PNAS
www.pnas.org

Supplementary Information for

The role of the striatum in incidental learning of sound categories

Sung-Joo Lim, Julie A. Fiez, Lori L. Holt

Sung-Joo Lim
E-mail: sungj.m.lim@gmail.com

This PDF file includes:

Supplementary text
Figs. S1 to S6
Tables S1 to S4
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1811992116

Supporting Information Text

Methods

Localizer task. We used a similar approach used by Leech et al. (1) to localize speech-selective responses in the l-STS, where they found learning-related changes after incidental sound category learning in the videogame. To ensure that participants remained alert during the task, we intermixed visuo-motor and auditory trials.

On each auditory trial, participants passively heard either a natural speech or nonspeech sound while seeing a fixation cross. The natural speech sound trials consisted of four repetitions of six English words and 24 acoustically-varying English syllables recorded by a female talker. For the nonspeech trials, participants heard four repetitions of six environmental sounds semantically-matched to the six English words and four repetitions of six randomly-chosen nonspeech sound exemplars used in the game training (Fig. 1). The English words and semantically-matched environmental sounds were selected from the localizer used by Leech et al. (1) to identify speech-selective regions. On visuo-motor trials (72 trials total), participants saw one of four distinct aliens or Gabor patches. Based on the color of the visual stimulus regardless of the type (i.e., alien or patch) participants pressed assigned buttons using the left or right index finger. Each trial was 1.5 s in duration, separated by varying intervals from 0 to 6 s. Presentation was controlled by E-Prime v2.0 (2).

Videogame training. We developed a variant of the Wade and Holt (3) videogame training compatible with in-scanner game play (Java 6.0 SDK) to examine online incidental auditory category learning using fMRI. Players navigated the game environment in order to orient toward the approaching alien using the right-hand number pad (1=LEFT, 3=RIGHT, 5=UP, 2=DOWN), and took appropriate game actions using button keys on the left-hand response glove. When action preparation keys (i.e., the index or third finger keys for either shooting or capturing) were pressed and held, a targeting graphic appeared in the center of the screen to aim at the alien (Fig. 1B). Pressing the left fourth finger key executed the game action determined by the preparation key. The mapping between the preparation keys and game action was counterbalanced across participants.

The game advanced by one level each time players successfully completed a predetermined number of game actions. As the game advanced to higher levels, the time window to take game actions was reduced; as aliens moved at a faster speed, participants must execute appropriate game actions more rapidly. When a participant failed to shoot/capture an alien five times (i.e., trials), a new round of the game automatically restarted, with the difficulty level adjusted based on the previous round. This ensured that participants continuously played the game in each functional run. Thus, participants went through multiple rounds of game training in the scanner and game difficulty was adjusted based on each participant's own performance.

Functional image acquisition and preprocessing. Prior to the functional scans, anatomical images were acquired using T1-weighted MP-RAGE sequence (repetition time (TR) = 1540 ms, echo time (TE) = 3.04 ms, field of view (FOV) = 256 mm, flip angle (FA) = 8 degrees) with 1 × 1 × 1 mm resolution, and T2-weighted structural images were collected (38 slices, TR = 6000 ms, TE = 73 ms, FOV = 200 mm, FA = 150 degrees). Whole-brain, T2*-weighted echo-planar images (EPI) sensitive to the blood-oxygen-level-dependent (BOLD) contrast were collected (TR = 1500 ms, TE = 25ms, FOV = 200 mm, FA = 60 degrees). Each volume had 29 axial slices (3.125 × 3.125 × 3.5 mm resolution) parallel to the anterior commissure–posterior commissure line. The localizer task acquired 321 volumes. The videogame training was conducted across three functional runs and each run acquired 480 volumes. The first two volumes of each run were discarded prior to analysis. Preprocessing and analysis were done using the Analysis of Functional NeuroImages (AFNI) (4). Preprocessing steps included slice-time correction, spatial realignment to the first volume, and normalization to Talairach space (5). Images were spatially smoothed with a 5.5 mm full-width at half maximum Gaussian kernel, and the signal for each voxel was scaled to a global mean of 100. The observed BOLD signal for each task (i.e., localizer and game training) was separately modeled by a general linear model (GLM). The baseline activity was modeled by trends from linear up to higher-degree polynomial terms to remove the slow signal drift (i.e., high-pass filter with an approximate cutoff of 240 s). The six additional continuous motion parameters were entered as covariates of no interest.

It is of note that due to a technical problem with our videogame training and log file recording, videogame trial event information (e.g., feedback delivery) was missing at random for some subjects. In such cases, we removed the functional volumes that correspond to these trials with missing game events. On average, the loss of functional volumes was 1.45%. We confirmed that this data loss was not different in the two participant groups ($t_{25} = 0.68$, $P = 0.50$), and that the rate of data loss across the whole sample or within each group was not related to the BOLD activation or striatal connectivity (all $|r|$s $<$ 0.32, Ps > 0.29).

