

ORIGINAL ARTICLE

The Role of the Human Auditory Corticostriatal Network in Speech Learning

Gangyi Feng ^{1,2}, Han Gyor Yi³ and Bharath Chandrasekaran⁴

¹Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China, ²Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China, ³Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94158, USA and

⁴Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Address correspondence to Bharath Chandrasekaran, Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA. Email: b.chandra@pitt.edu

Gangyi Feng and Han Gyor Yi contributed equally to this work

Abstract

We establish a mechanistic account of how the mature human brain functionally reorganizes to acquire and represent new speech sounds. Native speakers of English learned to categorize Mandarin lexical tone categories produced by multiple talkers using trial-by-trial feedback. We hypothesized that the corticostriatal system is a key intermediary in mediating temporal lobe plasticity and the acquisition of new speech categories in adulthood. We conducted a functional magnetic resonance imaging experiment in which participants underwent a sound-to-category mapping task. Diffusion tensor imaging data were collected, and probabilistic fiber tracking analysis was employed to assay the auditory corticostriatal pathways. Multivariate pattern analysis showed that talker-invariant novel tone category representations emerged in the left superior temporal gyrus (LSTG) within a few hundred training trials. Univariate analysis showed that the putamen, a subregion of the striatum, was sensitive to positive feedback in correctly categorized trials. With learning, functional coupling between the putamen and LSTG increased during error processing. Furthermore, fiber tractography demonstrated robust structural connectivity between the feedback-sensitive striatal regions and the LSTG regions that represent the newly learned tone categories. Our convergent findings highlight a critical role for the auditory corticostriatal circuitry in mediating the acquisition of new speech categories.

Key words: corticostriatal system, multi-modal imaging, MVPA, speech category learning

Introduction

A central computational task in speech perception is mapping inconstant acoustic signals into meaningful phonological categories (Diehl et al. 2004; Holt and Lotto 2008, 2010). Native speech sound categories are represented primarily along the bilateral superior temporal gyrus (STG) (Formisano et al. 2008; DeWitt and Rauschecker 2012; Bonte et al. 2014; Mesgarani et al. 2014; Arsenault and Buchsbaum 2015; Feng, Gan, et al. 2018). These neural representations emerge primarily via

unsupervised exposure to the statistical properties of one's native language, starting from prenatal stages and continuing to early childhood (Cheour et al. 1998; Nakahara et al. 2004; Vallabha et al. 2007; Garcia-Lazaro et al. 2011). In the mature brain, subregions of the STG reliably represent abstract category-level information despite the acoustic variability inherent in the raw signal, enabling effortless categorization of native speech sounds (Formisano et al. 2008; DeWitt and Rauschecker 2012; Mesgarani et al. 2014; Feng, Gan, et al. 2018).

The stable and robust representation of native speech sounds may come at the cost of reduced sensitivity to acoustic dimensions that bear critical linguistically relevant information in other languages (Kuhl et al. 2003). Although previous studies have suggested that speech category learning in adulthood engages the STG (Callan et al. 2003; Wang et al. 2003; Desai et al. 2008; Zhang et al. 2009), little is known about mechanistic details underlying the emergence of novel representations in the STG as a function of training (Ley et al. 2012; Karuza et al. 2014; Myers 2014).

We hypothesize that the striatum is a prime candidate for the source of training-induced neuroplasticity in the STG via bi-directional corticostriatal connectivity. The striatum shares extensive reciprocal projections with most areas of the cortex, including the temporal lobe (Yeterian and Pandya 1998; Cho et al. 2013). Through these connections, the striatum can facilitate plasticity in the response properties of cortical neurons even in adulthood (Parent and Hazrati 1995; Jung and Hong 2003). A crucial element of the cortical-striatal dynamics is the sensitivity of dopaminergic neurons in the striatum to reward signals, which can modify behavioral responses to external stimuli through the reinforcement learning process (Schultz et al. 1998; Schultz 1997, 1998, 2002; Calabresi et al. 2007). In a rat model, manipulating auditory corticostriatal connectivity can directly impact auditory decisions (Znamenskiy and Zador 2013; Xiong et al. 2015). In humans, it has been posited that the striatum may play a critical training signal that modulates emergent STG responses to novel speech categories (Lim et al. 2014). Behavioral experiments demonstrate that feedback is required for speech sound acquisition in adulthood, even if the feedback is implicitly reinforcing and not explicit in nature (Vallabha and McClelland 2007; Vlahou et al. 2012). Based on the previous studies that have shown robust engagement of the striatum during reinforcement learning, it has been hypothesized that striatum is engaged in feedback-driven speech category learning (Doupe and Kuhl 1999; McClelland et al. 2002; Formisano et al. 2008; Goudbeek et al. 2008). Functional neuroimaging studies have corroborated this prediction by showing that the striatum is sensitive to the presence and value of feedback during speech category learning (Tricomi et al. 2006; Yi et al. 2016). Specifically, within the striatum, activity in the putamen has been associated with optimal response strategy and individual success in non-native speech learning (Seger 2008; Seger and Miller 2010; Yi et al. 2016). However, the role of the striatum in guiding the neural representation of new speech categories and mediating learning outcome is poorly understood and is one of the focus of this study.

We utilized multi-modal neuroimaging methods to examine the functional role of corticostriatal circuitry in the emergence of novel category representation as well as the behavioral outcome. To this end, we trained adult native speakers of English to categorize non-native Mandarin Chinese lexical tones (see Fig. 1A for the visualization of the stimuli) using trial-by-trial corrective feedback (see Fig. 1B for the experimental procedure). In line with previous behavioral and neuroimaging studies, the tone category stimuli were natural productions from native speakers of Mandarin Chinese in the context of multiple syllables, allowing us to examine processes underlying acquisition of naturalistic speech categories with high acoustic variability in a controlled environment (Maddox and Chandrasekaran 2014; Chandrasekaran et al. 2014, 2015; Yi et al. 2016). As a function of sound-to-category training, we found that tone category representations emerged primarily in the left anterior STG (LaSTG). Crucially, the robustness of these novel representations closely

corresponded with behavioral response patterns, suggesting that the emerging representations are likely to be behaviorally relevant. Second, as training progressed, functional connectivity between the LaSTG and feedback-sensitive areas of the putamen became more differentiable across positive and negative feedback. Finally, we found robust structural connectivity between the LaSTG and the putamen. These results converge towards a system-level account of speech acquisition in adulthood: training results in the emergence of talker-invariant neural representations within a few hundred training trials involving feedback. Functional coupling between temporal lobe regions involved in sensory representation and the striatum, based on the underlying structural connectivity, mediate the acquisition of novel non-native speech categories in adulthood.

Materials and Methods

Participants

Young adult, native speakers of English were recruited from the greater Austin community. All participants underwent audiological testing using pure-tone audiometry and exhibited hearing thresholds of less than 25 dB HL at frequencies between 250 and 8000 Hz (octave steps). Potential participants ($N = 8$) were excluded if they reported a current or history of major psychiatric conditions, neurological disorders, hearing disorders, head trauma, or use of psychoactive medication. Included participants underwent a magnetic resonance imaging session ($N = 30$; 25 females [due to the unbalanced sex of the participants, a *post hoc* analysis was performed to assess the extent to which the behavioral performance in tone category learning differed as a function of sex; a mixed effects analysis of variance (ANOVA) was performed with the early vs. late stages as the within-subject independent variable and the sex (male or female) as the between-subject independent variable; the dependent variable was average accuracy; there was no main effect of sex of the participant [$F_{(1,28)} = 0.09$, $P = 0.771$]; the main effect of learning stage was significant [$F_{(1,28)} = 15.79$, $P < 0.001$]; the interaction between sex of the participant and the learning stage was not significant [$F_{(1,28)} = 0.65$, $P = 0.427$]]; right-handed; ages 18–32 years; mean age = 21.8, $SD = 3.7$), comprising the dataset reported in the current study. All participants were monetarily compensated for their time. All materials and protocols were approved by the Institutional Review Board of the University of Texas at Austin. Participants provided written informed consent before their participation in this study.

Stimulus

Natural exemplars ($N = 40$) of the four Mandarin tones (high flat, low rising, low dipping and high falling, see Fig. 1A) were produced in citation form by two native Mandarin speakers (originally from Beijing; one female) in the context of five monosyllabic Mandarin Chinese words (/bu/, /di/, /lu/, /mo/ and /mi/). These syllables were chosen because they also exist in the American English phonetic inventory. The stimuli were normalized for the RMS amplitude of 70 dB and the duration of 0.4 s (Wong et al. 2009; Perrachione et al. 2011). Five independent native Mandarin speakers correctly identified the four tones (categorization accuracy >95%) and rated the stimuli as highly natural.

