Brain-behavior relationships in incidental learning of non-native phonetic categories

Sahil Luthra[1], Pamela Fuhrmeister[2], Peter J. Molfese[3], Sara Guediche[4], Sheila E. Blumstein[5], & Emily B. Myers[1,2,6]

[1] University of Connecticut | Department of Psychological Sciences,
[2] University of Connecticut | Department of Speech, Language and Hearing Sciences
[3] National Institutes of Health
[4] Basque Center on Cognition, Brain and Language
[5] Brown University | Department of Cognitive, Linguistic and Psychological Sciences
[6] Haskins Laboratories

Contact:
sahil.luthra@uconn.edu (S. Luthra, corresponding author)
pamela.fuhrmeister@uconn.edu (P. Fuhrmeister)
peter.molfese@nih.gov (P. J. Molfese)
s.guediche@bcbl.eu (S. Guediche)
sheila_blumstein@brown.edu (S.E. Blumstein)
emily.myers@uconn.edu (E.B. Myers)

Keywords:
non-native phonetic perception
implicit learning
inferior frontal gyrus

Abstract

Research has implicated the left inferior frontal gyrus (LIFG) in mapping acoustic-phonetic input to sound category representations, both in native speech perception and non-native phonetic category learning. At issue is whether this sensitivity reflects access to phonetic category information *per se* or to explicit category labels, the latter often being required by experimental procedures. The current study employed an incidental learning paradigm designed to increase sensitivity to a difficult non-native phonetic contrast without inducing explicit awareness of the categorical nature of the stimuli. Functional MRI scans revealed frontal sensitivity to phonetic category structure both before and after learning. Additionally, individuals who succeeded most on the learning task showed the largest increases in frontal recruitment after learning. Overall, results suggest that processing novel phonetic category information entails a reliance on frontal brain regions, even in the absence of explicit category labels.

1. Introduction

Speech sounds have a complex internal structure, and in general, processing the fine-grained detail of these sounds relies on temporal brain regions such as the left superior temporal gyrus (LSTG; Desai, Liebenthal, Waldron, & Binder, 2008; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Mesgarani, Cheung, Johnson, & Chang, 2014; Myers, 2007). These temporal areas show tuning that is specific and structured according to the acoustic details of one's native language phonetic categories. However, a number of studies suggest that the perception of phonetic detail, even if largely supported by superior temporal cortex, is not entirely divorced from frontal brain regions. Individuals with Broca's aphasia, for instance, have shown subtle deficits in phoneme discrimination, though they make fewer errors than individuals with posterior brain damage (Blumstein, Baker, & Goodglass, 1977). This notion has also been supported by functional neuroimaging studies of native language perception, with frontal brain regions implicated in different aspects of acoustic-phonetic processing (Lee, Turkeltaub, Granger, & Raizada, 2012; Myers, 2007; Rogers & Davis, 2017; Xie & Myers, 2018). In particular, the left inferior frontal gyrus (LIFG) is sensitive to the proximity between an acoustic token and a phonetic category boundary (Myers, 2007; Myers, Blumstein, Walsh, & Eliassen, 2009) and responds to phonetic ambiguity in naturally-produced, continuous speech (Xie & Myers, 2018). While there are likely differentiable roles for frontal structures in the perception of speech, in general inferior frontal regions show evidence of abstraction away from low-level acoustic details in order to access category-level information about speech tokens (Chevillet, Jiang, Rauschecker, & Riesenhuber, 2013; Lee et al., 2012; Myers et al., 2009).

Further evidence for a role of frontal brain regions in speech perception comes from studies examining the acquisition of non-native phoneme categories. Non-native speech distinctions, especially those that are perceptually similar to existing native language categories, are very difficult to acquire in adulthood (Best & Tyler, 2007), with most adults falling short of native-like perceptual performance, even with targeted training (Golestani & Zatorre, 2009; Pruitt, Strange, Polka, & Aguilar, 1990; Strange & Dittmann, 1984). The extant research suggests that acquisition of new speech categories invokes processes in left frontal areas, among other neural systems. For

instance, Golestani and Zatorre (2004) showed that newly-learned non-native stimuli activated the bilateral IFG (*pars opercularis*) and LSTG relative to a noise baseline, and Myers and Swan (2012) showed that an area of the left middle frontal gyrus (MFG) immediately adjacent to Broca's area was sensitive to newly-acquired non-native category structure. One interpretation of these patterns is that non-native tokens activate emerging perceptual category information stored in the frontal lobe.