Localization of visual and motor ROIs. Using the same approach as in localizing speech-selective ROIs, we also localized additional visual and motor ROIs for each individual participant in order to explore the extent of striatal connectivity to these regions during the videogame training. For defining visual ROIs, we used the aliens > patch contrast; for localizing motor ROIs, we selected both positive and negative peak voxels in contrasting right- vs. left-hand motor responses for each individual participant, within the group-defined mask of the corresponding contrasts for the visual and motor ROIs (at voxel-wise $P <$ 0.05; cluster-corrected threshold of 263 voxels). Around each individual's peak activation voxel, we created an ROI using a 2.5 voxel radius sphere (~85 voxels).

Videogame task GLM. The GLM also had additional regressors of no interest. The response to gaming action motor-related events was estimated using a total of six regressors. Each separately modeled the response for each hand and for response

durations. The response duration and key-press behaviors varied considerably (e.g., continuous key press or discrete key strokes) because participants freely made motor responses in the videogame without any limits or restrictions. Thus, participants' motor responses were binned based on the response duration (1 TR, 2 TR, ≥3TR) for each hand to separately model the BOLD time series of brief, intermediate, and long motor responses. These regressors respectively modeled the BOLD response across 15 s, 16.5 s, and 18 s from the motor response onset. The hemodynamic response to feedback was estimated using two separate regressors, one for correct and the other for incorrect feedback. These regressors modeled the BOLD response across 15 s from the onset of feedback presentation.

Behavioral analysis. We examined whether the two groups (experimental vs. control) exhibited different category generalization from the videogame training. To this end, we analyzed trial-wise categorization accuracy using a logistic mixed-effects model implemented in *lme4* in R (v3.3.3). For each sound category type (offset vs. onset), we ran a separate model that included fixed factors of Group (experimental vs. control) and Sound Exemplar type (trained vs. novel), with participants as a random factor (also see Table S1 for the result of a full model testing all three factors). Any significant effects found in the model were followed up by post-hoc testing using differences of least-square means (*lsmeans* in R). Furthermore, we also evaluated whether the videogame training was effective in inducing category learning of the sounds, by using separate one-sample t-tests against chance level of performance (i.e., 0.25) for each participant group. Lastly, in order to evaluate whether the videogame training promotes auditory category learning (3), we performed a correlation analysis between participants' game performances and post-test category generalization performances for both category types.

Results

Further analysis of behavioral response patterns. Because there were four sound categories (two onset- and two offset-sweep categories), the chance categorization accuracy is at 0.25. However, it is of note that the two coarse category types (i.e., onset- vs. offset-sweeps) can be readily distinguished from each other based on differences in carrier frequencies and the timing of the frequency sweep (Fig. 1) without training. Thus, although participants in the control group experienced two sets of randomly sampled onset-sweep category exemplars, they should be capable of achieving above-chance categorization accuracy (i.e., >0.25) in novel generalization by coarsely distinguishing onset- vs. offset-sweep categories, even without making fine-grained distinctions within the onset- and offset-sweep category types.

Here, we examined whether the control group's novel generalization performance and in-scanner game performance can be explained by this ability to coarsely distinguish onset- vs. offset-sweep categories, rather than fine-grained distinction of categories within the onset and offset-sweep sound distributions. Based on the categorization response patterns to novel sound exemplars (Fig. S1A), we quantified the proportions of errors in distinguishing the coarse, onset vs. offset categories (i.e., between-category confusion errors) and the errors in distinguishing the two categories within each of the onset and offset category types (i.e., within-category confusion errors).

We found that the control group participants coarsely discriminated onset- vs. offset-sweep categories from each other, but exhibited high confusion in distinguishing between the two onset categories (Fig. S1A, right). As predicted, the control group's generalization performance as well as the in-scanner game performance can only be reliably explained by the between-category confusion error rates, but not by the within-category confusion errors (Fig. S1B–C, right). This pattern indicates that the control group's sound category learning of onset-sweeps is limited to coarse distinctions the onset-sweep versus offset-sweep categories.

On the contrary, the experimental group participants exhibited reliable categorization of all four categories (Fig. S1A, left). We found that the experimental group's generalization performance as well as the in-scanner game performance can be reliably explained by the proportions of response errors derived from between-category confusion and within-category confusion (Fig. S1B–C, left). This pattern indicates that the experimental group can discriminate onset- vs. offset-sweep categories from each other as well as the two categories within each category type.