Sound-to-Category Training Procedure

Sound-to-category training procedure closely followed a previous study (Yi et al. 2016). Participants performed a sound-to-category mapping task in the scanner while listening to the

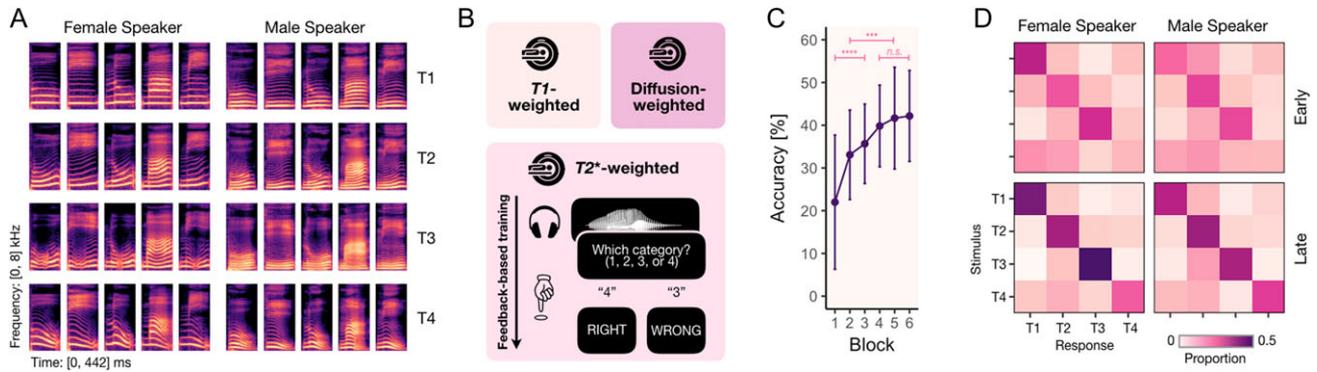


Figure 1 Stimuli, training procedures and behavioral learning performance. (A) Spectrograms for 4 tones contours produced by male and female speakers. T1: flat tone; T2: rising tone; T3, dipping tone and T4: falling tone. Within each speaker-tone combination, each of the five columns corresponds to the syllable context /bu/, /di/, /lu/, /mu/ and /mi/. (B) Imaging protocols and training procedure during functional magnetic resonance imaging (fMRI) scanning. In each trial, a stimulus was played for a fixed duration (442 ms). The participants were given a 2 s window to respond. Corrective feedback (“RIGHT” or “WRONG”, 750 ms) followed stimulus offset was presented after a jitter delay of 2–4 s. (C) Behavioral learning performance (accuracy; y-axis) across the six training blocks (x-axis). Error bars denote standard deviation across the 30 participants. Learning performance changed rapidly within the first three blocks (early stage) but become stable within the last three blocks (late stage), suggesting a plateau of performance in the latter. (D) Behavioral tone category confusion matrices for both male- and female-talker trials in the early and late learning stages.

speech sounds presented through headphones. Visual stimuli including the instructions and feedback were displayed via the in-scanner projector visible using a mirror attached to the head coil. Participants were equipped with a two-button response box in each hand. The experiment consisted of six contiguous scans, or “training blocks.” Before each block, participants were instructed to attend to the fixation cross on the screen. During each trial, an auditory stimulus was presented for 442 ms. Participants were instructed to categorize the sound into one of the four categories. Following the stimulus presentation and response, corrective feedback (i.e., “RIGHT” and “WRONG”) was displayed for 750 ms (see Fig. 1B). To efficiently model signals from stimulus presentation and feedback separately, we employed a jittered stimulus-feedback interval (2–4 s; feedback-stimulus interval: 1–3 s; pooled from a uniform distribution) (Dale 1999; Liu et al. 2001; Birm et al. 2002). If the participant failed to respond within the 2 s following stimulus onset, the response did not register and a cautionary feedback display was presented (i.e., “TIME”). Each stimulus was presented once within each block. The presentation order of the stimuli was pseudo-randomized into a sequence common to all participants but different across learning blocks.

Imaging Acquisition

The participants were scanned using the Siemens Magnetom Skyra 3 T MRI scanner at the Imaging Research Center of the University of Texas at Austin. Whole-brain T1-weighted anatomical images were obtained via MPRAGE sequence (repetition time = 2.53 s; echo time = 3.37 ms; field of view = 25 cm; 256 × 256 matrix; 1 mm × 1 mm voxels; 176 axial slices; slice thickness = 1 mm; distance factor = 0%). T2*-weighted whole-brain blood oxygen level-dependent (BOLD) images were obtained using a gradient-echo multi-band EPI pulse sequence (flip angle = 60° repetition time = 1.8 s; 166 repetitions; echo time = 30 ms; field of view = 25 cm; 128 × 128 matrix; 2 mm × 2 mm voxels; 36 axial slices; slice thickness = 2 mm; distance factor = 50%) using GRAPPA with an acceleration factor of 2. Diffusion-weighted anatomical images were obtained using the following parameters: repetition time = 8.3 s; echo time = 84 ms; field of view = 256 mm × 256 mm; 128 mm × 128 matrix; 64 axial slices; slice

thickness = 2 mm; distance factor = 0%; $b = 700 \text{ s/mm}^2$ in 64 directions.

Functional Magnetic Resonance Imaging Data Preprocessing

All functional imaging data were preprocessed using SPM8 (Wellcome Department of Imaging Neuroscience, London, UK; www.fil.ion.ucl.ac.uk/spm/). First, the T2*-weighted images were corrected for head movement based on the mean image. Then, the high-resolution T1-weighted images were linearly registered to the mean image and further normalized to a standard T1 template in the Montreal Neurological Institute (MNI) space using segmentation-normalization procedure. The realigned functional images were further smoothed with a Gaussian kernel of 6-mm full-width at half-maximum and then entered into the subject-level general linear model (GLM) analysis. Spatial smoothing and normalization procedures were not performed for the multivariate pattern classification (MVPC) analysis.

Univariate Activation Analysis

To identify brain regions that are related to feedback processes, we constructed GLMs at the subject level. Due to the jittered intervals between stimulus and feedback presentation, we were able to model the neural processes of the two types of events while minimizing signal overlap. The design matrix consisted of regressors for the stimulus, trial-by-trial feedback type (i.e., correct, incorrect or non-response) and six head movement parameters as well as the intercept. Temporal high-pass filtering (cutoff at 128 s) and AR1 autocorrelation correction methods were used to minimize the impact of low-frequency drifts. The resulting contrast images (e.g., correct vs. incorrect response during feedback) from the subject-level GLM analysis were then normalized into the MNI space and resampled to $2 \times 2 \times 2 \text{ mm}^3$ voxel size. In the group-level analysis, random-effect model (one-sample t-test) was used to localize the brain areas that were related to the correct versus incorrect feedback. All group-level statistical maps from the univariate analysis were thresholded at voxel-wise $P < 0.005$ initially, with cluster-level

family-wise error (FWE) corrected $P < 0.05$ as implemented in the SPM8 package.

MVPC Analysis

To conduct MVPC analysis on tone category (i.e., high flat, low rising, low dipping and high falling) and syllable identity (i.e., /bu/, /di/, /lu/, /mɑ/ and /mi/), we modeled single-trial brain responses at subject level. The unsmoothed functional images were analyzed for each participant in native space. The least squares single approach was used to model brain responses for each trial during stimulus presentation while controlling for the variance of other events in the same block (Mumford et al. 2012, 2014). Specifically, for each trial, a design matrix was constructed with a regressor of interest for a single trial during stimulus presentation; a regressor of non-interest consisted of other events (i.e., feedback presentation for that trial and stimulus and feedback presentation for the other trials), six head movement regressors and a session mean regressor for each learning block individually. Therefore, 240 subject-level GLM models were constructed totally for each subject. The t-statistic brain maps were calculated and further used for MVPC analysis (Misaki et al. 2010). The searchlight algorithm (Kriegeskorte 2006) with a linear support vector machine (SVM) classifier as implemented in the CoSMoMVPA toolbox (Oosterhof et al. 2016) and LIBSVM toolbox (Chang and Lin 2011) was employed to identify the locus of the neural representations of Mandarin tone categories and syllable identity respectively. At each voxel, the t values within a spherical searchlight (three-voxel-radius sphere, 93 voxels in average across spheres and participants) were extracted for each trial. Therefore, in each spherical searchlight, a $V \times T$ matrix of t values was constructed, where V refers to the number of voxels and T refers to the number of trials (e.g., 93×240). This matrix served as input for a SVM classifier for training and testing. To search for neural representations of tone category that are invariant to surface acoustic features (e.g., talker variability), we employed a leave-one-talker-out cross-validation (CV) procedure. The classifier was trained on the trials from one talker and was subsequently tested on the trials from another talker and vice versa. Thus, only the tone category (or syllable identity) information generalizable across talkers was informative to the classifier. Finally, mean classification accuracy was calculated and mapped back to the voxel at the center of each searchlight sphere.

Our aprior hypothesis is that the auditory corticostriatal system is involved in speech category learning. Therefore, we restricted our searchlight classification analysis to the bilateral superior temporal cortex, consisted of the bilateral STG and the superior temporal pole (see the outline in a render brain in Fig. 2A) that were defined by the atlas of automated anatomical labeling (Tzourio-Mazoyer et al. 2002; Rolls et al. 2015). The selection of the bilateral STG is based on the findings from previous studies using multivariate pattern analysis (Ley et al. 2012; Bonte et al. 2014; Mesgarani et al. 2014; Arsenault and Buchsbaum 2015; Feng, Gan, et al. 2018; Feng, Ingvallson, et al. 2018) that these areas are associated with speech perception and phonetic category representation, and tracing studies in animal models demonstrate strong connectivity between the superior temporal lobe and the striatum (Yeterian and Pandya 1998; Cho et al. 2013). We performed the abovementioned MVPC procedure across all voxels within the pre-defined STG and generated classification brain maps for the first three blocks (early stage of learning) and the last three blocks (late stage of learning), separately. For the group-level analysis, the

classification accuracy map for each participant was first normalized into the MNI space using the parameters estimated from the segmentation-normalization procedure and then entered into a one-sample t-test model tested against chance accuracy (tone: 1/4; syllable identity: 1/5). All statistical maps from MVPC analyses at the group level were thresholded at voxel-wise $P < 0.005$ initially, with cluster-level FWE-corrected $P < 0.05$ as implemented in the SPM8 package.

