Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task

Travis Wade[a] and Lori L. Holt
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

(Received 23 November 2004; revised 14 June 2005; accepted 11 July 2005)

This study examined perceptual learning of spectrally complex nonspeech auditory categories in an interactive multi-modal training paradigm. Participants played a computer game in which they navigated through a three-dimensional space while responding to animated characters encountered along the way. Characters' appearances in the game correlated with distinctive sound category distributions, exemplars of which repeated each time the characters were encountered. As the game progressed, the speed and difficulty of required tasks increased and characters became harder to identify visually, so quick identification of approaching characters by sound patterns was, although never required or encouraged, of gradually increasing benefit. After 30 min of play, participants performed a categorization task, matching sounds to characters. Despite not being informed of audio-visual correlations, participants exhibited reliable learning of these patterns at posttest. Categorization accuracy was related to several measures of game performance and category learning was sensitive to category distribution differences modeling acoustic structures of speech categories. Category knowledge resulting from the game was qualitatively different from that gained from an explicit unsupervised categorization task involving the same stimuli. Results are discussed with respect to information sources and mechanisms involved in acquiring complex, context-dependent auditory categories, including phonetic categories, and to multi-modal statistical learning. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.2011156]

PACS number(s): 43.71.−k, 43.71.An, 43.66.Ba, 43.66.Lj [ALF] Pages: 2618–2633

I. INTRODUCTION

Experience plays an essential role in shaping auditory perception in general, and speech perception in particular. However, there is a major complicating factor in characterizing this role experimentally; listeners come to the laboratory already shaped by considerable experience, the history of which may not be known to the experimenter. Since language experience cannot be controlled ethically, this is particularly troubling for speech perception. As a result, there are often limits on the certainty with which underlying learning or perceptual mechanisms can be inferred from patterns in adult (or even infant) perception. As a result, many long-standing questions concerning phonetic categories remain and current theories vary even in their most basic assumptions, including the very nature of perceptual objects (Diehl *et al.*, 2004; Fowler, 1986; Liberman and Mattingly, 1985; Lotto and Kluender, 1998; Nearey, 1997). In cases like this where ecological validity and experimental control are at odds, it can be useful to take a converging methods approach, for example examining adult and infant speech perception where control over experience is less realizable and, in addition, investigating experimental paradigms where strict control over the history of experience is possible. Along with studies of nonhuman animal perception and learning of speech sounds (e.g., Hauser *et al.*, 1998, 2000; Holt *et al.*, 1997, 2001; Kluender *et al.*, 1987, 1998; Sinnott *et al.*, 1976; Sinnott and Brown, 1997), human nonspeech auditory categorization designs provide an important controlled testing ground of the latter type. With properly selected nonspeech sounds, listeners' exposure to category instances can be carefully monitored. Observing the effects of this exposure during and after acquisition, then, aids in understanding the auditory and cognitive constraints upon sound category acquisition, and knowledge of these constraints in turn informs the examination of phonetic perception. Analogous lines of research have proven valuable in other domains; expert visual perception of non face objects, for example, has provided a new understanding of the cognitive and neural mechanisms involved in face perception (Gauthier and Tarr, 1997; Gauthier *et al.*, 1998; 1999a, b; Rossion *et al.*, 2002).

Observation of nonspeech category learning has already revealed some interesting and potentially informative parallels to speech categorization. However, interpretation of these results with respect to their link to phonetic categorization is challenged by the limited ecological validity of both the category stimulus distributions and the training methods used in studies thus far. Sounds and sound inventories for which learning has been observed are simpler by orders of magnitude than those involved in speech communication. Relatedly, the methods used to drive this learning have been limited to explicit training tasks demonstrably unlike anything encountered during speech category acquisition and considerably simpler than those used to affect non-face expertise in visual training studies (e.g., Gauthier *et al.*, 1998). The purpose of the present study is twofold: we describe a

[a]Address for correspondence: Posit Science, 114 Sansome Street, 5th floor, San Francisco, CA 94104. Electronic mail: Travis.Wade@positscience.com

method of exposing participants to novel nonspeech auditory categories that better resembles the natural category acquisition process and we explore the success of this method by investigating listeners' development of categories that, to varying degrees, involve some of the same challenges associated with the acquisition of naturally occurring speech categories. We first review a few relevant previous findings from nonspeech categorization studies.

A. Nonspeech auditory categorization

Several important parallels between speech and nonspeech categories have been observed. First, there is evidence that as a result of training human adults do, in fact, learn auditory categories that, like speech sounds, have nonlinear acoustic distributions. Lotto (2000), for example, taught listeners to categorize novel sounds as members of categories. Sounds varied continuously along one temporal and two spectral dimensions and categories overlapped thoroughly in all three acoustic dimensions. Category membership was determined by a complex rule and could depend on the acoustic characteristics of a given sound along any two of the three dimensions. Despite this complexity, however, listeners' categorization reliably improved over the course of training; they correctly labeled more sounds on the tenth day of training than on the first.

Also in line with phonetic acquisition, not all nonspeech acoustic distributions are learned with equal ease or in the same manner. Holt *et al.* (2004), for example, showed that general auditory perceptual discontinuities may interact with sound categorization. Listeners in this study learned sound categories defined along a single temporal dimension, tone onset time (TOT), a measure of the difference in onset time between two coterminous tones that has been previously used to model voice onset time (VOT) in speech (Pisoni, 1977). Nonspeech TOT categories were easiest to learn when the stimulus distributions defining the categories were positioned along the TOT dimension such that their boundary coincided with a temporal region (~20 ms) associated with increased discriminability in humans (Pisoni, 1977) and nonhumans (Kuhl and Miller, 1975) and known to underly a disproportionate number of phonetic distinctions in the world's languages (Keating, 1984; Lisker and Abramson, 1964). When the category distributions were shifted such that the natural peak in discriminability no longer coincided with the boundary between the category stimulus distributions, categories were more difficult to learn. Relatedly, Mirman *et al.* (2004) found qualitatively different learning patterns for categories as a function of the type of acoustic cue that differentiated category exemplars. Listeners who learned complex nonspeech sounds differing along a rapidly changing temporal dimension (amplitude rise time) were better at categorizing, but worse at discriminating, category exemplars than listeners who learned categories differing along a steady-state spectral dimension. This difference was observed even when pretraining sensitivity to the spectral and temporal cues was equalized across the two acoustic dimensions. This pattern of perception parallels differences observed in perception of steady-state (vowels and fricatives) and rapidly changing (consonants) speech sounds (e.g., Eimas, 1963).

In addition to the acoustic characteristics and distributional properties of sound categories, the category training procedure used also seems to have important consequences in auditory categorization. It is well known that, particularly for complex categories, the learner's task during training may affect resulting knowledge (e.g., Allen and Brooks, 1991; Ashby *et al.*, 1999, 2002). In nonspeech categorization, categorization-with-feedback designs have thus far been assumed to provide a reasonable approximation of the natural acquisition process. One notable finding in this respect was reported by Guenther *et al.* (1999) who examined listeners' ability to discriminate very similar within-category sounds as a function of different types of training. When training emphasized discrimination, listeners improved in detecting small acoustic differences among stimuli. However, when training emphasized categorization, they instead demonstrated "acquired similarity," becoming less sensitive to within-category acoustic differences. These findings may be informative regarding the structure of information encountered during acquisition, since perceptual warping of the acoustic space seems to accompany the categorization of speech sounds (Kuhl, 1991; Kuhl, 1992; cf. Lotto *et al.*, 1998).

The import of observations of nonspeech categorization is generally taken to be their qualitative similarity to patterns known to exist in speech perception. To the extent that effects in nonspeech acquisition and perception can be conclusively linked to similar effects in speech, nonspeech designs offer an important ground for investigating the limits and possible mechanisms governing categorization. However, also noteworthy are the differences between perception of speech and the nonspeech categories for which learning has been thus far observed. One striking difference involves the degree of competence typically reached in nonspeech training studies. Lotto (2000), for example, observed only 70% accuracy in participants' learning of two categories after ten hour-long sessions of intensive training; language users obviously must maintain many more categories much more accurately to achieve communication proficiency. Certainly, performance might continue to improve with more experience; language-acquiring infants receive thousands of hours of language exposure. However, with the impossibility of imposing this level of exposure in nonspeech training experiments comes the risk that the learning observed is fundamentally different from that involved in language acquisition (see Reber, 1989).

Another important (and potentially more manageable) difference involves the training method used in the studies. The categorization-with-feedback training commonly used in studies of nonspeech categorization is demonstrably quite unlike the processes by which humans are exposed to natural language sounds, and perhaps so fundamentally different as to preclude informative comparison. In the typical categorization training study participants sometimes undergo a short period of passive familiarization with the sounds to be learned and then hear a large number of exemplars from two

or (rarely) more categories. After each sound presentation, listeners press a button corresponding to a category label; following this response, some sort of visual feedback indicates whether the previously-heard sound corresponds to the same category as the button pressed.

Phonetic acquisition does not seem to involve explicit category labels, explicit response trials, or explicit feedback (e.g., Bruner, 1983; Jusczyk, 1997b). To the extent that category information accompanies category instances during (first) language acquisition, it involves complex correlations among acoustic event sequences and various visual, auditory, olfactory, tactile, and other events occurring in the environment. While it is far from clear how infants make use of this barrage of co-occurrences, it is known that human infants and adults are sensitive to statistical regularities at multiple levels (Jusczyk, 1997b; Kuhl *et al.*, 1992; Maye *et al.*, 2002; Saffran *et al.*, 1996; 1997, 1999). Moreover, there seem to be differences in the learning resulting from explicit training and that resulting from more incidental, implicit exposure. At least for patterns that are not particularly salient or easy to describe with simple rules, implicit exposure leads to faster, more accurate learning of the patterns than does explicit instruction (Reber, 1976; Reber *et al.*, 1980; Reber, 1989 for review). It has been suggested (e.g., Lacerda, 2003; Lacerda and Sundberg, 2004) that sensitivity to statistical regularities given rich, multimodal input from the environment results in the recognition of systematic patterns, including an awareness of phonetic categories that eventually interacts with other levels of linguistic processing.