While several studies have shown frontal recruitment for non-native learning, evidence points to increased reliance on temporoparietal structures as listeners become more proficient (see Myers, 2014 for review). For instance, individual success in learning has been associated with reduced activation of LIFG (Golestani & Zatorre, 2004; Myers & Swan, 2012) and increased recruitment of temporoparietal regions such as the bilateral angular gyri (AG) (Golestani & Zatorre, 2004). These findings can be taken as evidence that listeners may initially recruit frontal regions to process non-native sounds but that as listeners develop better-elaborated representations of the novel phonetic categories, processing of these sounds may increasingly recruit temporal regions associated with sensory perception. Under such a view, the early reliance on frontal regions may reflect access to articulatory codes or abstract category-level representations that can be used to guide perception, or else may reflect high demands on phonological working memory (Callan, Jones, Callan, & Akahane-Yamada, 2004; Golestani & Zatorre, 2004; Myers, 2014).

The interpretation of the role of frontal areas for native as well as non-native speech perception is complicated because many studies examining phonetic learning have used explicit tasks during scanning, such as phoneme categorization (Callan et al., 2004; Golestani & Zatorre, 2004). What is not clear is whether category-relevant neural activation is driven by the metalinguistic demands of the task or by speech perception *per se*. Indeed, Hickok and Poeppel (2000, 2004) have argued that the involvement of frontal brain structures in perceiving acoustic-phonetic detail is limited to situations in which participants must explicitly attend to sub-lexical details of the stimulus, as is required in phoneme identification tasks.

Nonetheless, frontal recruitment for phonetic learning has been observed in the absence of an explicit task. In a study by Myers and Swan (2012), participants were

exposed to a dental-retroflex-velar continuum (i.e., d̪a-ɖa-ga) and trained to categorize stimuli into two categories. Half of the participants learned that the category boundary was between the dental and retroflex tokens, and for the other half of the participants, the category boundary was between the retroflex and velar tokens. A short-interval habituation design (Zevin & McCandliss, 2005) was used during scanning: On every trial, participants heard a train of identical stimuli followed by a distinct stimulus, which either came from the same phonetic category as the preceding stimuli or came from the other category. Notably, participants were not asked to identify the category for the tokens they heard and instead only responded to occasional high-pitched catch trials. The bilateral MFG showed sensitivity to the learned category structure, suggesting a role for frontal regions in perceiving non-native phonemic distinctions even in the absence of an explicit identification task. However, it is important to note that the Myers and Swan (2012) study did use an explicit categorization task during training, so it is possible that participants were categorizing stimuli during the fMRI scan, despite not being required to do so.

Indeed, the vast majority of studies examining the perception of non-native phonemes have used training tasks in which participants are explicitly taught a category label that corresponds to each stimulus. This explicit information about category identity may reinforce the early, frontally-mediated stages of non-native phonetic learning (Myers, 2014). That is, the frontal activation associated with non-native phonetic learning *may specifically reflect a mapping between stimuli and category labels*, rather than reflecting (bottom-up) sensitivity to the underlying acoustic-phonetic category structure. As such, a more stringent test of a role for frontal regions in non-native phonetic learning would require the use of implicit paradigms during both the training and fMRI portions of the study, such that participants do not have labels for the categories being learned and therefore cannot categorize the stimuli, even implicitly.

In recent years, researchers have increasingly utilized implicit paradigms to train participants on novel categories. For instance, Leech, Holt, Devlin and Dick (2009) examined the neural underpinnings of implicit auditory learning using complex non-speech stimuli. Over the course of several training sessions, participants played a video game where auditory cues were diagnostic of whether an upcoming visual exemplar