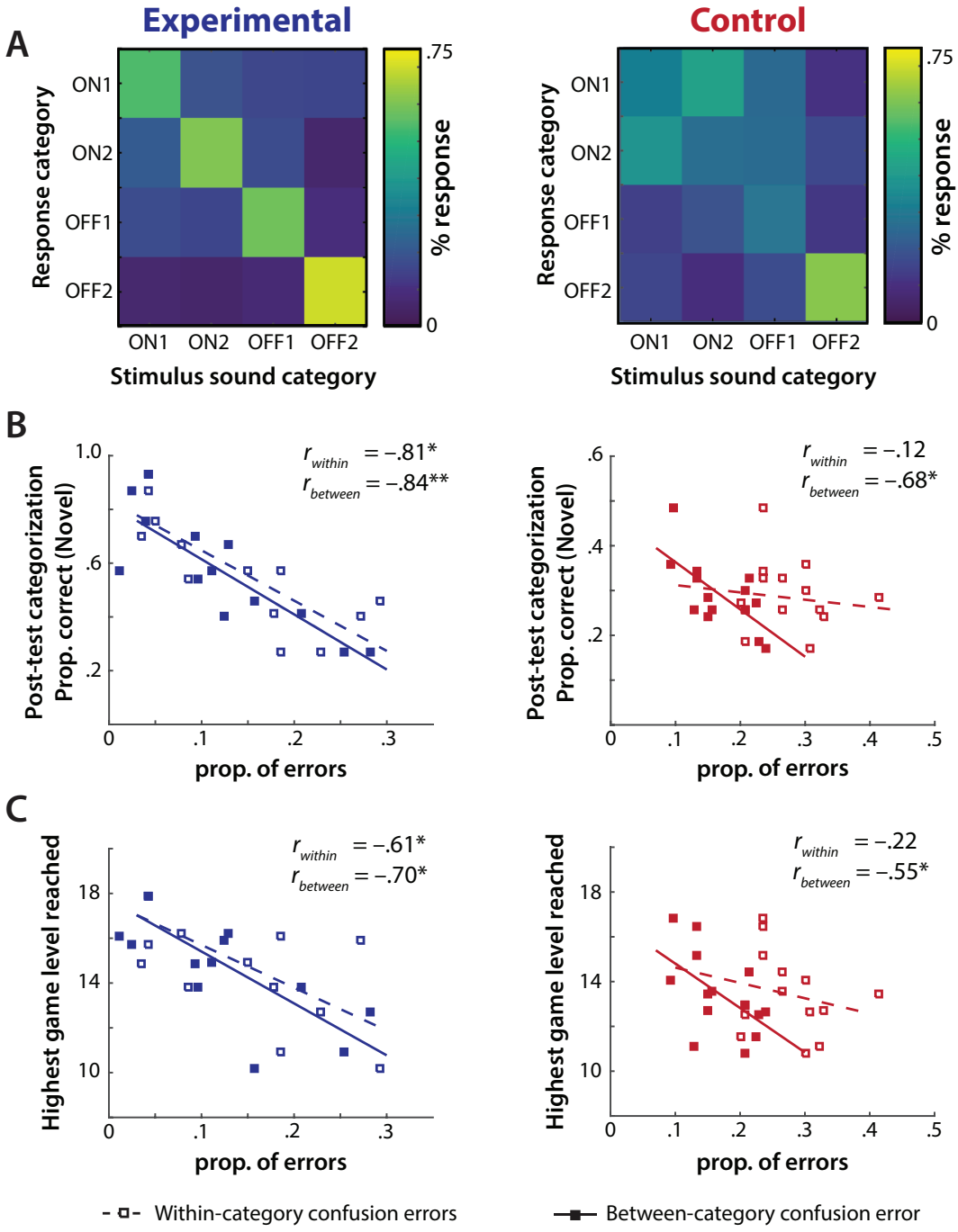

Fig. S1. Response patterns in categorizing novel generalization stimuli and the relationships between the two types of categorization errors to behavioral performances. (A) Average response pattern for categorizing novel generalization exemplars of the onset- and offset-sweep sound categories across participants in the experimental and control groups. The diagonal indicates the average proportion of correct categorization responses of each category. ON1 and ON2 indicate two onset-sweep categories; OFF1 and OFF2 denote two offset-sweep categories. (B) The relationship of the response error rates to the post-test generalization of category learning to novel onset-sweep category exemplars. Open and filled squares indicate individuals' proportions of errors in distinguishing within- and between- categories, respectively. The dashed and solid lines denote the linear fits of the within- and between- categories errors to the respective behavioral performance measures. Pearson's *r*-values show the relationships. (C) The relationship of the response error rates to in-scanner behavioral videogame performances is shown using the same illustration scheme in (B). $^{*}P < 0.05$, $^{**}P < 0.005$.