With the aim of addressing this issue from multiple, converging methods, the purpose of this study was to develop and test a method of exposing learners to auditory categories that better resembles the natural acquisition process in these respects. In the next section a training method is described that captures several essential aspects of phonetic acquisition. In this method, subjects play a computer game in which they must navigate through a three-dimensional space, performing actions specific to animated characters they encounter along the way. Each of these characters is associated with both a distinctive movement pattern and predetermined sound category; an exemplar of this category is presented repeatedly each time the character is encountered in the game. Participants are not informed of the nature of these sound categories, nor of their significance in the game task, and other sound effects including a repetitive, synthetic background music score are also present throughout the game. However, as the game progresses, the speed and difficulty of the required tasks increase so that quick identification of approaching characters by means of their characteristic sounds is, while never required or explicitly encouraged, of gradually increasing benefit to the player. Following one or more sessions of this type of exposure, listeners complete a categorization task involving explicit matching of sounds to characters encountered during the game. This training method was used to examine listeners' categorization of a somewhat larger (four category) inventory of sound categories composed of sound exemplars that incorporate somewhat more of the nonlinear, context-dependent nature of speech sounds than have exemplars employed in previous auditory categorization studies. Categorization patterns were compared across differences in category structure (experiment 1) and to categorization patterns resulting from an explicit, unsupervised categorization task that did not involve the interactive game (experiment 2).

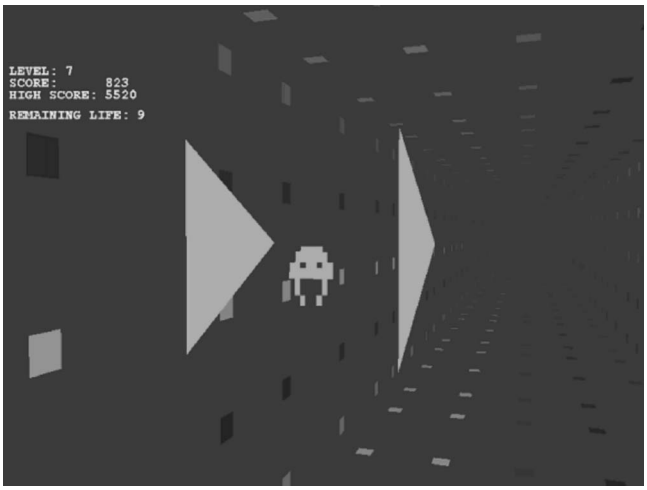

FIG. 1. Screenshot of typical game. The two-eyed figure in the center is an approaching irf-bat character; the orientation of the targeting graphic (two triangles) and background scene indicate that the player has adjusted the line of sight to the left to target the character.

B. Outline of the game

The design of the training task is conceptually, audiovisually, and ergonomically similar to that of typical commercial first-person shooting games. A screenshot of a typical game scene is shown in Fig. 1. For the duration of a game, the player moves forward at a constant perceived speed through a pseudo-three-dimensional tunnellike space. As the player progresses, he or she is periodically approached by animated *irf-bat* (interactive robot figure-based auditory training) characters, generally from the forward direction. The game involves four irf-bat characters that are easily distinguished from each other by shape, motion, and color patterns.

The player's tasks are to shoot and to capture these characters. Two of the four characters are *enemy irf-bats,* designated for shooting, and the remaining two are *friend irf-bats,* to be captured. These tasks are accomplished as follows, in a manner typical of similar games. Although the player moves uniformly forward throughout the game, it is possible to adjust the visual line of sight, to look or aim in any direction. This is accomplished using the LEFT, RIGHT, UP, and DOWN arrow keys with the right hand. Shooting and capturing require a combination of this movement and a multistep aiming process using the left hand. Specifically, when the Q key is pressed and held, a targeting graphic (the two triangles in Fig. 1) is illuminated in the center of the screen. This graphic is used to aim at enemy irf-bat characters in preparation for shooting. The player adjusts the sight line using the arrow keys, until a color change in the targeting graphic indicates that the irf-bat is currently on target for shooting. Finally, pressing the SPACE bar activates a shooting function. Capturing involves a similar process. When the

R key is pressed and held, a targeting graphic appears on the screen that aids the player in aiming at an irf-bat character using the arrow keys. Instead of pressing the SPACE bar, the subject must continually adjust the target using the arrow keys to keep the character in line as it approaches. As this is happening, the irf-bat character moves continuously forward, closer to the player's vantage point. Once the character reaches the player, with the capturing graphic displayed and on target, a capturing function is activated. If any irf-bat reaches the player's location without having been successfully shot or captured, it is said to have *escaped* and disappears from the player's line of sight.

Three variables figure prominently in the game's structure: *score, level*, and *life.* The player's primary objective is to acquire and maintain a high score. This is accomplished by shooting; each time the player successfully shoots an enemy irf-bat, the score increases, by an amount determined by (1) the current level and (2) the proximity of the character to the player at the time of shooting. More points are awarded for faster shooting and at higher levels of game play. The player advances one level each time a predetermined number of enemy characters (three in the present experiments) are successfully shot. Life is a measure of how many characters have recently escaped without being successfully shot or captured. At the beginning of a game, life is set at 10; it decreases by 2 each time a character escapes. This variable's value has no effect on the workings of the game, but the game terminates when it reaches zero. Its value increases by 1 each time a friend irf-bat is successfully captured and each time a level is completed. If the player shoots, instead of captures, a friend character, life is not increased and the score is decremented by a constant value. The values of each of these parameters in a given experiment is under the control of the experimenter.

C. Auditory category presentation in the game

In the present experiments, the game just described was used to present auditory categories that, like many speech sounds, are spectrally complex stimuli of a few hundred milliseconds duration. The game involves four sound categories, each of which possesses multiple exemplars and is associated with a single irf-bat character. The manner in which participants are exposed to these categories was designed to mirror several key aspects of natural phonetic acquisition.

Sound categories always co-occur with their associated characters. Each time a character appears visually in the context of the game, an exemplar of its corresponding sound category is presented auditorily. The sound is repeated continuously (with brief silences between repetitions) the entire time the character is active, i.e., from the time it is introduced until it is caught or captured or escapes. As a result, the auditory category (defined by multiple exemplars) tends to co-occur over the course of a game with both the visual image of the character and the distinctive motor/tactile patterns involved in shooting or capturing it. Whereas this combination is certainly simplistic compared to the rich set of visual, olfactory, auditory, kinesthetic, and other cues that may be correlated with speech sounds as they occur in the world, it represents much richer contextual support in category presentation than that typically present in explicit learning paradigms, where categories simply co-occur with their labels and feedback assignments. Additionally, repetition of a sound throughout a character's appearance is consistent with an apparently significant aspect of human language learners' early speech input; infant-directed speech appears to be repetitive (Fernald and Simon, 1984; Lacerda and Sundberg, 2004; Papousek *et al.*, 1985).

At the beginning of a game, exposure to these patterns of co-variance is expected to be fairly implicit, since knowledge of or attention to the acoustic exemplars is of no apparent consequence to game performance and no mention of their importance is made; participants are instructed only that they will be playing a video game. Characters appear near the center of the screen and approach the player slowly. Their accompanying sounds are merely part of a stylistically and texturally coherent game score, accompanied by background music and separate sound effects when the player shoots or captures a character or a character escapes. However, with each new level, the game becomes progressively more difficult in two ways. First, all motion in the game, including the approach of the characters and the targeting mechanism's responsiveness to the player's key presses, speeds up gradually. In addition, the characters begin approaching from locations that are gradually further displaced from the center of the screen. Each irf-bat character is associated with a single direction of origin (up, down, left, or right) from the center of the screen; on average (there is always an additional element of random noise), characters begin their approaches further in these directions as a game progresses.

These trends can be seen in the Fig. 2(a), which shows the starting locations of characters encountered over the course of a typical game. In this figure, the x and y dimensions represent the absolute distance in screen coordinates from the center of an approaching irf-bat character to the point at the center of the screen when the player is facing forward, a measure proportional to the apparent angular displacement of the character from the forward direction. Enemy characters approach the player in a straight line, and friend characters approach in a straight line infused with random jitter, so these locations remain nearly constant as long as a character is active. Each time a character is caught or captured or escapes, the player's line of sight is returned to the forward direction.

As shown in Fig. 2(a), as the game progresses, characters appear in more and more distal locations, requiring faster movement and hand-eye coordination on the part of the player. In the figure, the visible area of the game screen is represented by the range $(1,-1)$ in each dimension, so that the player's viewing frustum always extends only one unit in each direction from the current line of sight. Importantly, since this line is adjusted so that the player faces forward each time a character becomes inactive, only characters falling within the range $[(1, 1), (-1,-1)]$ are initially visible to the player; this range is represented by the shaded box. As shown in Fig. 2(a), at some point during a game (approximately level 8) the mean starting points of approaching characters move beyond this visible area. Since in this case an

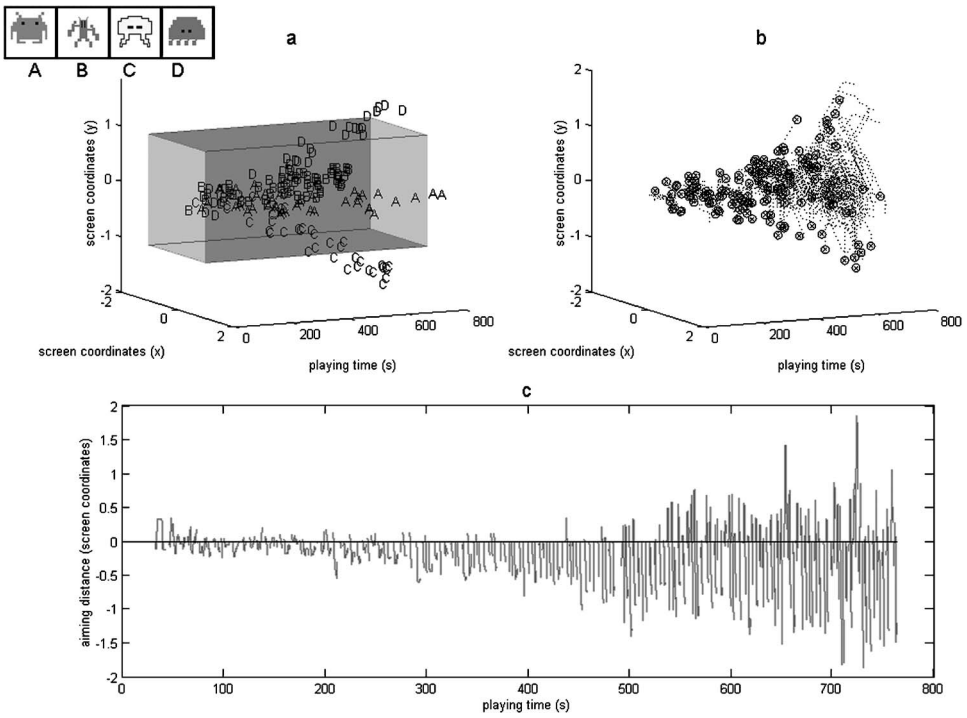

FIG. 2. Sample game. (a) Initial positions of irf-bat characters occurring during the game (inset shows individual characters), overlayed with the player's initial viewing frustum at each character's appearance. (b) Traces of player movements in response to the same characters. (c) Degree of player orientation toward characters over the course of the game, defined $((x_i^{\text{player}}-x_i^{\text{character}})^2+(y_i^{\text{player}}-y_i^{\text{character}})^2)^{1/2}-((x_0^{\text{player}}-x_0^{\text{character}})^2+(y_0^{\text{player}}-y_0^{\text{character}})^2)^{1/2}$.

irf-bat's sound is the only cue to its identity and location available to the player when it appears, the correlation between irf-bats, their sound patterns, and their typical starting locations is of increasing benefit to quick, accurate targeting. As starting points become still more distal, targeting becomes nearly impossible without quick categorization of sound patterns. Good performance at higher levels, then, requires a repeated, instantaneous, functionally oriented identification of sound categories that is generally not demanded of participants in explicit auditory category training studies.