was a member of one category (aliens to be captured) or another (aliens to be shot). Pre- and post-training fMRI sessions utilized an implicit oddball detection task, meaning that neither behavioral training nor the scanner task entailed explicit categorization. Results showed that better auditory learning was associated with increased reliance on STS post-training. More recently, Lim, Fiez, and Holt (2019) measured BOLD activity while participants played this incidental learning video game in the MRI scanner. The authors manipulated whether the non-speech auditory exemplars were organized into linearly separable categories (structured categories) or not (unstructured categories). Critically, the time course of activation in the basal ganglia – and more specifically, in the striatum – differed between structured and unstructured categories, consistent with a proposed role for the striatum in acquiring new behaviorally-relevant sound categories (Lim, Fiez, & Holt, 2014; Yi, Maddox, Mumford, & Chandrasekaran, 2016). While the authors focused their discussion on the striatum, this same pattern was also observed in a number of additional regions including the bilateral IFG. Further, striatal activity was positively correlated with changes in behavior and functionally connected to superior temporal sulcus. Taken together, such results suggest the involvement of a coordinated network of frontal, striatal, and temporal areas in auditory category learning, at least for non-speech sounds.

In general, incidental or implicit learning paradigms can yield successful non-native learning (Gabay & Holt, 2015; Lim & Holt, 2011), showing that consistent associations between category information and behaviorally relevant stimulus properties can increase sensitivity to novel sound distinctions. Vlahou, Protopapas, and Seitz (2012) used an incidental training paradigm to examine learning of two different sound categories. Native speakers of Greek heard two pairs of speech sounds (four sounds total) on every trial and were asked to identify whether tokens within the first pair or second pair differed in volume. Unbeknownst to subjects, one pair always consisted of two Hindi dental sounds while the other consisted of two Hindi retroflex sounds. Critically, the volume difference emerged only within the retroflex pair (i.e., the correct response always corresponded to the retroflex category). To ensure the task was appropriately challenging, the size of the volume difference within the retroflex pair was set adaptively, such that the task got harder (i.e., the volume difference got smaller) if

participants succeeded on easier levels. Following training, subjects' discrimination and identification abilities were tested explicitly. Vlahou and colleagues found that participants who completed the incidental learning task performed as well as or better than a group who received explicit training on the speech sounds, and both groups performed better than a group of naïve listeners. Thus, even though the incidental learning task itself did not require learning of the non-native phonemic contrast, the consistent temporal yoking of category-level information (the phonetic category difference) with a behaviorally relevant dimension (the volume difference) resulted in learning, consistent with other similarly structured studies of incidental learning (Seitz & Watanabe, 2005).

The aim of the current study is to examine the neural systems underlying the learning of a non-native phonetic category distinction using an incidental speech sound learning paradigm, specifically testing whether frontal regions are involved in non-native phonetic category learning in the absence of explicit category labels. In Experiment 1, we leveraged the incidental learning paradigm used by Vlahou et al. (2012) to promote non-native learning of the Hindi dental-retroflex contrast. Functional activation was measured with fMRI both before and after three days of incidental learning, allowing us to examine whether frontal brain regions are recruited for processing phonetic detail when participants are not explicitly aware that they are being exposed to two novel speech sound categories. In Experiment 2, we examined the extent to which behavioral gains over the course of the incidental learning sessions depend on consistent associations between the phonetic category structure and the task-relevant changes in volume.

2. Experiment 1

In Experiment 1, we collected fMRI data to measure changes in brain activity that occur after three days of an incidental learning task designed to induce sensitivity to a non-native phonetic category difference. Crucially, participants were not informed of the categorical structure of the stimuli until after all scanning was completed, at which point their sensitivity to the non-native phonetic category structure was assessed explicitly.

2.1. Methods

2.1.1. Participants

Twenty participants who self-identified as right-handed were recruited from the University of Connecticut campus community. All participants were native speakers of American English with no history of speech, language, hearing or neurological impairments. All participants received monetary compensation for participating. Due to experimenter error, two participants received inconsistent condition assignments during the incidental learning task and were therefore excluded from analyses, resulting in a final *n* of 18 (age: mean = 22.6, SD = 2.2; 14 female). All procedures were approved by the University of Connecticut Institutional Review Board, and all subjects provided informed consent prior to participating.

2.1.2. Stimuli

Voiced dental (/d̪ɑ/) and retroflex (/ɖɑ/) stops had been recorded for a previous study (Earle & Myers, 2015) by a native speaker of Hindi. For each type of stop, five unique productions were edited to the onset of the burst, equated for length and scaled to 70dB SPL mean amplitude. For each token, volume variants were created in Praat (Boersma & Weenink, 2017) by scaling the tokens to 70.5 dB, 71 dB, 72 dB, 73 dB and 74 dB. Additionally, a high-pitched variant was created for each token by increasing the F0 contour by 100 Hz in Praat.