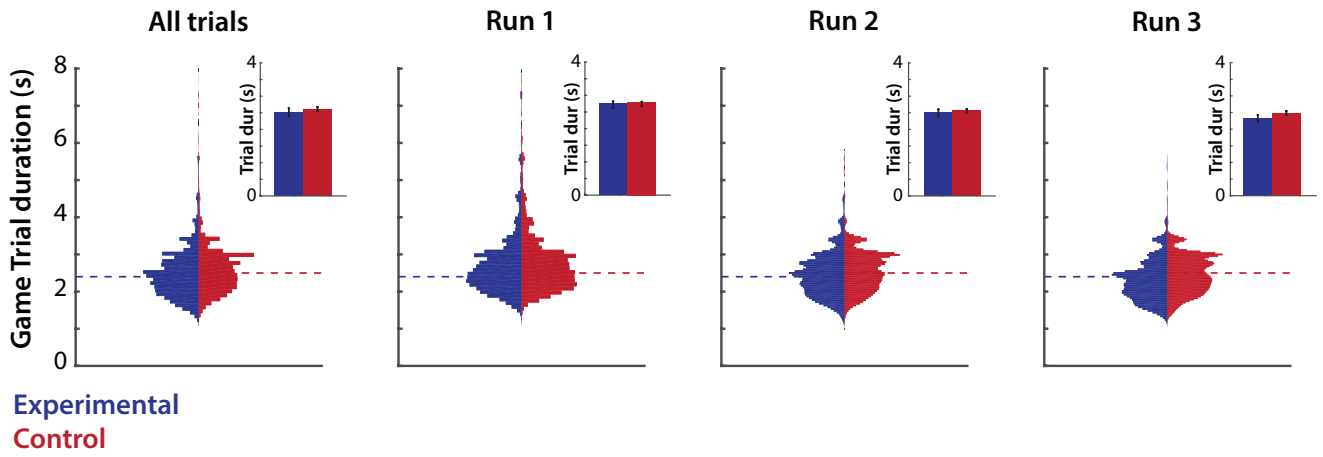

Fig. S2. Distributions of the game trial duration. Pooled distributions of all trials from all participants in experimental (blue) and control (red) groups. All trials across the three functional game training runs (left) and all trials within each functional run (three runs total) are illustrated. Dotted lines on each panel indicate the median game trial duration of each group. Inset bar graphs illustrate the mean duration of game trials across participants in each group. Error bars show ±1 SEM. There were no group differences in the game trial duration (independent samples t-tests; all ts < 1.46, all Ps > 0.157).