Figure 2(b) shows a typical player's responses to the set of irf-bat character stimuli. The dashed line shows the player's current aim trajectory (also in screen coordinates), and $\otimes$ symbols denote successful hits and captures. Despite a lack of previous exposure to the game or to the sound categories, this relatively successful player was able to maintain game play for several minutes after characters began originating from outside the viewing area, apparently using acquired knowledge of sound-character-location correlations to successfully target many characters.

D. Measurement of category acquisition

There are several means of measuring and characterizing participants' acquisition of sound categories with the game (henceforth, Irfbats) task. One ecologically attractive method is to observe participants' ability to translate acquired knowledge to successful game performance. Rough measures such as the absolute or average high score or level reached in a training session have proven to be effective and are discussed in the next section. More detailed information can be obtained by examining players' aiming responses to individual audiovisual stimuli encountered over the course of the game. Figure 2(c) is derived from the same player movements depicted in Fig. 2(b), portraying the direction of movement with respect to target characters. The y axis represents, over the course of the game, the Euclidean distance in screen coordinates between the current (at each moment during game play) line of sight and the character's location, compared to this same distance at the time the character first appeared:

$$((x_i^{\text{player}}-x_i^{\text{character}})^2+(y_i^{\text{player}}-y_i^{\text{character}})^2)^{1/2}-((x_0^{\text{player}}-x_0^{\text{character}})^2+(y_0^{\text{player}}-y_0^{\text{character}})^2)^{1/2}.$$

Negative values, therefore, indicate the player is adjusting the targeting device toward the character. As shown by the sequences of positive values in Fig. 2(c), the player makes periodic mistakes, adjusting the target in the direction opposite the character, particularly later in the game when characters are initially invisible, occasionally allowing characters to escape. However, even very late in the game, the majority of movement is toward the (often out-of-range) character. This measure was compared across games and participants in the experiments described below.

A more direct means of investigating acquired category knowledge is performance on an explicit postgame categorization task, in which players match sounds with pictures of characters encountered during game play. Such a test was also implemented in the present experiments. In the test, participants view a screen in which all characters are displayed with numbers while a sound is presented, and responses are made by pressing number keys (1–4) on the same keyboard used during game play.

II. EXPERIMENT 1

The purpose of experiment 1 was to measure and characterize the effectiveness of the Irfbats training paradigm in learning complex nonspeech auditory categories. Employing this method, participants were exposed to an inventory of sound categories intended to present categorization chal-

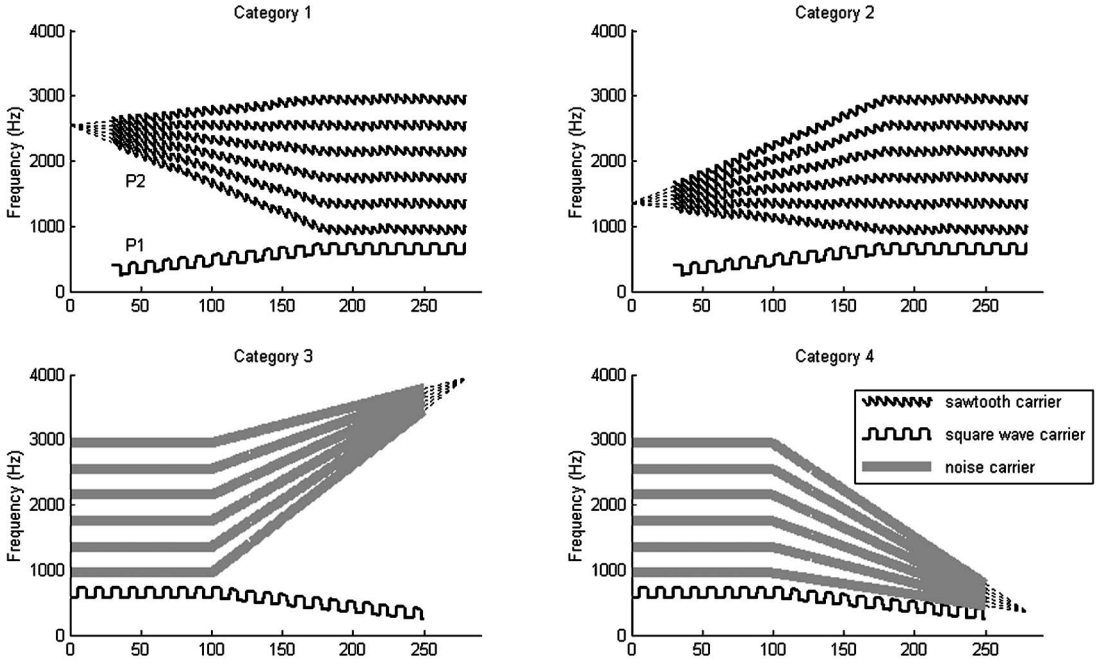

FIG. 3. Schematic representations category exemplars encountered during game play (each exemplar is comprised of the invariant P1 resonance and one P2 pattern.) Thin dashed lines show interpolation to P2 category loci for clarity.

lenges involving both spectral complexity and context-dependent category structure similar to those encountered in acquisition of phonetic categories.

A. Method

1. Stimuli

An inventory of four nonspeech categories, each composed of six stimulus exemplars, was used in training. Each stimulus was 250 ms in duration and had spectral peaks in two locations, *P*1 and *P*2. Category stimulus exemplars were differentiated by dynamic spectrotemporal patterns in P2, as shown in Fig. 3. For two of the categories, P2 onset frequency increased or decreased linearly over the initial 150 ms and then remained at a steady-state frequency for the following 100 ms. We refer to these stimuli as "onset" stimuli. The remaining two categories were "offset" stimuli with a symmetrical pattern; P2 frequency was constant for the first 100 ms and increased or decreased linearly to a final offset value across 150 ms. P1 frequency had a similar onset or offset pattern and was constant across stimuli within a category.

These stimuli modeled some of the spectrotemporal characteristics of speech signals in that P1 and P2 can be thought of as analogous to formant resonances of the vocal tract. Critically, however, the stimuli were perceptually very dissimilar to speech sounds and unlikely to have been perceived in a "speechlike" manner (Pisoni, 1987). No category involved a set of dynamic P2 patterns (category exemplars are described in detail below) that corresponded in absolute terms to observed formant patterns of any known set of phonologically equivalent speech sounds. Moreover, stimuli all possessed a complex fine temporal structure completely unlike that of speech. For all stimuli, the two spectral peaks were created by filtering two separate sources and combining the resulting waveforms additively. P1 was always derived from a square wave of periodicity 143 Hz. For onset stimuli, P2 was derived from a sawtooth wave of periodicity 150 Hz, and for offset stimuli it was derived from uniform random noise. (This last difference also helped to maximize the perceived difference between onset and offset categories).

To help ensure that stimuli were not perceived as speechlike, several naive observers, including some experiment 1 participants, were interviewed informally regarding the sounds. When asked for general impressions, observers invariably commented that they resembled "video-game sounds" or something similar; none mentioned speech sounds. When pressed to identify individual stimuli as speech sounds, responses were inconsistent across observers and did not reflect any relevant properties of the CV and VC sequences discussed above. Samples of the sounds are available online.[1]

Figure 3 shows schematized versions of P1 and P2 patterns of the six exemplars of each category encountered during the game task. The range of P2 patterns within categories was designed to reflect—to varying degrees—the same types of variability with which consonants are cued by formant transitions in the context of simple CV and VC sequences. Steady-state portions, roughly analogous to vowel place of articulation, varied in center frequency from 950 to 2950 Hz in 400-Hz steps within and across categories, thus carrying no first-order information to category membership. P2 transition trajectories were determined by a combination of (1) this steady-state frequency and (2) category-specific loci, toward or away from which P2 varied in frequency across time. As shown in Fig. 3, the beginning (offset stimuli) or end (onset stimuli) of P2 transitions always corresponded to the steady-state location; the transition itself spanned a linear trajectory of approximately[2] 83% of the distance from the

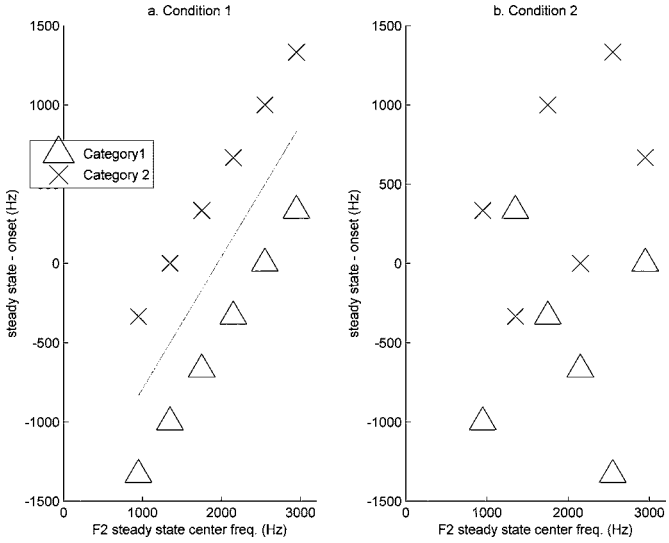

FIG. 4. Onset category stimuli shown in P2 steady state—onset trajectory space.

steady-state frequency to a constant target locus frequency for each category. This was intended to represent the way production (Delattre *et al.*, 1955, 1964) and perhaps perception (Sussman *et al.*, 1993, 1998) of consonant formant transitions varies depending on adjacent vowel context. For offset stimuli (categories 3 and 4), P2 loci were placed such that the categories would be easily distinguished based on only transition information. Since offset loci were either substantially higher (3950 Hz) or substantially lower (350 Hz) in frequency than the range of steady-state frequencies, the P2 trajectories of offset stimuli, while varying somewhat in slope and offset frequency, either always decreased in frequency or always increased in frequency within a category. For onset stimuli (categories 1 and 2), however, P2 loci were *within* the possible steady-state frequency range at 1350 and 2550 Hz and, as a result, no single invariant acoustic characteristic of onset transitions defined category membership. Category 1 onsets, for example, varied from steeply decreasing in frequency, to flat, to slightly increasing in frequency depending on the following steady-state frequency. As a result, the category 1 trajectories corresponding to the three highest P2 steady-state frequencies overlap completely in terms of slope with the category 2 trajectories preceding the lowest three steady-state frequencies.