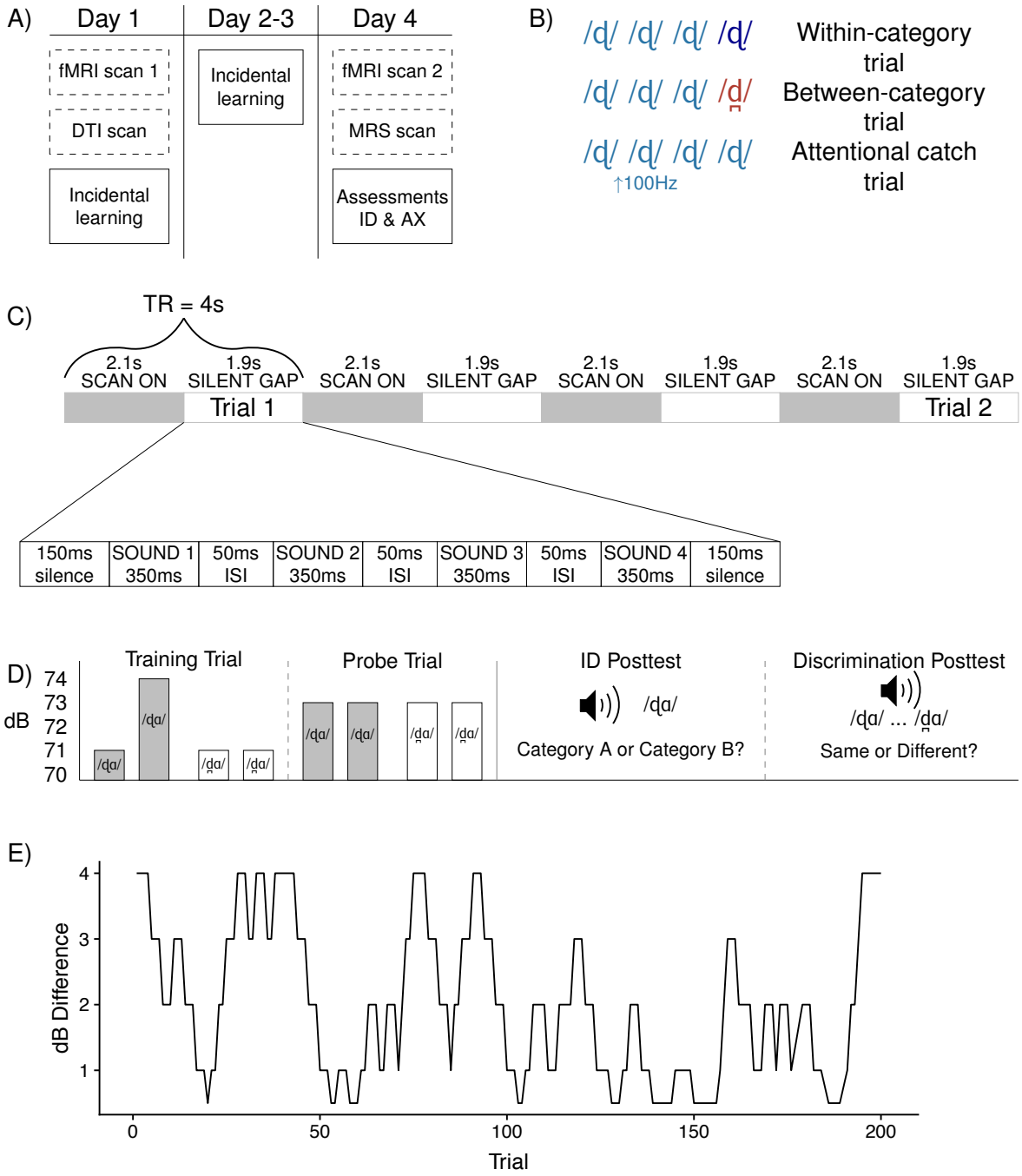

Figure 1. Procedure. **(A)** The full procedure took place over four days. In both Experiment 1 and Experiment 2, participants completed three days of incidental behavioral training and received explicit behavioral assessments (an identification task and an AX discrimination task) on the fourth day. Participants in Experiment 1 also completed MRI sessions (shown in dashed boxes) on Days 1 and 4. **(B)** A short-interval habituation task was used during scanning. On *within-category* trials, participants first heard three identical stimuli and then an acoustically distinct token from the same