Neural-Behavioral Correlation and Multidimensional Scaling

We employed representational similarity analysis (RSA) (Kriegeskorte et al. 2008; Kriegeskorte and Kievit 2013) to assess brain-behavioral consistency in tone category response patterns. Specifically, we examined the extent to which neural representations of tone categories that emerge in the STG subregions corresponded to the behavioral confusability of the tone categories. To achieve this, we calculated the similarity between the neural classification confusion matrix and behavioral confusion matrix. We first defined the brain regions that showed significant above-chance classification in the late blocks of training as regions of interest (ROIs). ROIs that were independent of the current dataset were also defined and constructed by re-analyzing data from a previous study (Yi et al. 2016) for result validation. The ROIs in the MNI space were then projected back to the native space for each participant. The multivoxel patterns were then extracted from each ROI separately for each trial. To generate a neural tone category confusion matrix, we conducted ROI-based tone classification analysis with leave-one-talker-out CV procedure. We then measured the similarity between the neural confusion matrix and the behavioral confusion matrix using the “corr2” function implemented in MATLAB 2016b. This brain-behavioral correlation analysis was conducted for each participant separately, and a one-sample t-test was performed on the subject-level correlation coefficients to determine the statistical significance at group level. To visualize the brain-behavioral consistency, we further performed an multidimensional scaling (MDS) analysis on the group-average confusion matrixes. The confusion matrixes were first converted into dissimilarity matrixes by averaging the elements across diagonal and then transform the matrixes into a two-dimensional space using “mdscale” function with the “metricstress” criterion as implemented in MATLAB 2016b.

Psychophysiological Interactions (PPI) Analysis

To examine the extent to which functional interaction (coactivation) between the left putamen (subregions in the striatum) and subregions in the STG changes as a function of feedback type (correct vs. incorrect) as well as the stage of learning (early vs. late), we performed a PPI analysis (Friston et al. 1997). The seed region (the left putamen) was defined by the conjunction between the anatomically defined putamen mask and the group-level activation contrast (i.e., correct > incorrect) map during feedback. The resulting putamen mask in the MNI space was then projected back to the native space based on the transformation matrix for each participant. To ensure selecting the functionally relevant part of the putamen for each participant, we defined the subject-specific putamen ROIs by selecting the voxels that were positive in the contrast of (correct > incorrect) during feedback. In addition, we defined the putamen seed ROIs that were independent of the current dataset for further

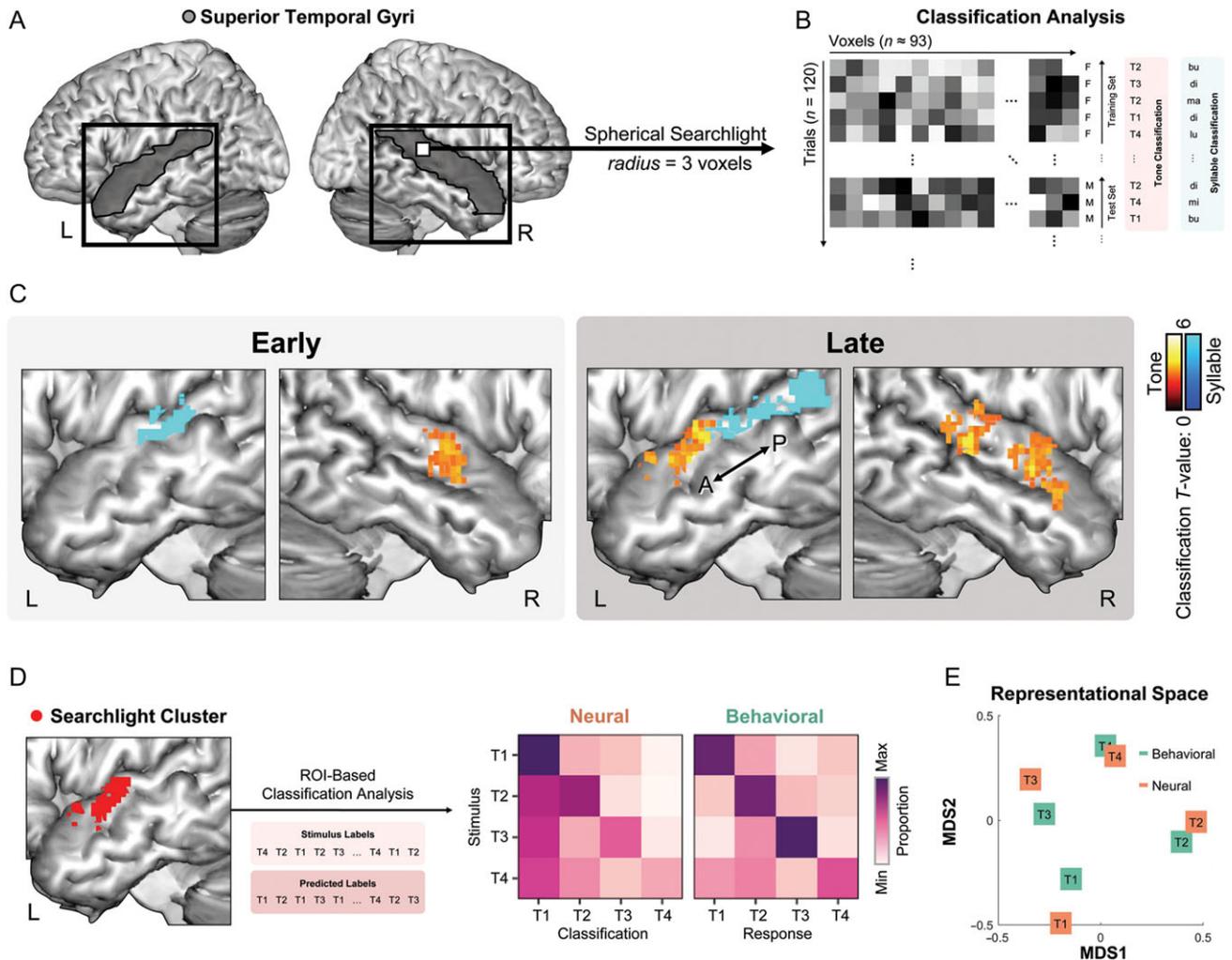


Figure 2. The emergence of tone categorical representation in the superior temporal cortex through training. (A) Searchlight MVPC analysis scheme within predefined STG ROI. (B) Leave-one-talker-out cross-validation procedure. A support vector machine classifier was trained on data from one talker and was tested on unseen data from another talker, and vice versa. We conducted the MVPC for the early (first three blocks) and late (last three blocks) blocks separately. (C) Searchlight MVPC brain maps during the early and late blocks of training. Syllable classification maps were projected onto the same render brain for comparison purpose. Cluster-level FWE-corrected $P < 0.05$. (D) Representational similarity analysis shows strong neural-behavioral correlations for tone confusion pattern in the LaSTG (see Results sections for details). (E) Multidimensional scaling further reveals a strong neural-behavioral consistency on tone category representation in a two-dimensional space.

validation, by re-analyzing the data from a previous study (Yi et al. 2016). Voxels that activated in the feedback contrast (correct > incorrect) across all blocks were selected (thresholded at cluster-level FWE-corrected $P < 0.05$; see Supplementary Fig. S3). We then extracted the subject-specific average activity time series in the putamen as the physiological regressor. We constructed a GLM for each subject using a design matrix with three regressors: 1) one psychological regressor (trial onsets of correct or incorrect response during feedback convolved with a canonical HRF), 2) one physiological regressor (signal from the seed region that was summarized as the first eigenvarieties of the time series) and 3) the interaction effect between the first and the second regressors. Specifically, we modeled each feedback type individually to examine how feedback valence changes the putamen-STG functional coupling (or how the STG activity was modulated by the putamen activity with feedback type). We also included the six head movement parameters as non-interest regressors. To examine the training effect, we conducted the above-described PPI analysis for early (first three

blocks) and late (last three blocks) blocks of training separately. Moreover, we performed the same PPI analysis to examine the functional coupling effect between the dorsolateral prefrontal cortex (DLPFC) and STG as a control analysis to examine the functional specificity of the putamen-STG circuitry.