Members of these categories, in particular the onset categories, thus lacked constant necessary and sufficient cues to category identity. This was intended to reflect the notoriously non-invariant nature of acoustic cues to many speech categories. However, also in line with phonetic categories (e.g., Lindblom *et al.*, 1992; Lindblom, 1996) and with many categories shown to be learnable in the visual domain (Ashby and Gott, 1988; Ashby and Maddox, 1990), onset stimulus categories were in fact linearly separable in a slightly higher dimensional space. Figure 4(a) shows P2 transition slope, a cue known to be useful in consonant discrimination (e.g., Liberman *et al.*, 1954), plotted against steady-state frequency for the onset categories (categories 1 and 2). As indicated by the dotted line, perfect discrimination between the two classes may be achieved by simple integration of information from these two sources. Thus, although there is no first-order acoustic information upon which to reliably base categorization decisions, there is higher-order structure that may be of use to learners. As a first step in evaluating the importance of this higher-dimensional relationship in category learning, a control condition (condition 2) was devised in which this relationship was absent. Stimuli for this condition are represented in Fig. 4(b); whereas categories 1 and 2 possess precisely the same P2 steady-state and trajectory ranges—and the same degree of cross-category overlap in these two dimensions—as the set just described (condition 1), the correlation between the two cues, rather than being determined by a category-specific locus, is pseudo-random. To the extent that information integration takes place as a result of category learning in game play, it was predicted that participants in condition 1 would outperform condition 2 participants during the game and/or in posttest category identification.

All sounds were created using Matlab (Mathworks, Inc.). Source signals were first generated at a sampling frequency of 22.05 kHz and filtered with an eighth-order elliptical bandpass filter with 2-dB peak-to-peak ripple, 50-dB minimum attenuation, and 500-Hz bandwidth. After filtering, all spectral peaks (P1 and P2 within and across categories) were equalized for rms amplitude, and 25-ms linear on-off ramps were applied. Finally, waveforms for each pair of formants were added together. Following synthesis, stimuli were inspected using spectrogram and waveform representations and found to closely match the intended parameters depicted in Fig. 3. A constant 50-ms silent interval separated repetitions of individual stimuli during game play.

2. Procedure

The game procedure used in training was identical to that described in the introduction; each of the four categories just described was associated with one of the four characters pictured in Fig. 2 and one direction of approach. Specifically, the two friend characters [A and B in Fig. 2(a)] progressively appeared from the right and left of the screen, respectively, and enemy characters [C and D in Fig. 2(a)] came from the bottom and top. Friend and enemy classes each included one onset and one offset category, such that P2 transition patterns (and not simply the onset-offset difference) took on the function of denoting character type. Friend characters A and B involved the two falling P2 patterns (high onset category 1 and low offset category 4, respectively) and characters C and D were matched with rising P2 patterns (categories 2 and 3).

Subjects were first familiarized with the game with a short tutorial program in which they were allowed to practice capturing and destroying stationary characters as the experimenter verbally explained the concepts of the game. No sounds were present during this familiarization session, and no mention was made of sound categories or their importance in the game task. Once subjects demonstrated to the experimenter's satisfaction that they understood the game procedure (familiarization typically took about 5 min), they were given a pair of headphones and informed that they

would play the game for 30 min, during which time their objective would be to achieve as high a score as possible. A *pause* screen (toggled by pressing the P key) was available to remind players of the details of the task, including summary instructions for shooting and capturing and descriptions of visual characters as friend or enemy. Participants were encouraged not to pause the game unless it was necessary. The participant was instructed to press the F1 key to begin a new game each time the life variable reached zero and a game terminated. Individual games ranged in length from tens of seconds to several minutes, and the number of games completed during the 30-min session ranged from 2 to 14.

Following the game session, subjects completed a categorization test in which they matched visual images of the irf-bat characters with exemplars of the four sound categories. Sound-category exemplars in the test were the 24 stimuli depicted in Fig. 3 and presented during game play and five novel sounds created to match the characteristic locus of each category. Novel stimulus P2 trajectories were determined by the same locus relationship depicted in Fig. 3, but had steady-state frequencies intermediate those of each pair of adjacent values used in training (1150–2750 Hz in five 400-Hz steps). Each stimulus was presented four times in the test, for a total of 176 trials, in random order. To provide maximal continuity with the game task, the same auditory and visual backgrounds were present during the test. A trial began with the simultaneous presentation of a sound stimulus and appearance of all four character images on the screen, arranged horizontally in an arbitrary constant order and accompanied by a number 1–4. As in the game, the sound repeated, with 30-ms silent gaps between repetitions, for 1.5 s or until the subject pressed a number 1–4 on a standard keyboard to register a response.

Game and test sessions took place in sound-attenuated booth using a laptop computer. All sounds were presented diotically over linear headphones (Beyer DT-150) at approximately 70 dB SPL.

3. Participants and design

Forty-two college students from the Carnegie Mellon University community reporting normal hearing participated in the experiment. Participants received undergraduate psychology credit for their participation.

To test whether any observed effects were due to the arbitrary mappings of categories to screen directions, visual characters, source signals, and stimulus types (onset versus offset) described above, two additional control subconditions were introduced in condition 1. Although the number of possible manipulations introduced by the richness of the game environment precluded fully factorial counterbalancing of game elements, we expected that these conditions would capture the effects of any major confounds. Condition 1b addressed the possibility that preference for a particular visual character or direction of origin might affect learning for some categories. In condition 1b the character correspondences described in Sec. II A 2 were reversed, with categories 1,2,3, and 4 occurring with characters C, A, B, and D, respectively. Since this manipulation involved a change in character assignment, location, and task (*shoot* or *capture*) for each sound category, it was predicted that any major bias patterns would be revealed in differences in performance between condition 1a (the originally described mappings) and condition 1b. Condition 1c was designed to test whether the difference in source signals between onset and offset categories would contribute to differences in learning. In this condition, the source-category type correspondences from condition 1a were reversed, so that onset category P2 resonances were derived from noise and offset P2s from the sawtooth source.

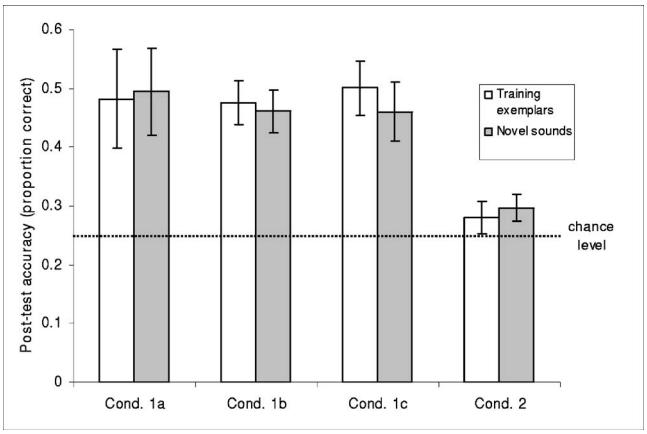

FIG. 5. Posttest accuracy across experiment 1 conditions (error bars show standard error of the mean).

Ten participants each were arbitrarily assigned to conditions 1a, 1b, and 1c. The remaining 12 participants were assigned to condition 2. (Condition 2 character-source-category correspondences were constant, identical to those of condition 1a).

B. Results

Participants gave a variety of reactions when faced with the posttest task of matching sounds to visual characters. The continuum of responses ranged from total surprise to relative confidence, although self-reported category knowledge did not always correspond with test performance.

1. Overall learning effects

Figure 5 summarizes overall categorization posttest performance across experiment 1 conditions. It was assumed that, if participants learned patterns of covariance between characters and sounds as a result of the game task, they would be able to match characters to sounds in a subsequent explicit identification task. When game characters cooccurred with onset categories defined by the structured variability patterns shown in Fig. 4(a), this learning indeed occurred. Subjects in conditions 1a, 1b, and 1c all performed reliably above chance level (25%) at posttest, for both previously heard [condition 1a: $t(9)=2.76$, $p=0.022$; condition 1b: $t(9)=5.92$, $p<0.001$; condition 1c: $t(9)=4.28$, $p=0.002$] and novel [condition 1a: $t(9)=3.3$, $p=0.009$; condition 1b: $t(9)=5.91$, $p<0.001$; condition 1c: $t(9)=5.29$, $p<0.001$] category exemplars. No overall differences in accuracy were observed between familiar and novel stimuli or between condition 1a, 1b, and 1c participants, indicating that

the learning effect was reproducible and robust over (at least) modest changes in the game procedure and the acoustics of sound category exemplars. Critically, condition 1 participants in both subgroups performed better than chance both for offset category stimuli, which were linearly separable by P2 trajectory [condition 1a: $t(9)=2.92$, $p=0.017$; condition 1b: $t(9)=7.65$, $p<0.001$; condition 1c: $t(9)=5.48$, $p<0.001$] and for onset category stimuli, which were not invariant in this manner [condition 1a: $t(9)=3.17$, $p=0.022$; condition 1b: $t(9)=4.31$, $p=0.002$; condition 1c: $t(9)=3.48$, $p=0.007$].