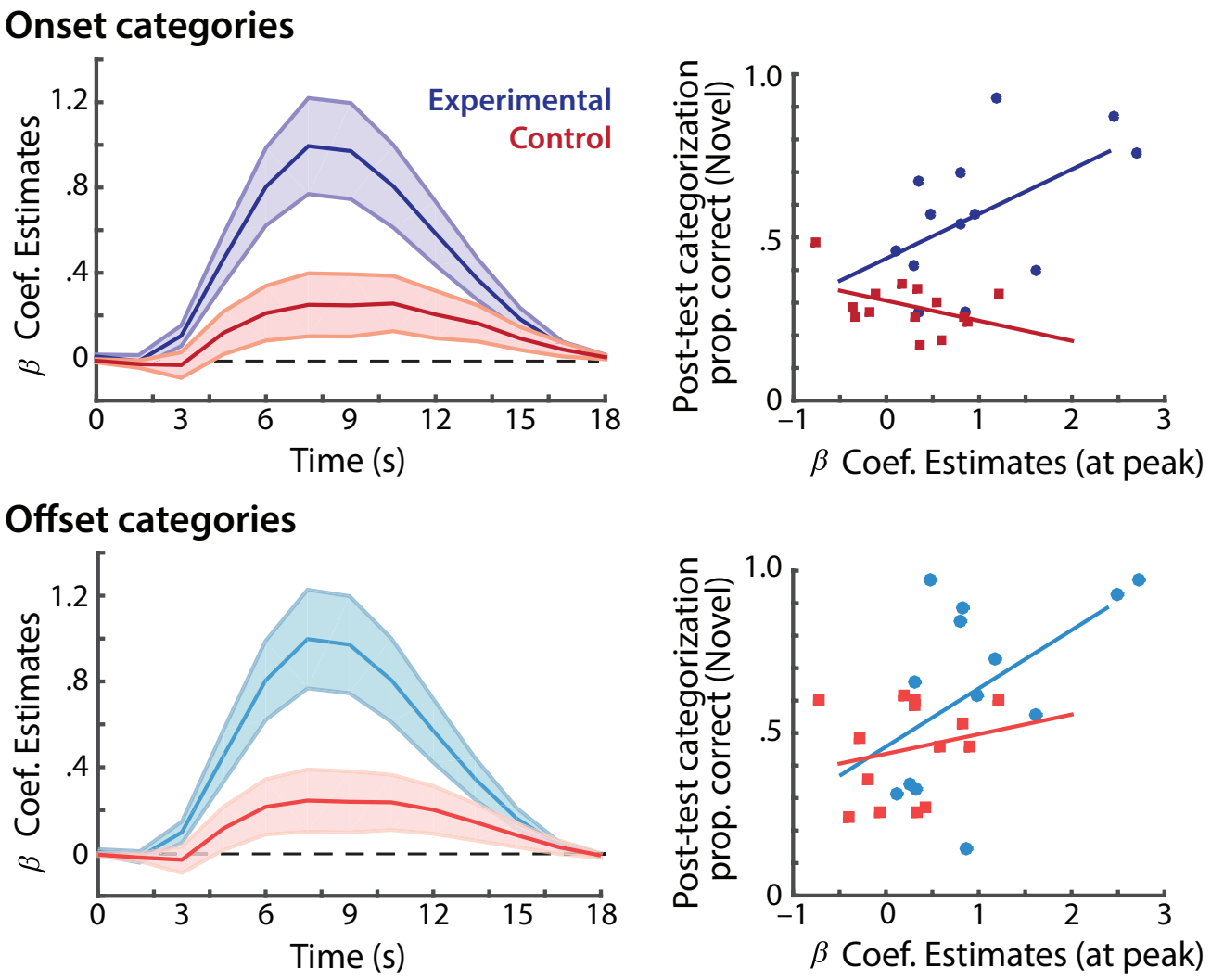

Fig. S3. Activation of the striatum (Fig. 3) in response to onset vs. offset sound category trials during the videogame training (left) and its relationship to behavioral generalization performance for categorizing novel exemplars of the corresponding sound category type (right). BOLD responses to the onset and offset category trials in the videogame training are separately modeled in the GLM. Shaded error bars show ±1 SEM.

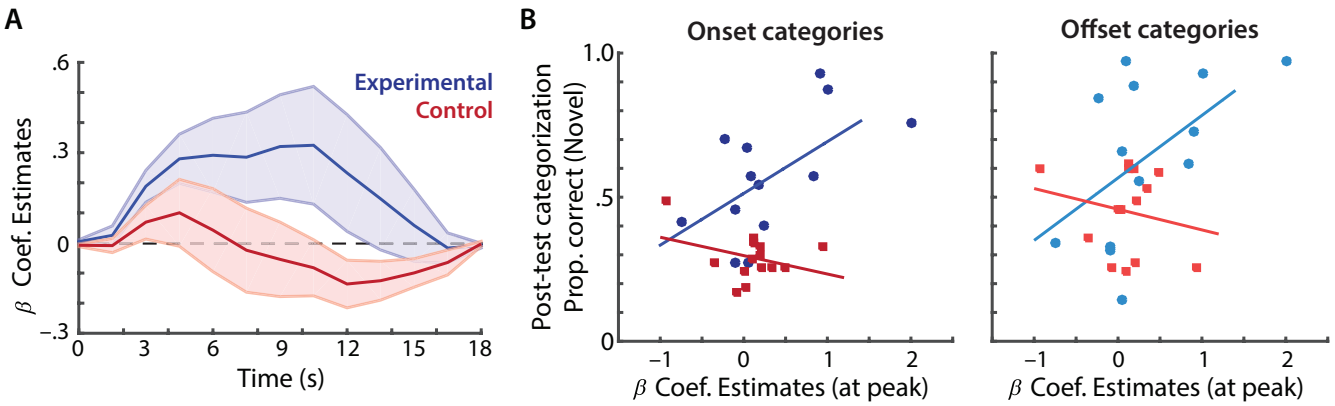

Fig. S4. The BOLD activation in the localized speech-selective left STS areas during the videogame training and its relationship to behavioral categorization performances. (A) The BOLD activation time course of speech-selective l-STS ROIs in response to the onset and offset sound category exemplars during the videogame training. The speech-selective l-STS ROIs were localized based on the voxel-wise speech > non-speech contrast results for individual participants (Fig. 5, left). Shaded error bars indicate ±1 SEM. (B) Relationship between category generalization performances and the l-STS recruitment (at peak) during the videogame training for each sound category type. A multiple linear regression analysis revealed a significant to trend-level Group × l-STS Activation effect on predicting category generalization (onset: β = 0.57, t = 2.12, P = 0.046; offset: β = 0.61, t =1.88, P = 0.073). Exploratory post-hoc analyses revealed significant/marginal positive relationships between the l-STS activation and behavioral performance in the experimental group (onset: r = .60, P = 0.029; offset: r = 0.54, P = 0.056), but not in the control group (all $|r|$s < 0.34, all Ps > 0.235).