Structural Connectivity Analysis

Diffusion-weighted images were corrected for eddy current and then brain-extracted using BET (Jenkinson et al. 2005). The pre-processed images were registered first to the native structural space and then to the MNI152 T1 template using ANTs (Avants et al. 2009; Schwarz et al. 2014). Fractional anisotropy (FA) images were created by performing DTIFIT on the preprocessed images (Behrens et al. 2003). Finally, BEDPOSTX was performed to build diffusion parameters distributions (Behrens et al. 2007). Probabilistic tracking was performed with a subregion in the STG (functionally defined brain region using searchlight MVPC analysis) as the seed and the putamen (functionally defined

subregions in the left hemisphere using univariate activation analysis, correct versus incorrect feedback; masked with binarized putamen masks from the Harvard–Oxford subcortical atlas as the waypoint target. For each participant, the following parameters were used: 50 000 samples; curvature threshold = 0.2; maximum number of steps = 2000; step length = 0.5 mm (Behrens et al. 2007). After probabilistic tracking had been completed, each of the tracks for individual participants was independently thresholded in the top 95% percentile, leaving only 5% of the voxels with the highest intensity values. These maps were subsequently binarized, transformed from the native diffusion space to the standard space and added across all participants (Javad et al. 2014; Darki and Klingberg 2015). Finally, a group-level map was constructed by retaining only the voxels with higher intensity values than 15, which corresponded to at least 50% of the 30 participants from whom a given voxel survived the 95% percentile threshold (Darki and Klingberg 2015). This group-level threshold retained 2.5% of the voxels (Saur et al. 2008, 2010). The resulting group map was binarized and registered back to the native diffusion space. FA values for each participant were calculated within this group-wise map.

Results

Behavioral Training Outcomes

With sound-to-category training, participants learned to categorize Mandarin tone categories significantly above-chance. In the first block, the mean accuracy across the participants ($N = 30$) was 22% (range: 0–40%; $SD = 9\%$; chance level = 25%). In the final block, the mean accuracy was 42% (range: 13–100%; $SD = 24\%$). A paired t-test demonstrated that final block accuracy was significantly higher than the initial block accuracy ($t_{(29)} = 5.01$, $P < 0.001$). In addition, accuracy from the last three blocks was significantly higher than that from the first three blocks ($t_{(29)} = 3.99$, $P < 0.001$). Within the first three blocks, accuracy was significantly different, suggesting that learning process occurred rapidly ($F_{(2, 58)} = 15.86$, $P < 0.001$). However, accuracy was not significantly different across the last three blocks ($F_{(2, 58)} = 0.76$, $P = 0.472$), indicating that learning performance was relatively stable over the later training blocks (Fig. 1C). Therefore, we defined the first three blocks as “early” learning and the last three blocks as “late” learning. Moreover, the behavioral tone category confusion patterns were similar across the male and female talkers used in the training (see Fig. 1D).

The Emergence of Speech Category Representations in the STG

We performed a searchlight MVPC analysis with cross-talker (leave-one-talker-out) CV procedure (Feng, Gan, et al. 2018) to examine the emergence of talker-invariant representations of the non-native speech categories within the bilateral STG (see Fig. 2A). The MVPC analyses were conducted for the early (the first three blocks; each block consists of 40 trials) and late (the last three) blocks of the training separately (see Fig. 2B for the MVPC analysis procedure). During the initial three blocks (early), we found that region with significantly above-chance classification performance in tone category was located restricting in the right anterior STG (RaSTG) (peak MNI coordinates: $x = 64$, $y = -6$, $z = 0$; cluster size: 127 voxels). No significant cluster was found in the left hemisphere (Fig. 2C, left panel). During the last three training blocks (late), we found several brain areas within the bilateral STG where classification performance was significantly above the chance level (Fig. 2C, right panel). These regions included the LaSTG

($x = -62$, $y = -6$, $z = 4$; cluster size: 131 voxels), RaSTG ($x = 58$, $y = -2$, $z = -8$; cluster size: 88 voxels) and the right posterior STG (RpSTG) ($x = 66$, $y = -24$, $z = 2$; cluster size: 91 voxels). In comparison with the tone classification, we also examined the classification of (native) syllable identity (/bu/, /di/, /lu/, /ma/ and /mi/) and found significantly above-chance classification accuracy in the bilateral posterior STG during both early and late training blocks (Fig. 2C; see Supplementary Fig. S1 for the whole-brain searchlight results).

To further validate these findings, we conducted two separate ROI analyses with different approaches. Using these approaches, we explicitly tested our findings that tone classification accuracy significantly increased for the late blocks relative to early blocks for left STG but not the right STG. In these ROI analyses, we ensured that the ROI definition is independent of ROI-based classification analysis. First, we conducted the ROI analysis with leave-one-subject-out procedure; that is, ROI definition was based on $N-1$ (i.e., 29) subjects and ROI-based classification analysis was conducted on the held-out subject. To do so, each single-trial brain activation map was normalized into MNI space for each subject first. For each held-out iteration, searchlight classification analysis within the STG was employed for each of the $N-1$ subjects. A group-level t-test ($n = 29$) was conducted to determine the above-chance voxels for each hemisphere of the STG. Voxels that were survived at the cluster-level FWE-corrected $P = 0.05$ were selected and defined as ROIs. Furthermore, the ROI-based classification analysis was then conducted for the held-out subject with the ROI that defined by the other 29 subjects previously. This iteration process was repeated 30 times. Finally, a group-level t-test ($N = 30$) was conducted to determine the statistical significance of the ROI analysis. This embedded leave-one-subject-out ROI analysis approach ensures that the ROI definition and ROI-based classification analysis are independent. Note that, here the ROI definition was only conducted using trials in the last three blocks (late) instead of all blocks because of we hypothesized that robust (and stable) brain representation of category emerged in the late blocks (consistent with a behavioral plateau during the later blocks). In prior work, we have shown that in the earliest block, participants often employ a “random-responder” strategy and are in the process of mapping out the sound-to-button press; these issues could contaminate the trained model. Our priori hypothesis and observation are based on our previous work (e.g., Yi et al. 2016) and behavioral learning patterns (see Fig. 1C and behavioral results section). By using this unbiased leave-one-subject-out ROI analysis, we found that left STG showed significant increased tone classification accuracy in the last three blocks compared with that in the first three blocks (late vs. early: $t_{(29)} = 2.34$, $P = 0.026$; early vs. chance: $t_{(29)} = -0.16$, $P = 0.876$; late vs. chance: $t_{(29)} = 3.56$, $P = 0.001$). Of particular interest is that (unlike the left STG) the tone classification accuracy in the right STG was significantly above-chance in both early and late stages (early vs. chance: $t_{(29)} = 4.56$, $P < 0.001$; late vs. chance: $t_{(29)} = 3.565$, $P = 0.001$), and the classification accuracy did not change over learning stages (late vs. early: $t_{(29)} = -0.44$, $P = 0.662$). Second, to further validate (and replicate) the above ROI analysis finding, we also defined STG ROIs independent of the current dataset. To do so, we re-analyzed data from a previously published study (Yi et al. 2016, *Cerebral Cortex*, $N = 23$) using the same searchlight MVPC approach within the bilateral STG. Yi et al. (2016) used the identical speech training paradigm (and similar scanning protocol) as the current experiment, which makes it a perfect localizer dataset for ROI definition. We found that two brain regions showed significantly above-chance tone classification accuracy

(FWE-corrected $P < 0.05$ at cluster level) in the late blocks but not in the early blocks (see Supplementary Fig. S2). These regions included the LaSTG and the right middle portion of STG (RSTG). We defined the two regions as ROIs to conduct ROI-based classification analysis on the current dataset. Again, we found that the LaSTG showed significantly increased tone classification accuracy in the last three blocks compared with that of the first three blocks (late vs. early: $t_{(29)} = 2.17$, $P = 0.017$; early vs. chance: $t_{(29)} = 2.04$, $P = 0.023$; late vs. chance: $t_{(29)} = 4.96$, $P < 0.001$). In contrast, tone classification accuracy of the RSTG was not significantly changed over learning stages (late vs. early: $t_{(29)} = -0.55$, $P = 0.707$; early vs. chance: $t_{(29)} = 4.16$, $P < 0.001$; late vs. chance: $t_{(29)} = 3.43$, $P < 0.001$).

The Neural-Behavioral Similarity in Tone Confusion Patterns

To examine the behavioral relevance of the emergent representation in the LaSTG, we assessed the neural-behavioral similarity in tone confusion patterns by using RSA (Kriegeskorte et al. 2008). We computed the correlation between the neural confusion matrices, derived from the MVPC analysis performed in the LaSTG, and the behavioral confusion matrices, derived from each participant's behavioral responses during training (Fig. 2D). The neural and behavioral matrices were significantly correlated ($t_{(29)} = 2.91$, $P = 0.007$) (Fig. 2D, middle panel) compared to chance at group level. This finding indicates that the individual differences in neural confusion pattern of newly acquired speech categories are associated with behavioral confusion patterns. For visualization of the brain-behavior relationship, we conducted MDS analyses on both the neural and behavioral data. We qualitatively confirmed a high degree of similarity in the dimensions underlying tone category representations derived from neural and behavior data (Fig. 2E).