Condition 2 participants heard onset categories that were not structured reliably in the higher-dimensional acoustic space defined by P2 trajectory and steady-state frequency dimensions [Fig. 4(b)], and did not demonstrate learning. These participants did not differ reliably from chance in posttest identification of either novel [$t(11)=2.1$; $p=0.06$] or previously encountered [$t(11)=1.08$; $p=0.302$] stimuli. A one-way between-subjects ANOVA revealed that condition 1 participants reliably outperformed condition 2 participants in accurately categorizing both novel [$F(1,40)=11.69$; $p=0.001$] and previously encountered [$F(1,40)=13.5$; $p=0.001$] stimuli. Moreover, the difference observed was rather large; the difference in overall *p(c)* between conditions 1 and 2 was 19%, corresponding to an effect size $d=1.12$ (conservatively using the larger of the two observed standard deviations, from condition 1). This effect size corresponds to a power of approximately 0.8 with the smaller condition 2 sample size $n=12$, so it seems the moderate subject numbers were sufficient.

Since only condition 1a participants' sound-character pairing perfectly matched that of condition 2 participants (condition 1b differed in sound-character and condition 1c in source-category correspondences), this comparison was repeated including only condition 1a and condition 2 participants. The results were similarly reliable [novel stimuli: $F(1,20)=7.63$, $p=0.012$; familiar stimuli: $F(1,20)=6.003$, $p=0.024$].

2. Category comparisons

The top two rows of Fig. 7 show responses to each exemplar of each of the four categories at posttest across steady-state frequencies. Here several important trends in categorization are apparent. First, while condition 1 participants tended to recognize stimuli from all four categories reliably, condition 2 participants performed uniformly at chance level. Interestingly, this was true even for offset categories 3 and 4, for which exemplars were identical to those heard by condition 1 listeners. It would seem that the lack of structure in the onset categories discouraged condition 2 participants from making use of any audio-visual correspondences in the game.

To test for effects of the complexity of category distinctions on their learning for condition 1 subjects, a 3(training condition; 1a, b, c) $\times$ 2 (category type; onset versus offset) mixed model ANOVA compared subjects' sensitivity (%hits − %false alarms) to categories of each type. A main effect of category type was observed [$F(1,27)=24.25$; $p<0.001$], indicating that the unidimensionally defined offset categories were recognized slightly more accurately than the onset categories. This was not at all surprising given the more difficult nature of the two-dimensional onset distinction. Additionally, while no training condition main effect was observed ($F<1$), the training condition $\times$ category type interaction reached significance [$F(2,17)=4.6$; $p=0.019$]. *Posthoc* comparisons indicated that this effect was due to a slightly greater category type difference in condition 1c, where noise replaced a sawtooth wave as the source for the onset P2 resonance. Whereas learning occurred for both groups, the harmonic source thus seems to have been a slightly better carrier of P2 information than the noise source.

Finally, inspection of category 1 and 2 identification by condition 1 subjects suggests that, while the structured variability in the onset stimulus cues helped the listeners to establish the two categories, the high-level distinction was not learned perfectly. Specifically, category 1 accuracy tends to decline at higher P2 steady-state frequencies (i.e., where P2 is falling), while category 2 accuracy is worst for the lowest steady states (where P2 is rising). This suggests that (some) listeners may have been relying too heavily on the slope of the P2 transition and not compensating maximally for the steady-state part of the locus rule. Still, however, the fact that condition 1 subjects outperformed condition 2 subjects indicates that the higher-order structure played a role in learning, even in the brief (30-min) training period employed.

3. Game performance effects

In characterizing the effect of the Irfbats task on listeners' categorization, it was informative to examine performance during the game as well as at posttest. Due to errors in test administration, game performance data from two participants (one in condition 1a, one from condition 1b) were not recorded for comparison with posttest scores. Figure 6 shows posttest accuracy plotted against three measures of success at the game task for the remaining 40 listeners: mean high score achieved, mean high level attained, and the mean aiming distance measure pictured in Fig. 2(c).

Figure 6 illustrates a few important patterns in game performance across participants and groups. First, condition 1 participants (games with structured onset category variability) outperformed condition 2 participants (random onset variability) not only at posttest but also during the game itself. On average, condition 1 participants achieved higher scores [$F(1,38)=8.7$, $p=0.005$; means (s.d.) for conditions 1, 2: 40 230 (14 763), 26 998 (7012)], reached higher levels of the game [$F(1,38)=8.5$, $p=0.006$; means: 11.5 (2.7), 9.01 (1.7)], and navigated more accurately toward characters throughout the game [$F(1,38)=4.3$, $p=0.045$; means: −33.02(12.1), −24.9(9.05)] than did condition 2 participants. Moreover, game performance and category learning accuracy appear to be related. For condition 1 participants, all three performance measures correlate reliably with post test accuracy; players who demonstrated success (higher score and level values) and accuracy (lower aiming difference values) during the game also exhibited better category learning at posttest. It is not surprising that these correspondences did not hold for condition 2 subjects, who did not

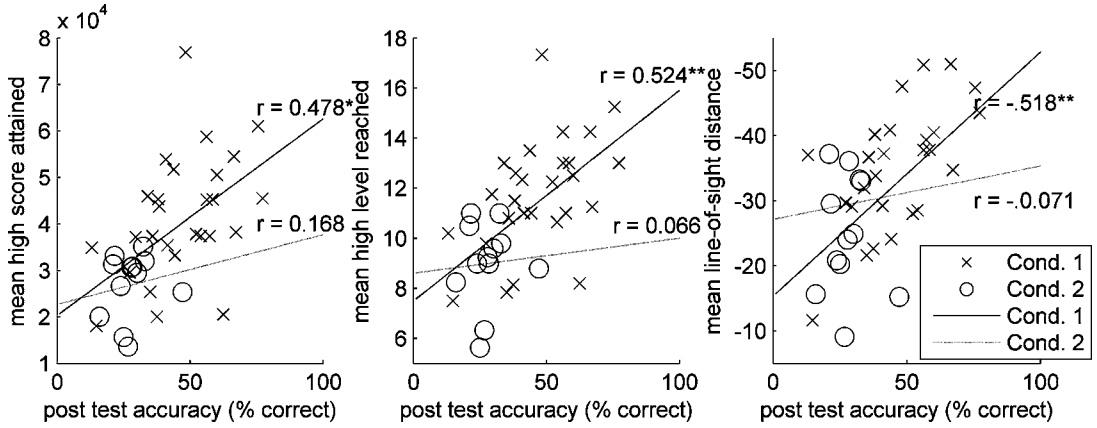

FIG. 6. Experiment 1 posttest accuracy plotted against game performance measures. Regression lines indicate correlation patterns between measures for each participant group * indicates correlation $p<0.05$, ** indicates $p<0.01$.

demonstrate learning overall, although this might also be due to the smaller total number of subjects in condition 2.

Thus, the higher-order structure imposed on condition 1 stimuli affected category learnability and learning in the Irfbats exposure task was sensitive to differences in this structure. At posttest, condition 1 participants (1) demonstrated learning for both simple and more complex category exemplar distributions and (2) reliably outperformed condition 2 participants, who did not learn overall. Furthermore, it seems likely that the interactive, incidental nature of the game was important in achieving this result, since success at the Irfbats task was related to posttest performance both within and between participant groups. This indicates that the game task was not merely a superficial addition to an unrelated incidental learning task. Game performance was also affected by category learning and sensitive to differences in category structure.

III. EXPERIMENT 2

Experiment 1 demonstrated that simple audio-visual correlation patterns in a game task enabled listeners to learn complex sound categories, whether the category distinctions involved unidimensionally invariant acoustic cues (categories 3 and 4) or required integration of two or more cues (categories 1 and 2). In beginning to characterize the type of learning that took place in familiar terms, it is important that both of these distinction types were tested. Previous studies have shown that training conditions affect learning differentially depending on the complexity of category distinctions. Ashby *et al.* (1999), for example, observed visual categorization under unsupervised conditions, where observers grouped visual patterns into a given number of categories. Optimal categorization occurred only when stimuli were unidimensionally separable, and not when distinctions involved more than one dimension. Similarly, Ashby *et al.* (2002) found an interaction between category structure and the way category information was presented during training. "Observational training," in which category labels were presented to participants simultaneously with training stimuli, and "feedback training," in which corrective feedback was provided after stimuli (and sometimes after a subject response), led to similar learning for simple, unidimensional category distinctions. However, for categories whose recognition required integration of information along two separate visual dimensions (length and orientation), feedback training provided a substantial advantage. These differences were tentatively taken as evidence of the learning systems involved, namely whether a simple explicit system was complemented by more complex, implicit learning over the course of training (Ashby *et al.*, 1998).

The game used here combines elements of all of these methods of category exposure, as (perhaps) does natural language acquisition. At various points in the game, the relative order of occurrence of auditory stimuli, visual characters, and participant responses reflect elements of unsupervised, observational, feedback, and other types of training. As a first step in comparing learning in the game task to that previously observed in more controlled designs, and in general to further characterize the effects of the game on learning, an additional experiment measured learning of the same sound categories used in experiment 1, in more explicit circumstances and absent the game task. Specifically, an unsupervised training design was used, in which listeners were told the number of sound categories and exposed repeatedly to category exemplars but given no category label information or other feedback. Essentially, this was expected to reveal which aspects of category structure were critically presented by the structure of the game environment and those that were more self-evident given minimal instruction and explicit, deliberate comparison of salient acoustic properties of the stimuli.

A. Method

1. Stimuli

Stimuli were identical to those used in experiment 1. Only the six exemplars of each category presented during experiment 1 game play and depicted in Fig. 3 were used in exposure and testing.