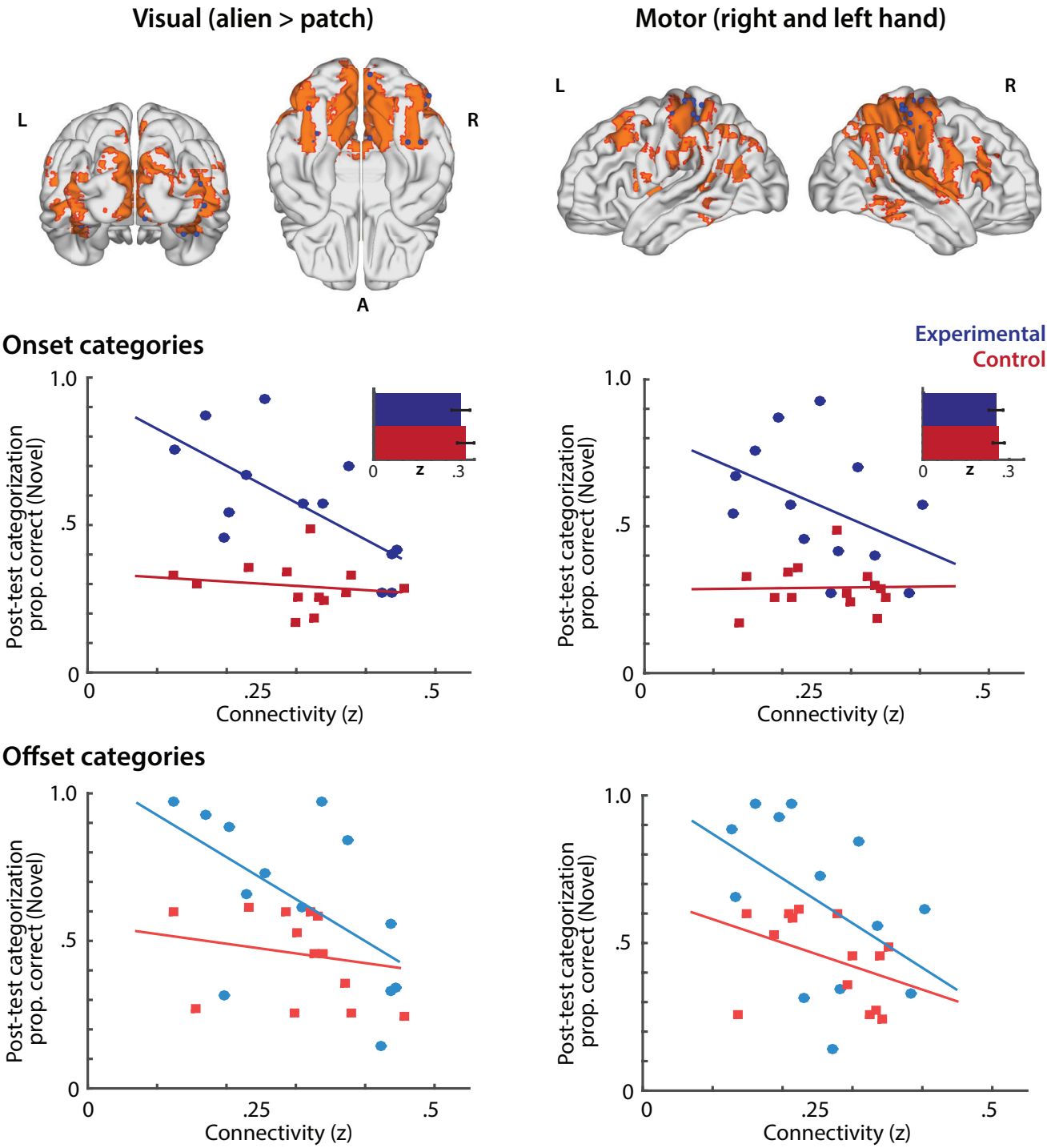

Fig. S5. The findings from an exploratory analysis on striatal connectivity patterns to the visual and motor (right- and left-hand response) localized regions (top), and their relationship to behavioral performance for categorizing novel onset and offset category exemplars in the post-test (middle and bottom panels, respectively). Striatal connectivity was extracted from individually defined visual and motor ROIs (blue spheres) using the same approach as in localizing speech-selective l-STS ROIs (SI Methods). For both visual and motor ROIs, multiple linear regression analyses revealed non-significant group differences in predicting the offset sound category learning with striatal connectivity (Group x Striatal Connectivity: ts < 1.54, Ps > 0.137). For predicting onset category generalization, only the visual ROI exhibited a significant effect of Group × Striatal Connectivity: $t = 2.48$, $P = 0.021$).