Feedback-Sensitive Brain Activations in the Striatum

Consistent with our previous study that has shown feedback sensitivity in the striatum during speech category learning (Yi et al. 2016), we found that the bilateral putamen, extending to the head of the caudate nucleus and nucleus accumbens (cluster in the left hemisphere, peak MNI: $x = -18$, $y = 8$, $z = 4$, cluster size = 269; cluster in the right hemisphere, peak MNI: $x = 18$, $y = 8$, $z = -8$, cluster size = 404), were significantly more activated during positive feedback (correct trial) than negative feedback (incorrect trial). We further conducted ROI analyses with two independently pre-defined putamen ROIs to investigate the learning stage (early vs. late) by feedback type [correct vs. incorrect] interaction. The bilateral putamen ROIs were defined by re-analyzing the data from our previous study (Yi et al. 2016). Voxels that activated in the feedback contrast (correct > incorrect) across all blocks were selected (thresholded at cluster-level FWE-corrected $P < 0.05$). We further restricted the activated voxels within the putamen masks that defined by the Harvard-Oxford subcortical atlas (see Supplementary Fig. S3 for visualization of the putamen activations in the two separate datasets). The ROI-based univariate activation analysis showed that main effects of feedback type were significant for bilateral putamen (left putamen: $F_{(1,29)} = 11.73$, $P < 0.001$; right putamen: $F_{(1,29)} = 31.12$, $P < 0.001$) while main effects of learning stage were not significant (left putamen: $F_{(1,29)} = 1.96$, $P = 0.303$; right putamen: $F_{(1,29)} = 0.86$, $P = 0.587$). The stage-by-feedback interaction effects were not significant for either hemisphere (left putamen: $F_{(1,29)} = 0.77$, $P = 0.618$; right putamen: $F_{(1,29)} = 4.15$, $P = 0.088$). These

results suggest that feedback processing-related putamen activations are stable across learning blocks.

In addition, we observed that several brain regions related to error monitoring, including the DLPFC (cluster in the left hemisphere peak MNI: $x = -20$, $y = 58$, $z = 18$, cluster size = 253 voxels; right hemisphere: $x = 26$, $y = 44$, $z = 20$, cluster size = 525 voxels) and bilateral temporal-parietal junction that extending to the Heschl's gyrus (left hemisphere peak MNI: $x = -58$, $y = -20$, $z = 12$, cluster size = 281 voxels; right hemisphere: $x = 48$, $y = -20$, $z = 10$, cluster size = 340 voxels) as well as anterior cingulate cortex (ACC, peak MNI: $x = -8$, $y = 8$, $z = 36$) were significantly more active during incorrect response relative to correct response during feedback. In keeping with our predictions regarding auditory cortico-striatal circuitry, in the next sections we examined functional and structural connectivity between the putamen and the LaSTG. These brain regions were identified (using ROI as well as whole-brain analyses) on the basis of MVPC and univariate analyses to encode information related to the non-native speech categories (LaSTG) and performance feedback (bilateral putamen), respectively.

Functional Connectivity between Striatum and STG during Feedback Processing

We examined the functional interactions between the striatum and the STG during feedback processing by employing psychophysiological interaction (PPI) analysis. To this end, we first identified voxels within the left putamen that were significantly activated in the contrast of [correct - incorrect] feedback, based on an independent anatomical mask derived from the Harvard-Oxford subcortical atlas (L Putamen, see Fig. 3A). Since the striatal activation pattern included the ventral striatum, we used this approach to restrict the analyses to the putamen (based on our apriori hypothesis). Next, we identified voxels within the left STG that showed greater-than-chance decoding for tone categories in the late stage of learning (i.e., LaSTG; see Fig. 3A). Using the left putamen as the seed region, we assessed effective connectivity between the left putamen and LaSTG by calculating PPI effect for each type of feedback (correct and incorrect trials during feedback) (Friston et al. 1997), across early (blocks 1-3) and late (blocks 4-6) blocks. The PPI analysis can reveal the extent to which the functional coactivation between the above two regions could be modulated by our experimental conditions. Across blocks, incorrect feedback increased functional coupling between the left putamen and the LaSTG compared to that of correct feedback (the main effect of feedback type: $F_{(1,28)} = 9.98$, $P = 0.004$, repeated measured ANOVA; see Fig. 3B for visualization of the feedback-related PPI effect in a representative subject). We also found a significant interaction effect between learning stage (early vs. late) and feedback type (correct vs. incorrect) ($F_{(1,28)} = 11.02$, $P = 0.002$). Planned t-tests revealed that during incorrect feedbacks, the PPI effect was higher in the late stage relative to the early stage ($t_{(29)} = 2.39$, $P = 0.024$) while the trend was opposite during correct trials ($t_{(29)} = -2.05$, $P = 0.050$) (Fig. 3C). In addition to the left putamen, we examined feedback-associated effective connectivity between the left DLPFC and the LaSTG. The frontal cortex is argued to be critically involved in the supervised tuning of emergent representations of novel speech sounds (Myers 2014). In our study, we identified a significantly activated cluster in the left DLPFC for the feedback (incorrect-correct) contrast, which is relatively anterior and inferior to the cluster reported in the opposite contrast in a previous study utilizing the same training procedures (Yi et al. 2016). Using this DLPFC mask as

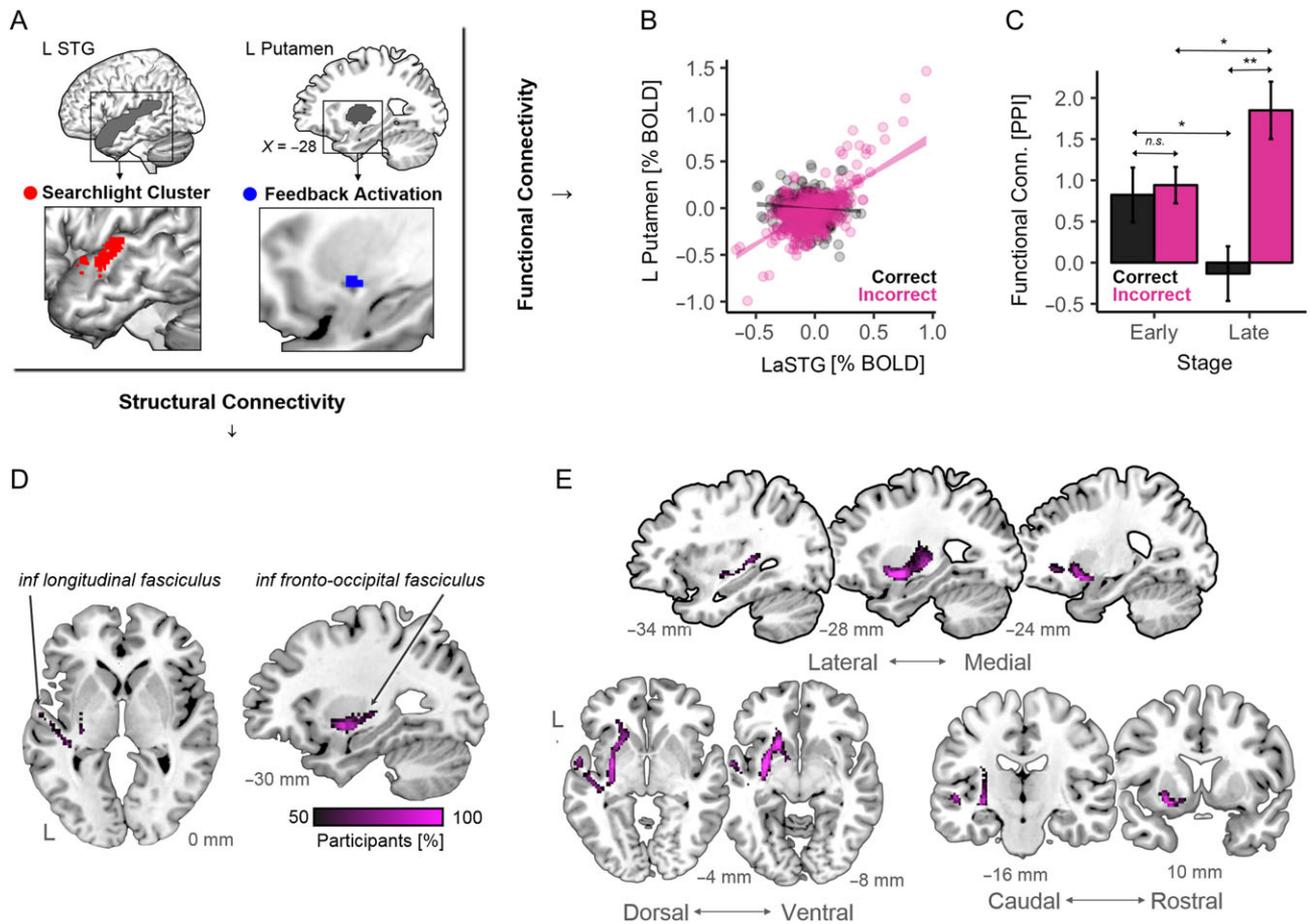


Figure 3. Auditory corticostriatal functional and structural connectivity were associated with speech category learning. (A) In all connectivity-based analyses, the cortical (LaSTG) and striatal (putamen) regions of interest (ROI) were defined using the MVPC and univariate activation analyses, respectively. (B) ROI-based psychophysiological interaction (PPI) analysis showed increased functional coupling between the left putamen and LaSTG during incorrect feedback (magenta) compared to correct feedback (black) across blocks for a representative participant. (C) the feedback-related PPI effect (i.e., incorrect > correct) increased over the early and late blocks during training. Notes: * $P < 0.05$; ** $P < 0.01$; error bars: s.e.m. (D) Probabilistically estimated structural corticostriatal pathways. Across 30 participants, probabilistic tracking was performed between the LaSTG and the putamen, leaving only top 5% voxels for each participant. (E) Brain slices from three directions showing the auditory corticostriatal fiber pathways.

the seed and the LaSTG cluster as the target, we additionally performed PPI analyses for feedback (correct vs. incorrect). We could not find evidence that the functional coupling between DLPCF and LaSTG changed across early and late stages either during incorrect trials ($t_{(29)} = 0.37$, $P = 0.71$) or correct trials ($t_{(29)} = 1.56$, $P = 0.13$). Thus, the effective connectivity analyses reported here are consistent with an increased modulation of STG activity by the putamen based on error monitoring throughout training but fail to support similar patterns using the same analysis for the frontal cortex.