2. Procedure

The categorization task consisted of five familiarization blocks alternated with five transfer blocks in a single session. Each block consisted of three repetitions of each of the six

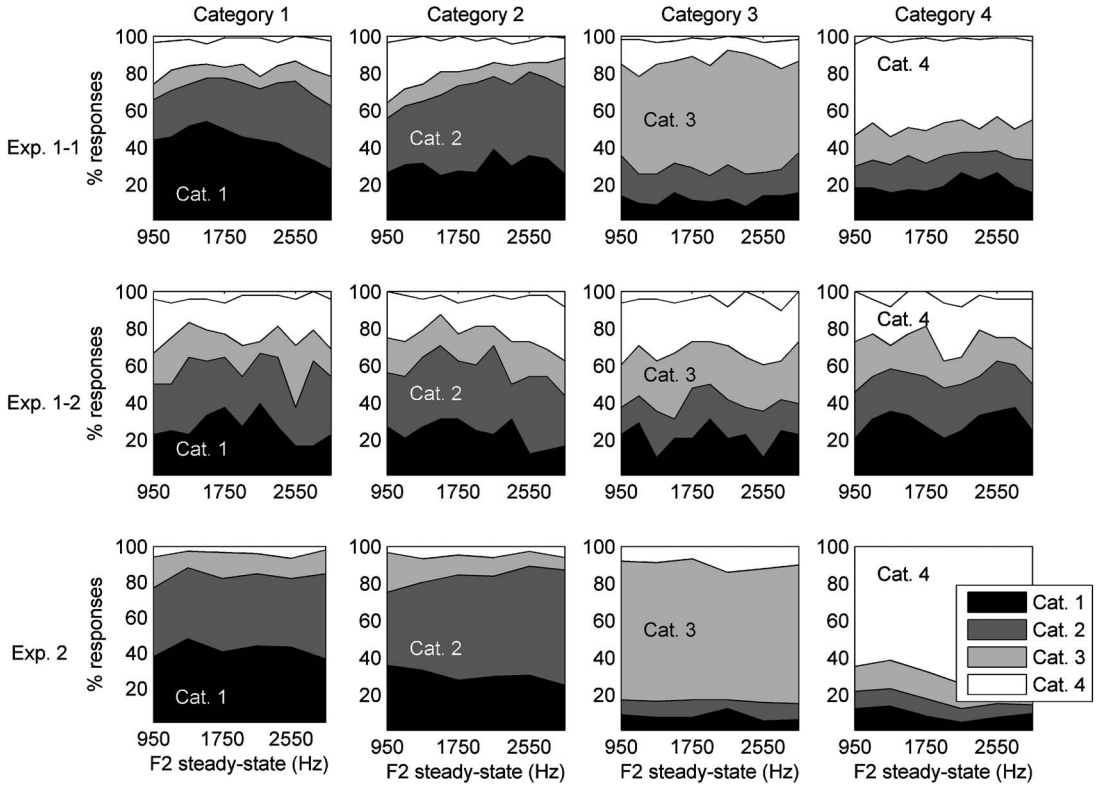

FIG. 7. Response patterns across test stimuli for experiment 1, condition 1 (a and b collapsed; top), condition 2 (middle), and experiment 2 participants. Labeled areas correspond to intended category responses; total area less than 100% indicates missed (timed-out) responses.

exemplars of each category, for a total of 720 stimuli (360 familiarization, 360 transfer). This was in keeping as closely as possible with session length and stimulus exposure experienced in experiment 1 (experiment 1 participants, on average, encountered 367.3 characters during their 30 min of game play). Participants heard isolated sound stimuli in sound-attenuated booths over headphones at approximately 70 dB SPL. Stimuli were presented in random order within blocks, using *ALVIN*, a software system recently developed by Hillenbrand and Gayvert (2005). In familiarization blocks, participants were instructed to listen to each stimulus carefully and learn as much as possible about the sounds in order to reliably divide them into four categories, pressing a button labeled "continue" after each trial. In transfer blocks, they were instructed to press one of four numbered buttons, depending on which of the four arbitrarily labeled categories they chose to assign to the sound's category. Participants were urged to be as consistent as possible and were informed that perfect performance was possible, following Ashby *et al.* (1999).

3. Participants

Ten college students from the Carnegie Mellon University community reporting normal hearing participated in the experiment. Participants received undergraduate psychology credit for their participation.

B. Results

Each participant's response patterns were first translated into a set of category labels by choosing the set of response-to-stimulus category mappings that maximized overall accuracy (% correct). Resulting overall accuracy scores averaged 59.9%, well above chance (25%) performance [$t(9) = 6.003$; $p < 0.001$] and in fact somewhat higher than overall experiment 1, condition 1 (a,b; condition 1c involved a different set of sounds) game participants' scores [$F(1,28) = 3.41$; $p = 0.076$]. This last result was not especially surprising or indicative of fundamental differences in learning. Experiment 2 involved an explicit auditory categorization task, while exposure to sounds in experiment 1 was purely incidental (participants did not even know they were learning sounds), so direct comparison of overall accuracy after a single session is not particularly informative. More critical was the interaction of training type and whether category distinction cues were unidimensional (offset stimuli) or integrative (onset stimuli) in nature.

The bottom row of Fig. 7 shows responses to each exemplar of each of the four categories averaged across the five transfer blocks. A qualitatively different categorization pattern seemed to result from explicit unsupervised training. Experiment 2 participants performed quite well on the linearly separable offset categories, but confused the two cue-integrating onset categories (1 and 2) with each other at near-chance level. To illustrate this pattern more clearly, Fig. 8 shows the difference between correct responses and locus-differing competitor responses for onset and offset categories across testing conditions (zero indicates chance-level performance). This comparison demonstrates that, while experiment 2 participants labeled offset categories fairly accurately, they responded to onset categories with almost no tendency to discriminate between competitors. A

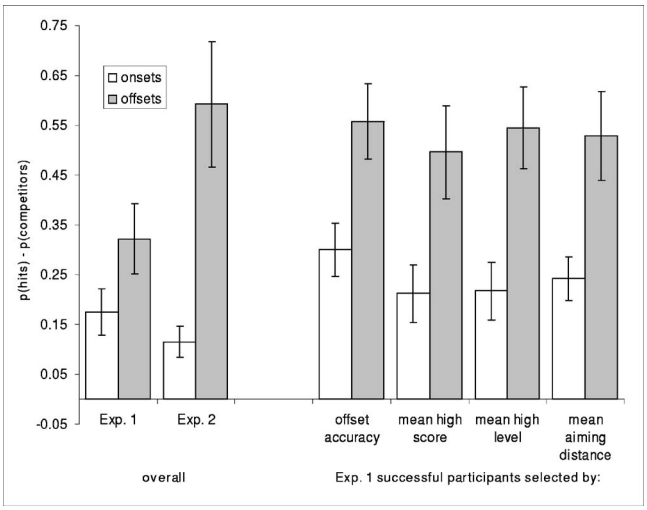

FIG. 8. Onset and offset distinction accuracy (correct responses minus locus-varying competitor responses) across training conditions (left) and for experiment 1 subsets (right) defined by removal of ten worst-performing participants on four measures of game training success (error bars show standard error of the mean).

2(training condition) $\times$ 2(category distinction type) mixed model ANOVA with training condition as a between-subjects factor revealed that onset categories were most difficult overall [$F(1,28)=26.4$; $p<0.001$]. The training condition effect did not reach significance; critically, however, a training $\times$ category type interaction was observed [$F(1,28)=7.5$; $p=0.011$]. Relative to overall performance, unsupervised categorization participants had considerably more difficulty with onset stimuli compared to offset stimuli.

One interpretation of this difference is that game participants, like observers trained with feedback in previous experiments (e.g., Ashby *et al.*, 1999, 2002), were relatively better able to learn the higher-dimensional onset distinction than listeners explicitly comparing the sounds in the unsupervised categorization task. However, the training $\times$ category interaction cannot be conclusively attributed to an advantage on complex distinctions for game participants, since absolute performance levels precluded direct comparison of responses across groups. The unsupervised training might instead have given experiment 2 listeners a special advantage in categorizing the offset categories, although it is unclear what the source of this advantage might have been. As a first step in addressing this issue, it was observed that the Irfbats task introduced at least two sources of variability to the category identification task that were not introduced by experiment 2's simple, constant procedure. Absolute average performance in experiment 1 posttest was probably deflated by subsets of participants who either (1) due to initial difficulty with the game task failed to advance sufficiently during 30 min of play for much learning to occur or (2) were surprised or confused by the posttest task and did not optimally display their acquired knowledge. Although study involving longer-term game play will be required to solve these problems completely, for the present data it was useful to equate performance across the tasks by considering subsets of experiment 1 participants who were successful at the game task and able to demonstrate acquired category knowledge in the explicit posttest. The right portion of Fig. 8 shows onset and offset identification patterns for Irfbats participants selected based on several possible criteria. Offset category identification accuracy was selected as a measure of success at the posttest task, since considering experiment 2 results it should have been fairly easy for successful participants and since it did not directly involve the more critical onset performance. For game success, the mean high score, level, and aiming speed measures presented above were considered. The performance of successful players was similar across criteria; they performed similarly to experiment 2 participants for offset categories and tended to *outperform* experiment 2 participants for onset categories. (The onset difference reached significance only when the offset score or mean distance criterion was used.) Thus, it does seem that, at least for the more successful participants, Irfbats-style feedback provided an advantage for learning the complex onset category distinction.

IV. GENERAL DISCUSSION

This study introduced Irfbats, an interactive video game developed for use in investigating the acquisition of auditory categories. While playing Irfbats, participants were incidentally exposed to sound category exemplars in the presence of other richly correlated multimodal cues. Observations of postgame sound categorization and patterns in game performance allowed for measurement of the types of auditory distributions that are learnable in the absence of explicit feedback.

A. Posttest categorization

Postgame sound categorization patterns demonstrated that even without explicit feedback Irfbats players can learn spectrally complex non-invariant auditory categories within a rather short period (30 mins) of incidental exposure when higher-dimensional acoustic cue relationships are present. In experiment 1, the inventory of sounds presented during the game included both categories possessing distinctive, invariant spectral cues (rising versus falling offset patterns in categories 3 and 4, as shown in Fig. 3), and categories that were not unidimensionally separable (onset categories 1 and 2). Much has been made of the significance of non-invariant category cues with respect to speech perception (Delattre *et al.*, 1955; Kluender *et al.*, 1987; Stevens and Blumstein, 1981). Like the speech categories they modeled, the non-invariant category distributions were linearly separable in a higher-dimensional space when two separate acoustic cues (onset trajectory and steady-state frequency) were integrated. Experiment 1 demonstrated that participants exposed to such categories exhibited robust learning over the course of a single 30-min game session. In a posttest they were able to match visual characters from the game reliably to sound category exemplars encountered during game play, as well as to novel sound exemplars drawn from the same category distributions. Another group of participants heard categories that possessed the same distributions of onset trajectories and steady states but lacked structured second-order cues to cat-

egory membership. These listeners did not differ from chance in their posttest categorization responses.

This finding is of potential significance concerning the types of knowledge required for context-dependent speech perception. Though, like many phonetic categories of the world's languages, onset categories 1 and 2 could not be differentiated by any single invariant acoustic cue, listeners (experiment 1, condition 1) learned the sounds, even without feedback or instructions to learn the categories. Together with the fact that nonhuman animals can learn similarly non-invariant speech categories (Kluender *et al.*, 1987), this finding is consistent with an account of speech perception that exploits general learning mechanisms for phonetic acquisition (Diehl, *et al.*, 2004) rather than specialized processes (e.g., Liberman and Mattingly, 1985; 1989; Trout, 2001).