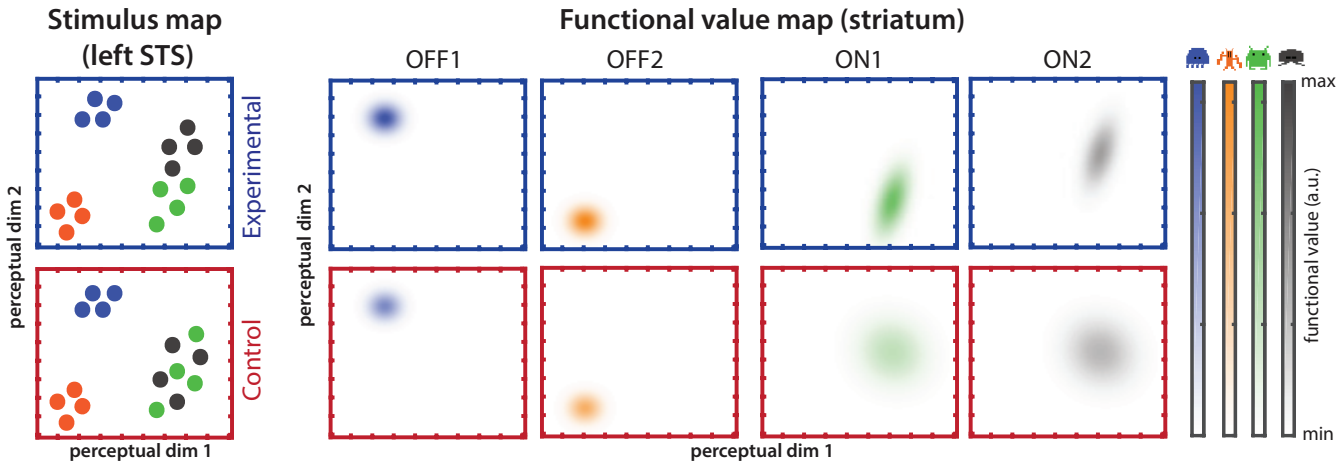

Fig. S6. Schematic illustrations of the proposed sound category learning driven by the incidental videogame training. Sound category exemplars are represented in a perceptual space in the auditory l-STS area based on their acoustic features (left; see (6)). The striatum learns value-based stimulus representations (i.e., functional utility of the sound information in the game). The striatum accrues information on each game trial for all listeners. However, when the stimuli comprise a coherent category space, the experimental group builds accumulated knowledge about the sound category-level information that can be generalized across exemplars within each category. On the contrary for the control group, knowledge about each onset sound exemplar does not accumulate in a statistically coherent manner. Thus, although the striatum learns functional value of the onset sounds in the control group condition, the knowledge is not as useful in guiding behavior on subsequent trials in the videogame, or generalizing to novel exemplars.

Table S1. Mixed-effects logistic modeling results of the behavioral post-test categorization accuracy

Model	Predictor	*df*	χ^2	*P*
Full model	Group	1	11.73	**0.00062**
	Sound category type	1	77.15	**2.20×10^{-16}**
	Exemplar type	1	25.36	**4.76×10^{-7}**
	Group × Sound category type	1	2.12	0.15
	Group × Exemplar type	1	3.16	0.076
	Sound category type × Exemplar type	1	1.16	0.26
	Group × Sound category type × Exemplar type	1	6.61	**0.010**
Offset sounds	Group	1	6.81	**0.0091**
	Exemplar type	1	8.77	**0.0031**
	Group × Exemplar type	1	0.47	0.49
Onset sounds	Group	1	17.72	**2.57×10^{-5}**
	Exemplar type	1	17.74	**2.82×10^{-5}**
	Group × Exemplar type	1	9.73	**0.0018**

Note: Significant effects are highlighted in bold.

Table S2. Striatal clusters exhibiting Group × Time course across three videogame training runs

	Regions	Talairach Coordinates			*F*-value	Cluster size, voxels
		x	y	z		
Run 1	L Caudate body	–10	2	17	3.91	11
	L Putamen	–22	–7	11	5.21	7
	L Putamen	–22	2	–4	3.70	6
	R Putamen	26	8	–4	6.66	33
Run 2	L Caudate body	–10	2	17	4.21	12
	L Putamen	–28	–13	–4	4.90	12
	L Putamen	–19	14	–7	5.00	9
	L Putamen	–22	–7	11	4.00	7
	R Putamen	26	8	–4	6.80	40
Run 3	L Caudate body	–10	2	17	4.32	10
	L Putamen	–19	14	–7	5.72	10
	L Putamen	–22	–7	11	3.73	6
	R Putamen	26	8	–4	5.97	38

Note: voxel-wise $P \leq 0.005$ corrected at alpha=0.05 with cluster threshold of six voxels. L, left; R, right.