To validate the above findings and further examine the specificity of the left putamen-LaSTG PPI effect, we conducted additional control PPI analyses with pre-defined brain regions. We first defined ROIs by re-analyzing data from our previous study (Yi et al. 2016) with the searchlight tone category classification approach. Two regions were found and selected as ROIs (i.e., LaSTG and RSTG, see Supplementary Fig. S2). We conducted ROI-based PPI analyses on the current dataset with the two ROIs and found that left putamen-LaSTG PPI effect for feedback type (incorrect vs. correct) significantly increased across learning stages (late [incorrect–correct] vs. early [incorrect–correct]: $t_{(29)} = 2.22$, $P = 0.014$; early [incorrect–correct]: $t_{(29)} = 0.90$, $P = 0.186$; late [incorrect–correct]: $t_{(29)} = 3.47$, $P < 0.001$). In

contrast, the left putamen-RSTG feedback-related PPI effect (incorrect vs. correct) did not significantly increase (late [incorrect–correct] vs. early [incorrect–correct]: $t_{(29)} = 1.39$, $P = 0.090$; early [incorrect–correct]: $t_{(29)} = 1.24$, $P = 0.110$; late [incorrect–correct]: $t_{(29)} = 2.70$, $P = 0.003$; see Supplementary Fig. S4). In addition, using right putamen as seed region, we further found that the feedback-related PPI effect [incorrect vs. correct] were not significantly different between early and late blocks for either the right putamen-LaSTG (early vs. late: $t_{(29)} = 0.16$, $P = 0.683$; early: $t_{(29)} = 3.05$, $P = 0.003$; late: $t_{(29)} = 3.26$, $P = 0.002$) or the right putamen-RSTG circuitry (early vs. late: $t_{(29)} = 1.22$, $P = 0.216$; early: $t_{(29)} = 2.77$, $P = 0.007$; late: $t_{(29)} = 4.39$, $P < 0.001$). These results altogether suggest that left putamen-LaSTG PPI effect is (specifically) sensitive to learning.

Linking Corticostriatal Pathways to Cortical Representations and Learning Outcomes

We used probabilistic fiber tracking to identify the auditory corticostriatal fiber pathways and examine the extent to which pre-existing structural integrity of the pathways related to the cortical representation of speech categories and behavioral learning outcomes. Within the striatum, we identified the

voxels within the putamen that showed higher functional activation for feedback during correct trials relative to incorrect trials. Within the STG, we identified the cluster that showed higher-than-chance decoding for tone categories in the late stage of training, revealed using the searchlight MVPC analysis (i.e., LaSTG). Using the LaSTG mask as the seed and the left putamen mask as a waypoint target, we probabilistically estimated corticostriatal white-matter pathways. The resultant pathways were thresholded for the top 5% voxels in individual participants and then subjected to the group threshold of 50% of the participants (Saur et al. 2008, 2010; Javad et al. 2014; Darki and Klingberg 2015). The lateral most parts of the estimated pathway close to the seed LaSTG region corresponded to the inferior longitudinal fasciculi (Fig. 3D). Near the ventral portion of the target region of putamen, the pathway extended to the inferior frontal-occipital fasciculi (see Fig. 3E for multi-slice visualization of the auditory-striatum structural pathway). Moreover, the individual differences in FA values of this pathway were associated with the individual differences of behavioral outcomes and MVPC classification accuracy in the LaSTG (see detailed correlation results in Supplementary Material).

Discussion

Talker-invariant neural representations of new speech categories emerged in the left STG within a few hundred trials of sound-to-category training. Emergent representations of the learned speech categories in the left STG were related to individual participants' behavioral response patterns, suggesting that the representational neural plasticity is behaviorally relevant and related to training, rather than repeated exposure. In the striatum, the bilateral putamen was activated more for correct relative to incorrect feedback. Functional coupling between the feedback-sensitive regions within the left striatum and the representational regions of the LaSTG was greater for incorrect relative to correct feedback, and this difference increased in the late stage of training, suggesting that the corticostriatal functional connectivity plays an important role in fine-tuning emerging speech representations. Finally, fiber tractography showed robust structural connectivity in the pathway of the inferior fronto-occipital fasciculus and the inferior longitudinal fasciculus that connecting the feedback-sensitive striatal regions and the left STG regions that represent the newly learned tone categories. These findings altogether point to an important role of the auditory corticostriatal system in the emergence of talker-invariant neural representations of newly acquired speech categories in adulthood.

The Learning-Induced Emergence of Neural Representations of Newly Acquired Speech Categories

Electrophysiological studies across species have shown that novel sound category learning alters the spatiotemporal properties of the neural responses in the auditory cortex (Brown et al. 2004; Ohl and Scheich 2005; Zhang et al. 2009). Novel sound categories can be decoded from the multivoxel response patterns of the human primary auditory cortex after training (Ley et al. 2012). Previous activation-based functional magnetic resonance imaging studies have found increased activation in the bilateral STG after categorization training or sound-to-word mapping training (Callan et al. 2003; Wang et al. 2003; Golestani and Zatorre 2004; Desai et al. 2008; Leech et al. 2009; Myers and Swan 2012). Here, we show that short-term sound-to-category training is associated with the emergence of novel speech

category representations in the LaSTG. It is worth noting that these novel STG representations are invariant to surface acoustic features and spatially distinct from regions representing native syllables. These findings suggest that short-term training experience shapes the talker-invariant neural representation of novel speech sound category in the left STG. This contrasts with the functional response patterns in the right STG which yielded significantly reliable decoding of tones irrespective of training stage. One possibility is that the response patterns in the right hemisphere reflect pitch processing that differs from the left hemisphere counterpart, in which the left STG reflect category-specific pitch distinctions that emerge as a function of category learning (Zatorre and Gandour 2008). This possibility is consistent with our recent findings that the right STG activation patterns encode pitch height instead of pitch direction information, while the left STG is sensitive to both pitch height and direction in a group of native speakers of Mandarin (Feng, Gan, et al. 2018).

Previous studies examining training-related neuroplasticity have primarily focused on the outcome of training (Callan et al. 2003; Wang et al. 2003; Golestani and Zatorre 2004; Desai et al. 2008; Leech et al. 2009; Ley et al. 2012; Myers and Swan 2012). Less is known regarding the process underlying the emergence of neural representations of speech categories as a function of training, and how the representational structure of newly acquired speech categories relates to behavior. Here, we show not only that the neural representation of novel speech categories changes as training progresses but also that there is a strong association between the patterns of neural category representations and behavioral categorization of the stimuli. Our findings advance the understanding of the representational neural plasticity during the learning process and clarify the relationship with individual variability in brain-behavioral response patterns.

Feedback-Sensitive Subregions of the Striatum Modulate the Acquisition of Non-Native Speech Categories

While infants can acquire speech categories in a mostly unsupervised manner, adult speech learning is feedback dependent (Doupe and Kuhl 1999; McClelland et al. 2002; Goudbeek et al. 2008). However, the underlying cognitive and neural mechanisms of how corrective feedback guides the acquisition of new speech categories are poorly understood. Here, we used corrective feedback to signal accuracy of a given categorization response (i.e., RIGHT vs. WRONG) on a trial-by-trial basis, enabling the participant to construct and update their internal representation of speech sounds and learn to abstract them into categories irrespective of surface acoustic variabilities. We found that the striatum is more activated with the positive feedback (i.e., RIGHT) than with the negative feedback (i.e., WRONG) during training. This finding suggests that the striatum plays a significant role in sound-to-reward mapping (Yi et al. 2016). Our results are consistent with an emerging view regarding the key role of the basal ganglia in music and speech processing (Salimpoor et al. 2011, 2013; Lim et al. 2014).

In the current study, we provide a comprehensive account of the striatal involvement in feedback-dependent speech category learning by through multiple analytical approaches. Results from univariate activation analysis and multivariate pattern analysis were integrated to investigate the functional interaction between the feedback-sensitive striatum regions and category-sensitive superior temporal lobe regions during learning. Prior work on music processing demonstrates

increased functional coupling between auditory regions and the ventral striatum as a function of rewarding music (Salimpoor et al. 2013). Here, we posited that the striatum exhibits dynamic coupling with the auditory cortex during error monitoring. In support of our hypothesis, we found that functional coupling between the striatum and category representation areas in the STG was associated with error processing. Specifically, negative feedback triggered by incorrect categorization responses was associated with higher putamen-STG functional coactivation relative to positive feedback triggered by correct categorization responses. We did not find a significant change in coupling between the left DLPFC and the left STG. This is a relevant finding because DLPFC regions are highly active during feedback processing and have been hypothesized to be involved in shaping STG responsivity to novel speech categories (Myers 2014). Previous studies have shown that striatum activity is tuned to rewards (Schultz et al. 1998; Schultz 1997, 2002; Yi et al. 2016) and the striatum could selectively release or inhibit the cortex to allow for selection of a motor (Humphries et al. 2006) or cognitive strategy (Frank 2005). During feedback-based speech categorization training, it is possible that the striatum selectively tunes the activity in the auditory and motor cortices to perpetuate the selection of category responses that maximize reward. This process can be understood as the result of recursive operations between the striatum and the cortex. Recursive operations refer to a set of iterative processes during which the results from a single iteration are fed back through the loop for further processing and elaboration. Learning new speech categories may depend on recursive, bootstrapping functional interactions in cortico-striatal circuits during sound-to-category mapping (Seger and Miller 2010). Hence, in the context of speech category learning, recursive operations may facilitate the fine-tuning of talker-invariant category representations in the auditory cortex by selecting appropriate motor strategies which yield positive feedback (Seger 2008). During speech category learning, interactions between the striatum and auditory cortex might enable the integration of perceptual and feedback signals, mediating the shift from novice to skilled behavioral performance. This may be the putative basis for the strong correspondence between the robustness of emergent representations and behavioral response patterns.