Learning in the game was also compared with that resulting from an explicit unsupervised categorization task in experiment 2. An interaction involving category type was observed, such that game participants showed relatively more learning for the non-invariant onset categories requiring cue integration. This is informative in beginning to characterize the type of learning that resulted from the game. With respect to category exemplar distribution effects, the categorization responses of game participants were more like learners trained with explicit feedback in previous studies; exposure through game play seems to provide a similar advantage to feedback in learning complex, information-integrating category structures (Ashby *et al.* 1999, 2002), whereas observation and unsupervised training are helpful only for simpler, invariant categories. Additional study will be required to fully characterize the effects of the interactive game task on the types and extent of learning. In particular, comparison of category knowledge resulting from extended, explicit feedback training with that of expert game players will be necessary.

B. Interactive task effects

The structure of the Irfbats game was designed to model some of the interactions and multimodal correlations through which listeners may come to recognize the acoustic regularities underlying the sound distributions of a native language. Repeated presentations of category exemplars co-occurred with distinct visual events and motor tasks inherent to the game. Although sound categorization was not explicitly mentioned to or required of the players, it was of increasing benefit to achieve high levels of game performance as play progressed.

Indeed, it does not appear that the game task provided only a superfluous precursor to the more standard categorization posttest; rather, several characteristics of players' game performance proved to be good predictors of posttest categorization accuracy. Participants' ability to achieve higher scores, reach higher levels, and navigate successfully in the game environment was related to their knowledge of sound categories, and also was sensitive to differences in the non-invariant category distributions. As participants acquired knowledge of the sound categories relevant to the game task, skill at the task increased. Likewise, as increased skill enabled players to reach more difficult stages of the game, continued success demanded increasingly efficient, accessible knowledge of sound categories. This pattern is consistent with the interactive processes known to be important in category learning in other species (e.g., Baptista and Petrinovich, 1986; Eales, 1989) and thought to underly the phonetic and phonological acquisition processes (Bruner, 1983; Kuhl *et al.*, 2003; Lacerda, 2003; Lacerda and Sundberg, 2004).

Developing a precise mechanistic account of how experience shapes the acquisition of phonetic categories is challenging because it is impossible to attain full control over the histories of listeners' experience with speech. Even very young infants possess significant experience with, and demonstrate sensitivity to, the sound structure of the native language (e.g., Jusczyk, 1997a; Kuhl *et al.*, 1992). In domains like this where control over the input is elusive, a converging methods approach can be useful in balancing the competing demands of experimental control and ecological validity. One perspective on this issue is that understanding how human listeners extract information from the auditory environment can be informative about the constraints and mechanisms the system brings to phonetic (speech) categorization. Presently, there exists a limited literature describing general auditory (nonspeech) categorization (e.g., Guenther *et al.*, 1999; Holt *et al.*, 2004; Lotto, 2000; Mirman *et al.*, 2004), so there is still much that can be learned about how listeners categorize acoustic stimuli using these more traditional methods. Nevertheless, we believe that the present paradigm has value in providing an intermediate step along the continuum from experimental control to naturalistic observation. To be sure, this video game environment is a considerable step removed from the rich structure present in learning speech categories. Even so, Irfbats appears to capture some of the characteristics of correlation among multiple cues, function-based categorization, and exploration of an environment that likely accompany phonetic category acquisition. Just 30 min of incidental exposure to the multimodal statistical regularities present in the game was sufficient to promote category learning for a set of stimulus exemplars that model one of the central challenges in speech categorization, the lack of invariance.

Recent studies support the utility of examining nonspeech learning to understand potential mechanisms available to speech perception. Adults, infants, and nonhuman primates exhibit sensitivity to the statistical regularities present in sequences of speech syllables (Hauser *et al.*, 2000; Saffran *et al.*, 1996, 1997) and statistical learning of this sort appears to be deployed for both speech and nonlinguistic sounds (Saffran *et al.*, 1999). Thus, there is evidence that general (statistical) learning mechanisms may be operative across linguistic and nonlinguistic sound classes. The Irfbats paradigm relates to this literature in that it likewise requires sensitivity to statistical regularity for learning to be observed. However, the present work differs in a couple of important ways from previous approaches.

For the most part, investigation of statistical learning has been limited to single modalities. Statistical learning appears to be operative for both auditory (e.g., Saffran *et al.*, 1997)

and visual (Fiser and Aslin, 2002; Kirkham *et al.*, 2002) stimuli, but there thus far has been little investigation of how multimodal regularities are learned. It is, of course, presumed that learners take advantage of informative regularity where it exists, even if it occurs cross-modally (e.g., Massaro, 1987). The present paradigm provides a means of addressing this assumption explicitly and gaining an understanding of potential constraints upon the types of multimodal regularities that are learnable. Here, we have demonstrated that listeners acquire sound categories for spectrally complex, non-invariant acoustic exemplars through incidental exposure in a video game environment provided that they possess second-order acoustic regularity and they covary with multimodal perceptual/motor cues introduced in the game.

The Irfbats game is also differentiated from previous studies of statistical learning by how it assesses learning. Investigations of statistical learning have primarily relied upon measures of familiarity to assess whether participants learned statistical regularities presented in an experiment; adult participants, for example, are above chance at judging whether stimuli are consistent (familiar) or inconsistent (unfamiliar) with the regularities they encountered incidentally in previous exposure (Saffran *et al.*, 1997), and extensions of this same paradigm are used with human infant and nonhuman primate listeners (Hauser *et al.*, 2000; Saffran *et al.*, 1996). The results of the present studies move a step beyond familiarity. Participants were able to apply what they learned in the course of the game to the task of categorization.

C. Conclusions

The present results provide evidence that adult listeners can solve what has been thought to be a rather difficult auditory categorization challenge, acquiring categories for spectrally non-invariant acoustic exemplars, in a task in which categorization is largely incidental. Observation of learning was contingent upon the existence of a higher-dimensional acoustic relationship between the non-invariant cues and the presence of rich statistical regularity with other perceptual/motor cues provided in game play. Moreover, the categorization behavior of participants who played the game was demonstrably different from that of participants who merely classified the stimuli in an unsupervised categorization task. We interpret these results as evidence of human adults' general capacity to make use of informative statistical regularities in the input in interactive, functionally oriented situations and suggest that understanding more about the manner by which listeners discover sound structure in the environment can instruct us about the general learning mechanisms that may be brought to bear on phonetic categorization.

Some caution might be warranted in making strong claims regarding phonetic categorization or multimodal learning based on the present results, however. First, there is the possibility that the onset categories 1 and 2 in experiment 1, condition 1 did in fact possess invariant acoustic cues. As shown in Fig. 3, the initial energy distributions of category 1 exemplars did involve a P2 center frequency region (roughly 2.06–2.67 kHz over the first 25 ms) that was uniformly higher than that of category 2 exemplars during the same time frame (1.23–1.84 kHz). Listeners, then, could conceivably have identified categories based on the spectral properties of this distinctive region alone, disregarding the remainder of the onset patterns. Indeed, it has been similarly proposed in speech that the overall spectral shape of voiced consonants' initial bursts, rather than the non-invariant following formant transitions, are responsible for their context-independent recognition (Blumstein and Stevens, 1980). Dynamic formant patterns, though, have long been considered a salient acoustic property of voiced stops (e.g., Cooper *et al.*, 1952; Delattre *et al.*, 1955; Liberman *et al.*, 1954). In fact, language users have a persistent tendency to label consonants primarily based on formant transitions even after extensive training emphasizing instead the role of burst spectra (Francis *et al.*, 2000). It is especially unlikely that listeners in the present experiment classified category exemplars based only on their initial spectra, since unlike consonants they did not begin with acoustically prominent burst patterns but instead with brief onset ramps. It seems reasonable, therefore, to characterize categories 1 and 2 as having a "lack of invariance" of the same type that phonetic categories exhibit. Nonetheless, additional study in which the initial locations of similar dynamic spectral peaks are further removed from category-specific loci would help to clarify this issue. It would also be informative to examine the effects of differing types and degrees of "unstructured" variability like that employed in experiment 1, condition 2. Onset-trajectory–steady-state correspondences for the present study were a single pseudo-random distribution; further manipulation of these correspondences—perhaps involving their variation independent of absolute initial P2 ranges—could help to uncover more precisely the dependence of categorization on higher-order acoustic structure.

Another cause for caution in interpreting the results of experiment 1 is the overall level of categorization accuracy observed at posttest. As shown in Fig. 6, some individual participants performed quite well; overall, however, like the Japanese Quail in Kluender *et al.*'s (1987) phonetic category learning study and the human listeners in Lotto's (2000) complex nonspeech learning study, even condition 1 participants demonstrated far-from-perfect recognition. Judging by the results of the present study, this deficit seems more likely a methodological issue than a fundamental learning constraint. That is, it is expected that achieving expert-level performance may require more than a single 30-min learning session. In experiment 1, none of the participants approached the level of performance (score, level, etc.) that the experimenter reached after several hours of game play. This room for improvement indicates first of all that the game is well suited for longer-term studies. Furthermore, the correlation between game performance and posttest accuracy illustrated in Fig. 6 suggests that as skill level at the game continues to increase, so too will category knowledge. Pilot data suggest that listeners may begin to recognize categories like those described above at near 100% accuracy after as little as an hour of game play. Further study, perhaps involving hours of

game play over multiple sessions, will help in further characterizing the learnability of non-invariant sound classes like those explored here.

Finally, care must be taken in interpreting the correlation between game performance and learning as evidence that it was precisely the interactive nature of the task that helped drive learning. While learning sound classes had clear consequences for game performance, further study will be necessary to determine how truly interactive the process was. Training conditions, probably also over longer training periods, in which various potentially informative aspects of the Irfbats design (e.g., character-response, character-direction, and category-character contingencies) are altered or withheld will have to be compared to address this issue and to evaluate the relative importance of each type of information.

In sum, the Irfbats game paradigm appears to capture characteristics of auditory category learning that may be essential to learning natural sound categories, including phonetic categories. Moreover, it provides a learning environment in which to empirically manipulate experience to mechanistically address the bases of complex auditory categorization.

ACKNOWLEDGMENTS

This work was supported by a James S. McDonnell Foundation award for Bridging Mind, Brain, and Behavior to LLH, NIH Grant No. 5 RO1 DC04674-02 to LLH, a grant from the Bank of Sweden Tercentenary Foundation, and by a fellowship from the NIH Postdoctoral Training Grant on "Individual Differences in Cognition." The authors thank Christi Adams, Sara Kapner, Kelly Bunch, and Brian Mathias for help in conducting the experiments.