Table S3. Brain areas exhibiting Group × Time course effect collapsed across all functional runs from a whole-brain analysis

Regions	BA	Talairach Coordinates			*F*-value	Cluster size, voxels
		x	y	z		
R Precentral gyrus/MFG		44	–2	48	9.89	1535
L Medial FG	6	–7	2	57	6.08	
R Cingulate gyrus		17	–25	36	5.71	
R Cingulate gyrus	24	11	–4	48	5.00	
R IFG		55	8	27	8.53	
L IFG		–32	17	–16	5.79	
R Insula		50	8	3	6.88	
L Insula		–38	–14	12	6.94	
R Caudate body		17	8	9	8.00	
L Caudate body		–11	–2	18	5.25	
R Putamen		26	8	–4	7.35	
L Putamen		–26	13	–3	5.64	
L Putamen		–28	-4	–6	5.03	
L Thalamus		–7	–14	–7	3.17	
L Parahippocampus	34	–17	–5	–18	5.79	
L Ant Cingulate/Insula		–14	22	24	7.37	162
R SFG		26	55	30	7.02	185
L SFG	9	–20	53	34	7.15	64
R Medial FG	10	5	56	–7	6.66	38
L Medial FG		–19	53	–7	4.58	23
L Medial FG	8	–7	38	38	7.14	34
R MFG/PFG	9	32	26	32	4.98	29
L Precentral gyrus	6	–41	–4	32	10.21	103
L IFG	10	–43	38	2	4.74	34
R MTG/ITG		38	–2	–34	5.71	44
R Parahippocampus gyrus		32	–26	–16	8.23	209
L Medial temporal/cerebellum		–13	–25	–19	5.81	35
L Medial temporal/cerebellum		–32	–8	–31	5.18	68
R Cerebellum		23	–40	–25	5.01	26
L Cerebellum		–28	–85	–25	6.41	72
R STG	22	50	–10	2	5.34	32
R STG/SMG		67	–43	–4	6.81	112
R SPL	7	17	–67	56	5.74	42
L SPL		–40	–49	62	4.17	24
R Intraparietal sulcus		29	–49	38	6.53	96
L IPL	40	–54	–49	50	7.48	91
L Precuneus	31	–7	–49	35	5.44	31
L Precuneus	19	–31	–64	41	5.62	101
R Cuneus/MOG		26	–94	2	7.51	46
L Cuneus	18	–1	–76	20	5.71	69
L Lingual gyrus	19	–7	–88	–13	3.88	29
L Post cingulate		–28	–61	11	8.24	92

Note: voxel-wise $P \leq 0.005$ corrected at alpha=0.05 with cluster threshold of 23 voxels. L, left; R, right; STG, superior temporal gyrus; SMG, supramarginal gyrus; MTG, middle temporal gyrus; ITG, inferior temporal gyrus; FG, frontal gyrus; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; SPL, superior parietal lobule; IPL, inferior parietal lobule; MOG, middle occipial gyrus; ant, anterior; post, posterior; BA, Brodmann area.

Table S4. Multiple linear regression results in predicting the generalization performance of each category type (onset and offset) assessed in the post-test with the peak striatal activation during the videogame training

Sound category type	Factors	β	t	P
Onset sounds	Group	0.26	1.35	0.19
	Striatal Activation	–0.17	0.75	0.46
	Group × Striatal Activation	0.70	2.37	**0.027**
Offset sounds	Group	0.048	0.18	0.86
	Striatal Activation	0.18	0.59	0.56
	Group × Striatal Activation	0.36	0.92	0.37

Note: Significant effects are highlighted in bold.

References

1. Leech R, Holt LL, Devlin JT, Dick F (2009) Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J Neurosci* 29(16):5234–5239.
2. Schneider W, Eschman A, Zuccolotto A (2002) *E-Prime: User's Guide*. (Psychology Software Inc., Pittsburgh, PA).
3. Wade T, Holt LL (2005) Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J Acoust Soc Am* 118(4):2618.
4. Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29(3):162–173.
5. Talairach J, Tournoux P (1988) *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. (Thieme Medical, New York).
6. Emberson LL, Liu R, Zevin JD (2013) Is statistical learning constrained by lower level perceptual organization? *Cognition* 128(1):82–102.