In addition, we found that significantly increased functional coupling during error monitoring (i.e., incorrect vs. correct trials during feedback), as evidenced by PPI analysis was only observed in the late stage of learning. However, we did not find such feedback type by learning stage interaction effect for the univariate activation in the striatum. These intriguing findings suggest that feedback monitoring supported by the striatum regions is associated with emergent representational plasticity and the functionality of the striatum may be different across different learning stages. Early learning may be reliant on positive feedback supported by striatum activation. Due to the fact that the striatum has rich input and output connections with prefrontal cortices that related to rule learning and subcortical regions that associated with memory formation (e.g., hippocampus). It is possible that the interaction between these regions contributes to speech learning in the early (novice) stage although our current data does not speak to this possibility. In contrast, during later stage of learning, successful learning may be more depended on error monitoring supported by putamen-STG coactivation during this stage.

Alternatively, a predictive coding mechanism (Schultz 1997; Schultz et al. 1997; Diederen et al. 2016, 2017) can be considered

as another explanation for the learning stage-by-feedback type PPI interaction. Learning has been proposed to be associated with changes in the prediction about future events such as rewards or feedbacks (Schultz 1997). Learners' brain is constantly generating and updating hypotheses that predict feedback valances/outcomes. The learning stage-by-feedback PPI interaction effect may relate to the degree of prediction error at different stages. In the late stage of learning, prediction error may increase when learners encountered negative feedback that did not match learners' (more robust) prediction as compared with early learning. This putative prediction error-related PPI modulation between striatum and STG may facilitate the fine-tuning of talker-invariant category representations in the auditory cortex so that appropriate strategies could be employed for yielding positive feedback.

Structural Integrity of the Auditory Striatal–Cortical White-Matter Pathway

The functional neuroimaging results as discussed above suggest that there are significant functional interactions between the cortical (STG) and striatal (putamen) regions that related to the acquisition of novel non-native speech categories. To examine the structural basis of the observed striatal–cortical connectivity, we assayed the white-matter fiber pathways between the LaSTG (defined using MVPC analysis) and the putamen (identified using univariate activation analysis). In the resulting tractography that was highly reliable across the participants, the inferior fronto-occipital fasciculus extended across the anteroposterior axis of the ventral putamen, connecting to the inferior longitudinal fasciculus that reached the LaSTG from the dorsomedial to ventrolateral axis. While the relatively modest sample size of the current study precludes us from making direct inferences about the relationship between the integrity of this striatal–cortical pathway and the behavioral learning performance, the characterization of the structural connectivity provides an insight into the neural infrastructure for the dynamic functional coupling that was observed between the putamen and the LaSTG during speech category learning.

The present results establish the behavioral relevance of corticostriatal/striatocortical connectivity in speech category learning. Corticostriatal dysfunction has been linked to various speech and language disorders, such as aphasia (Brunner et al. 1982; Damasio et al. 1982), disruption of temporal patterns in speech production (Volkman et al. 1992) and stuttering (Giraud et al. 2008). These previous studies have provided a compelling background for the consideration of the integral role that the corticostriatal networks play in speech perception (Kotz and Schwartze 2010). Here, we addressed the fundamental mechanisms underlying the role of this system in speech learning (Lim et al. 2014). Animal models show that subregions of the superior temporal cortex, including the core and belt auditory areas, project to the striatum in a topographically systematic manner (Borgmann and Jürgens 1999; Jung and Hong 2003). We posit that the corticostriatal–cortical white-matter fiber pathways provide the structural infrastructure for building critical neural representation patterns of new speech categories throughout training.

Conclusion

By characterizing the category learning-induced plasticity in neural representation of speech category, feedback-related brain activation and functional connectivity during learning, as

well as anatomical striatal–neocortical fiber pathways, our study provides new insights and fundamental knowledge into the mechanisms through which corticoatrial circuit facilitates the emergence of neural representation of speech categories and mediates behavioral performance. More broadly, our study elucidates fundamental neurobiological mechanisms underlying speech acquisition.

Supplementary Material

Supplementary material is available at *Cerebral Cortex* online.

Funding

Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Numbers R01DC015504 and R01DC013315 (to B.C.).

Notes

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. *Conflict of Interest*: None declared.

References

- Arsenault JS, Buchsbaum BR. 2015. Distributed neural representations of phonological features during speech perception. *J Neurosci*. 35:634–642.
- Avants BB, Tustison N, Song G. 2009. Advanced normalization tools (ANTS). *Insight J*. 2:1–35.
- Behrens TEJ, Berg HJ, Jbabdi S, Rushworth MFS, Woolrich MW. 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage*. 34:144–155.
- Behrens TEJ, Woolrich MW, Jenkinson M, Johansen-Berg H, Nunes RG, Clare S, Matthews PM, Brady JM, Smith SM. 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn Reson Med*. 50:1077–1088.
- Birn RM, Cox RW, Bandettini PA. 2002. Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. *Neuroimage*. 15:252–264.
- Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E. 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J Neurosci*. 34:4548–4557.
- Borgmann S, Jürgens U. 1999. Lack of cortico-striatal projections from the primary auditory cortex in the squirrel monkey. *Brain Res*. 836:225–228.
- Brown M, Irvine DRF, Park VN. 2004. Perceptual learning on an auditory frequency discrimination task by cats: association with changes in primary auditory cortex. *Cereb Cortex*. 14:952–965.
- Brunner RJ, Kornhuber HH, Seemüller E, Suger G, Wallech CW. 1982. Basal ganglia participation in language pathology. *Brain Lang*. 16:281–299.
- Calabresi P, Picconi B, Tozzi A, DiFilippo M. 2007. Dopamine-mediated regulation of corticoatrial synaptic plasticity. *Trends Neurosci*. 30:211–219.
- Callan DE, Tajima K, Callan AM, Kubo R, Masaki S, Yamada RA. 2003. Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage*. 19:113–124.
- Chandrasekaran B, Koslov SR, Maddox WT. 2014. Toward a dual-learning systems model of speech category learning. *Front Psychol*. 5:825.
- Chandrasekaran B, Yi HG, Blanco NJ, McGeary JE, Maddox WT. 2015. Enhanced procedural learning of speech sound categories in a genetic variant of FOXP2. *J Neurosci*. 35:7808–7812.
- Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2:1–27.
- Cheour M, Ceponiene R, Lehtokoski A, Luuk A, Allik J, Ahlo K, Naatanen R. 1998. Development of language-specific phoneme representations in the infant brain. *Nat Neurosci*. 1:351–353.
- Cho YT, Ernst M, Fudge JL. 2013. Cortico-amygdala-striatal circuits are organized as hierarchical subsystems through the primate amygdala. *J Neurosci*. 33:14017–14030.
- Dale A. 1999. Optimal experimental design for event-related fMRI. *Hum Brain Mapp*. 8:109–114.
- Damasio AR, Damasio H, Rizzo M, Varney N, Gersh F. 1982. Aphasia with nonhemorrhagic lesions in the basal ganglia and internal capsule. *Arch Neurol*. 39:15–24.
- Darki F, Klingberg T. 2015. The role of fronto-parietal and fronto-striatal networks in the development of working memory: a longitudinal study. *Cereb Cortex*. 25:1587–1595.
- Desai R, Liebenthal E, Waldron E, Binder JR. 2008. Left posterior temporal regions are sensitive to auditory categorization. *J Cogn Neurosci*. 20:1174–1188.
- DeWitt I, Rauschecker JP. 2012. Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci USA*. 109:E505–E514.
- Diederer KMMJ, Spencer T, Vestergaard MDD, Fletcher PCC, Schultz W. 2016. Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron*. 90:1127–1138.
- Diederer KMJ, Ziauddeen H, Vestergaard MD, Spencer T, Schultz W, Fletcher PC. 2017. Dopamine modulates adaptive prediction error coding in the human midbrain and striatum. *J Neurosci*. 37:1708–1720.
- Diehl RL, Lotto AJ, Holt LL. 2004. Speech perception. *Annu Rev Psychol*. 55:149–179.
- Doupe AJ, Kuhl PK. 1999. Birdsong and human speech: common themes and mechanisms. *Annu Rev Neurosci*. 22:567–631.
- Feng G, Gan Z, Wang S, Wong PCM, Chandrasekaran B. 2018. Task-general and acoustic-invariant neural representation of speech categories in the human brain. *Cereb Cortex*. 28:3241–3254.
- Feng G, Ingvalson EM, Grieco-Calub TM, Roberts MY, Ryan ME, Birmingham P, Burrowes D, Young NM, Wong PCM. 2018. Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients. *Proc Natl Acad Sci USA*. 115:E1022–E1031.
- Formisano E, DeMartino F, Bonte M, Goebel R. 2008. “Who” is saying “what”? Brain-based decoding of human voice and speech. *Science*. 322:970–973.
- Frank MJ. 2005. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J Cogn Neurosci*. 17:51–72.
- Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ. 1997. Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*. 6:218–229.
- Garcia-Lazaro JA, Ahmed B, Schnupp JWH. 2011. Emergence of tuning to natural stimulus statistics along the central auditory pathway. *PLoS One*. 6:e22584.