[1]http://www.psy.cmu.edu/~lholt/php/gallery_irfbats.php

[2]The 25-ms onset-offset amplitude ramps may have caused endpoint percepts to be somewhat further from the locus than illustrated in Fig. 3.

Allen, S. W. and Brooks, L. R. (**1991**). "Specializing the operation of an explicit rule," J. Exp. Psychol. Gen. **120**, 3–19.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (**1998**). "A neuropsychological theory of multiple systems in category learning," Psychol. Rev. **105**, 442–481.

Ashby, F. G., Queller, S., and Berretty, P. M. (**1999**). "On the dominance of unidimensional rules in unsupervised categorization," Percept. Psychophys. **61**, 1178–1199.

Ashby, F. G., Maddox, W. T., and Bohl, C. J. (**2002**). "Observational versus feedback training in rule-based and information-integration category learning," Mem. Cognit. **30**, 666–677.

Ashby, F. G., and Gott, R. E. (**1988**). "Decision rules in the perception and categorization of multidimensional stimuli," J. Exp. Psychol. Learn. Mem. Cogn. **14**, 33–53.

Ashby, F. G., and Maddox, W. T. (**1990**). "Integrating information from separable psychological dimensions," J. Exp. Psychol. Hum. Percept. Perform. **16**, 598–612.

Baptista, L. F., and Petrinovich, L. (**1986**). "Song development in the white-crowned sparrow: social factors and sex differences," Anim. Behav. **34**, 1359–1371.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Bruner, J. (**1983**). *Child's Talk; Learning to Use Language* (Norton, New York).

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1964**). "Formant transitions and loci as acoustic correlates of place of articulation in American fricatives," Studia Lingusti. **18**, 104–121.

Diehl, R., Lotto, A., and Holt, L. L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Eales, L. A. (**1989**). "The influences of visual and vocal interaction on song learning in zebra finches," Anim. Behav. **37**, 507–508.

Eimas, P. D. (**1963**). "The relation between identification and discrimination along speech and nonspeech continua," Lang Speech **6**, 206–217.

Fernald, A., and Simon, T. (**1984**). "Expanded intonation contours in mothers' speech to newborns," Dev. Psychol. **20**, 104–113.

Fiser, J., and Aslin, R. N. (**2002**). "Statistical learning of higher-order temporal structure from visual shape sequences," J. Exp. Psychol. Learn. Mem. Cogn. **28**, 458–467.

Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective," J. Phonetics **14**, 3–28.

Francis, A. L., Baldwin, K., and Nusbaum, H. C. (**2000**). "Effects of training on attention to acoustic cues," Percept. Psychophys. **62**, 1668–1680.

Gauthier, I., and Tarr, M. J. (**1997**). "Becoming a 'Greeble' expert: exploring mechanisms for face recognition," Vision Res. **37**, 1673–1682.

Gauthier, I., Behrmann, M., and Tarr, M. J. (**1999a**). "Can face recognition really be dissociated from object recognition?" J. Cogn Neurosci. **11**, 349–370.

Gauthier, I., Williams, P., Tarr, M. J., and Tanaka, J. (**1998**). "Training 'Greeble' experts: A framework for studying expert object recognition processes," Vision Res. **38**, 2401–2428.

Gauthier, I., Tarr, M., Anderson, A., Skudlarski, P., and Gore, J. (**1999b**). "Activation of the middle fusiform face area increases with expertise in recognizing novel objects," Nat. Neurosci. **2**, 568–573.

Guenther, F. H., Husain, F. T., Cohen, M. A., and Shinn-Cunningham, B. G. (**1999**). "Effects of categorization and discrimination training on auditory perceptual space," J. Acoust. Soc. Am. **106**, 2900–2912.

Hauser, M. D., Agnetta, B., and Perez, C. (**1998**). "Orienting asymmetries in rhesus monkeys: The effect of time-domain changes on acoustic perception," Anim. Behav. **56**, 41–47.

Hauser, M. D., Newport, E. L., and Aslin, R. N. (**2000**). "Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins," Cognition **75**, 1–12.

Hillenbrand, J. M., and Gayvert, R. T. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res. **48**, 45–50.

Holt, L. L., Kluender, K. R., and Lotto, A. J. (**1997**). "Discrimination of single-formant stimuli by chinchillas (*Chinchilla villidera*)," J. Acoust. Soc. Am. **102**, 3188.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2001**). "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?" J. Acoust. Soc. Am. **109**, 764–774.

Holt, L. L., Lotto, A., and Diehl, R. (**2004**). "Auditory discontinuities interact with categorization: implications for speech perception," J. Acoust. Soc. Am. **116**, 1763–1773.

Jusczyk, P. W. (**1997a**). "Finding and remembering words: Some beginnings by English-learning infants," Curr. Dir. Psychol. Sci. **6**, 170–174.

Jusczyk, P. W. (**1997b**). *The Discovery of Spoken Language* (MIT, Cambridge, MA).

Keating, P. A. (**1984**). "Phonetic and phonological representation of stop consonant voicing," Language **60**, 286–319.

Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (**2002**). "Visual statistical learning in infancy: Evidence for a domain general learning mechanism," Cognition **83**, B35–B42.

Kluender, K. R., Diehl, R. L., and Killeen, P. R. (**1987**). "Japanese quail can learn phonetic categories," Science **237**, 1195–1197.

Kluender, K. R., Lotto, A. J., Holt, L. L., and Bloedel, S. L. (**1998**). "Role of experience for language-specific functional mappings of vowel sounds," J. Acoust. Soc. Am. **104**, 3568–3582.

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

Kuhl, P. K. (**1992**). "Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes," in *Developmental Psychoacoustics*, edited by L. A. Werner and E. W. Rubel (American Psychological Association, Washington, DC), pp. 293–332.

Kuhl, P. K., and Miller, J. D. (**1975**). "Speech perception by the chinchilla:

Voiced-voiceless distinction in alveolar plosive consonants," Science **190**, 69–72.
Kuhl, P. K., Tsao, F.-M., and Liu, H., -M. (**2003**). "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning," Proc. Natl. Acad. Sci. U.S.A. **100**, 9096–9101.
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (**1992**). "Linguistic experience alters phonetic perception in infants by six-months of age," Science **255**, 606–608.
Lacerda, F. (**2003**). "Phonology: An emergent consequence of memory constraints and sensory input," Read. and Writ.: An Interdiscip. J. **16**, 41–59.
Lacerda, F., and Sundberg, U. (**2004**). "An ecological theory of language learning," J. Acoust. Soc. Am. **116**, 2523.
Liberman, A., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (**1954**). "The role of consonant-vowel transitions in the stop and nasal consonants," Psychol. Monogr. **68**, 1–13.
Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.
Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.
Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.
Lindblom, B., Brownlee, B. D., and Moon, S.-J. (**1992**). "Speech transforms," Speech Commun. **11**, 357–368.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-linguistic study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.
Lotto, A. J. (**2000**). "Language acquisition as complex category formation," Phonetica **57**, 189–196.
Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**, 602–619.
Lotto, A. J., Kluender, K. R., and Holt, L. L. (**1998**). "Depolarizing the perceptual magnet effect," J. Acoust. Soc. Am. **103**, 3648–3655.
Massaro, D. W. (**1987**). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Erlbaum, Hillsdale, NJ).
Maye, J., Werker, J. F., and Gerken, L. (**2002**). "Infant sensitivity to distributional information can affect phonetic discrimination," Cognition **82**, B101–B111.
Mirman, D., Holt, L. L., and McClelland, J. L. (**2004**). "Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues," J. Acoust. Soc. Am. **116**, 1198–1207.
Nearey, T. M. (**1997**). "Speech perception as pattern recognition," J. Acoust. Soc. Am. **101**, 3241–3254.
Papousek, M., Papousek, H., and Bornstein, M. H. (**1985**). "The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech," in *Social Perception in Infants*, edited by T. M. Field and N, Fox (Ablex, Norwood, NJ), pp. 269–295.
Pisoni, D. B. (**1977**). "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," J. Acoust. Soc. Am. **61**, 1352–1361.
Pisoni, D. B. (**1987**). "Auditory perception of complex sounds: Some comparisons of speech vs nonspeech signals," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ), pp. 247–256.
Reber, A. S. (**1976**). "Implicit learning of synthetic languages: The role of instructional set," J. Exp. Psychol. [Hum Learn] **2**, 88–94.
Reber, A. S. (**1989**). "Implicit learning and tacit knowledge," J. Exp. Psychol. **118**, 219–235.
Reber, A. S., Kassin, S. M., Lewis, S., and Cantor, G. W. (**1980**). "On the relationship between implicit and explicit modes in the learning of a complex rule structure," J. Exp. Psychol. [Hum Learn] **6**, 492–502.
Rossion, B., Gauthier, I., Goffaux, V., Tarr, M., and Crommelinck, M. (**2002**). "Expertise training with novel objects leads to left-lateralized face-like electrophysiological responses," Psychol. Sci. **13**, 250–257.
Saffran, J. R., Aslin, R. N., and Newport, E. L. (**1996**). "Statistical learning by 8-month-old infants," Science **274**, 1926–1928.
Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (**1999**). "Statistical learning of tone sequences by human infants and adults," Cognition **70**, 27–52.
Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (**1997**). "Incidental language learning: Listening (and Learning) out of the corner of your ear," Psychol. Sci. **8**, 101–105.
Sinnott, J. M., and Brown, C. H. (**1997**). "Perception of the American English liquid /ra-la/ contrast by humans and monkeys," J. Acoust. Soc. Am. **102**, 588-602.
Sinnott, J. M., Beecher, M. D., Moody, D. B., and Stebbins, W. C. (**1976**). "Speech sound discrimination by monkeys and humans," J. Acoust. Soc. Am. **60**, 687-695.
Stevens, K. N. and Blumstein, S. E. (**1981**). "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the Study of Speech*, edited by P. D. E. J. L. Miller (Erlbaum Hillsdale, NJ), pp. 1–38.
Sussman, H. M., Hoemeke, K. A., and Ahmed, F. S. (**1993**). "A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation," J. Acoust. Soc. Am. **94**, 1256–1268.
Sussman, H. M., Fruchter, D., Hilbert, J., and Sirosh, J. (**1998**). "Linear correlates in the speech signal: The orderly output constraint," Behav. Brain Sci. **21**, 241–299.
Trout, J. D. (**2001**). "The biological basis of speech: What to infer from talking to the animals," Psychol. Rev. **108**, 523–549.