- Giraud AL, Neumann K, Bachoud-Levi AC, vonGudenberg AW, Euler HA, Lanfermann H, Preibisch C. 2008. Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. *Brain Lang.* 104:190–199.
- Golestani N, Zatorre RJ. 2004. Learning new sounds of speech: reallocation of neural substrates. *Neuroimage.* 21:494–506.
- Goudbeek M, Cutler A, Smits R. 2008. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Commun.* 50:109–125.
- Holt LL, Lotto AJ. 2008. Speech perception within an auditory cognitive science framework. *Curr Dir Psychol Sci.* 17:42–46.
- Holt LL, Lotto AJ. 2010. Speech perception as categorization. *Atten Percept Psychophys.* 72:1218–1227.
- Humphries MD, Stewart RD, Gurney KN. 2006. A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J Neurosci.* 26:12921–12942.
- Javad F, Warren JD, Micallef C, Thornton JS, Golay X, Yousry T, Mancini L. 2014. Auditory tracts identified with combined fMRI and diffusion tractography. *Neuroimage.* 84:562–574.
- Jenkinson M, Pechaud M, Smith S. 2005. BET2: MR-based estimation of brain, skull and scalp surfaces. In: Eleventh annual meeting of the organization for human brain mapping, Toronto, p. 167.
- Jung Y, Hong S. 2003. Corticostriatal connections of the superior temporal regions in the macaque monkey. *Korean J Biol Sci.* 7:317–325.
- Karuzza EA, Emberson LL, Aslin RN. 2014. Combining fMRI and behavioral measures to examine the process of human learning. *Neurobiol Learn Mem.* 109:193–206.
- Kotz SA, Schwartze M. 2010. Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends Cogn Sci.* 14:392–399.
- Kriegeskorte N. 2006. Information-based functional brain mapping. *Proc Natl Acad Sci USA.* 103:3863–3868.
- Kriegeskorte N, Kievit RA. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci.* 17:401–412.
- Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Front Syst Neurosci.* 2:4.
- Kuhl PK, Tsao F-M, Liu H-M. 2003. Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc Natl Acad Sci USA.* 100:9096–9101.
- Leech R, Holt LL, Devlin JT, Dick F. 2009. Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J Neurosci.* 29:5234–5239.
- Ley A, Vroomen J, Hausfeld L, Valente G, DeWeerd P, Formisano E. 2012. Learning of new sound categories shapes neural response patterns in human auditory cortex. *J Neurosci.* 32:13273–13280.
- Lim S-J, Fiez J a., Holt LL. 2014. How may the basal ganglia contribute to auditory categorization and speech perception? *Front Neurosci.* 8:1–18.
- Liu TT, Frank LR, Wong EC, Buxton RB. 2001. Detection power, estimation efficiency, and predictability in event-related fMRI. *Neuroimage.* 13:759–773.
- Maddox WT, Chandrasekaran B. 2014. Tests of a dual-system model of speech category learning. *Bilingualism.* 17:709–728.
- Mcclelland JL, Fiez JA, Mccandliss BD. 2002. Teaching the /r/-/l/ discrimination to Japanese adults: behavioral and neural aspects. *Physiol Behav.* 77:657–662.
- Mesgarani N, Cheung C, Johnson K, Chang EF. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science.* 343:1006–1010.
- Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010. Comparison of multivariate classifiers and response normalizations for MVPA. *Neuroimage.* 53:103–118.
- Mumford JA, Davis T, Poldrack RA. 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage.* 103:130–138.
- Mumford JA, Turner BO, Ashby FG, Poldrack RA. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage.* 59:2636–2643.
- Myers EB. 2014. Emergence of category-level sensitivities in non-native speech sound learning. *Front Neurosci.* 8:238.
- Myers EB, Swan K. 2012. Effects of category learning on neural sensitivity to non-native phonetic categories. *J Cogn Neurosci.* 24:1695–1708.
- Nakahara H, Zhang LI, Merzenich MM. 2004. Specialization of primary auditory cortex processing by sound exposure in the “critical period”. *Proc Natl Acad Sci USA.* 101:7170–7174.
- Ohl FW, Scheich H. 2005. Learning-induced plasticity in animal and human auditory cortex. *Curr Opin Neurobiol.* 15:470–477.
- Oosterhof NN, Connolly AC, Haxby JV. 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front Neuroinform.* 10:27.
- Parent A, Hazrati L. 1995. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res Rev.* 20:91–127.
- Perrachione TK, Lee J, Ha LYY, Wong PCM. 2011. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J Acoust Soc Am.* 130:461–472.
- Rolls ET, Joliot M, Tzourio-Mazoyer N. 2015. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage.* 122:1–5.
- Salimpoor VN, Benovoy M, Larcher K, Dagher A, Zatorre RJ. 2011. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nat Neurosci.* 14:257–262.
- Salimpoor VN, Van DenBosch I, Kovacevic N, McIntosh AR, Dagher A, Zatorre RJ. 2013. Interactions between the nucleus accumbens and auditory cortices predict music reward value. *Science.* 340:216–219.
- Saur D, Kreher BW, Schnell S, Kümmerer D, Kellmeyer P, Vry M-S, Umarova R, Musso M, Glauche V, Abel S, et al. 2008. Ventral and dorsal pathways for language. *Proc Natl Acad Sci USA.* 105:18035–18040.
- Saur D, Schelter B, Schnell S, Kratochvil D, Küpper H, Kellmeyer P, Kümmerer D, Klöppel S, Glauche V, Lange R, et al. 2010. Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. *Neuroimage.* 49:3187–3197.
- Schultz W. 1997. Dopamine neurons and their role in reward mechanisms. *Curr Opin Neurobiol.* 7:191–197.
- Schultz W. 1998. Predictive reward signal of dopamine neurons. *J Neurophysiol.* 80:1–27.
- Schultz W. 2002. Getting formal with dopamine and reward. *Neuron.* 36:241–263.
- Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science.* 275:1593–1599.

- Schultz W, Tremblay L, Hollerman JR. 1998. Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology*. 37:421–429.
- Schwarz CG, Reid RI, Gunter JL, Senjem ML, Przybelski SA, Zuk SM, Whitwell JL, Vemuri P, Josephs KA, Kantarci K, et al. 2014. Improved DTI registration allows voxel-based analysis that outperforms Tract-Based Spatial Statistics. *Neuroimage*. 94:65–78.
- Seger CA. 2008. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci Biobehav Rev*. 32:265–278.
- Seger CA, Miller EK. 2010. Category learning in the brain. *Annu Rev Neurosci*. 33:203–219.
- Tricomi E, Delgado MR, McClelland JL, Fiez JA. 2006. Performance feedback drives caudate activation in a phonological learning task. *J Cogn Neurosci*. 18:1029–1043.
- Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*. 15:273–289.
- Vallabha GK, McClelland JL. 2007. Success and failure of new speech category learning in adulthood: consequences of learned Hebbian attractors in topographic maps. *Cogn Affect Behav Neurosci*. 7:53–73.
- Vallabha GK, McClelland JL, Pons F, Werker JF, Amano S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proc Natl Acad Sci USA*. 104:13273–13278.
- Vlahou EL, Protopapas A, Seitz AR. 2012. Implicit training of nonnative speech stimuli. *J Exp Psychol Gen*. 141:363–381.
- Volkman J, Hefter H, Lange HW, Freund HJ. 1992. Impairment of temporal organization of speech in basal ganglia diseases. *Brain Lang*. 43:386–399.
- Wang Y, Sereno J a, Jongman A, Hirsch J. 2003. fMRI evidence for cortical modification during learning of Mandarin lexical tone. *J Cogn Neurosci*. 15:1019–1027.
- Wong PCM, Perrachione TK, Gunasekera G, Chandrasekaran B. 2009. Communication disorders in speakers of tone languages: etiological bases and clinical considerations. *Semin Speech Lang*. 30:162–173.
- Xiong Q, Znamenskiy P, Zador AM. 2015. Selective corticostriatal plasticity during acquisition of an auditory discrimination task. *Nature*. 521:1–16.
- Yeterian EH, Pandya DN. 1998. Corticostriatal connections of the superior temporal region in rhesus monkeys. *J Comp Neurol*. 399:384–402.
- Yi HG, Maddox WT, Mumford JA, Chandrasekaran B. 2016. The role of corticostriatal systems in speech category learning. *Cereb Cortex*. 26:1409–1420.
- Zatorre RJ, Gandour JT. 2008. Neural specializations for speech and pitch: moving beyond the dichotomies. *Philos Trans R Soc B Biol Sci*. 363:1087–1104.
- Zhang Y, Kuhl PK, Imada T, Iverson P, Pruitt J, Stevens EB, Kawakatsu M, Tohkura Y, Nemoto I. 2009. Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. *Neuroimage*. 46:226–240.
- Znamenskiy P, Zador AM. 2013. Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature*. 497:482–485.