phonetic category. On *between-category* trials, participants heard three identical stimuli and then a stimulus from the contrastive phonetic category. Participants did not make overt responses to either within-category or between-category trials; instead, they only responded to occasional *attentional catch* trials, in which one of the four tokens in the stimulus train was presented at a higher pitch (raised by 100 Hz). **(C)** A sparse sampling design was used for the scanner task, whereby stimuli were presented in silent gaps that fell between scans. **(D)** To induce sensitivity to the phonetic category distinction, the incidental learning task used *training trials* in which participants heard four speech sounds and had to decide whether the first pair or second pair contained a volume difference. Unbeknownst to participants, one pair consisted of two dental tokens and the other of two retroflex tokens. In Experiment 1, the volume difference was consistently associated with one phonetic category, and Experiment 2 tested whether learning depended on this consistent association. The second half of each learning session also included some *probe trials*, in which neither pair contained a volume difference; of interest was whether participants would make their selection in line with previous trials. On the fourth day of the experiment, participants completed ID and discrimination posttests, as shown. **(E)** The incidental learning sessions used an adaptive staircase structure. Participants needed to respond correctly to three consecutive trials before they moved to the next difficulty level (smaller volume difference), and an incorrect response moved them back to the previous difficulty level (larger volume difference). A sample participant's trajectory is shown here; a threshold value is computed as the average of the inflection points on the staircase within a block of 50 trials. Note that it should be very challenging to detect very small volume differences based on volume information alone, so it behooves participants to leverage the consistent association between phonetic category information and volume information to succeed on the volume task. As such, we assume that lower thresholds on the volume task reflect incidental learning of the phonetic category distinction.

2.1.3. Procedure

The experiment took place over four consecutive days; the full schedule of tasks is displayed in Figure 1A, and the procedure for each task is described in detail below. On the first day, subjects participated in an initial fMRI scan and then completed their first session of the incidental learning task. On days 2 and 3, participants completed additional sessions of the incidental learning task. On day 4, participants completed another fMRI scan, after which they were informed that they had been exposed to two speech sounds from Hindi. At this point, a behavioral posttest was conducted to explicitly assess identification and discrimination for the two categories.

fMRI sessions

Scanning took place on a 3-T Siemens Prisma using a 64-channel head coil. Anatomical images were acquired sagittally using a T1-weighted magnetization-prepared rapid acquisition gradient echo (MP-RAGE) sequence (TR = 2300 ms, TE = 2.98 ms, FOV = 256 mm, flip angle = 9 degrees, voxel size = 1 mm x 1 mm x 1 mm). Diffusion-weighted images were acquired during the first scan session and magnetic resonance spectroscopy data were collected during the second session; those data are not presented here. Functional images were acquired with a T2*-weighted EPI sequence using a slow event-related design (TR = 4.0 seconds [2.1 seconds scan with a 1.9 sec delay], TE = 25 ms, FOV = 192 mm, flip angle = 90 degrees, slice thickness = 2.5 mm, in-plane resolution = 2 mm x 2 mm). Thirty-six slices were acquired per TR, and slices were acquired in an interleaved, ascending fashion. A sparse sampling method was used to ensure that auditory information was presented during silent intervals between scans (Edmister, Talvage, Ledden, & Weisskoff, 1999; Figure 1C), and trials were presented during the silent gap every 3 TRs, yielding a stimulus onset asynchrony of 12 seconds. In each scan session, participants completed five functional runs of 36 trials each; 110 volumes were acquired per run, and each run lasted approximately 7.5 minutes.

A short-interval habituation paradigm was used for the in-scanner task. On each trial, participants heard four stimuli in quick succession with a 50-ms ISI (Figure 1B). The first three tokens on a given trial were always repetitions of a single token. On *within-category* trials, the fourth token was a different production of the same syllable, spoken by the same speaker. On *between-category* trials, the fourth token was from the other phonetic category. Because subjects did not have to make judgments about the category membership of the stimuli, the design provided an implicit measure of neural responses to phonetic category information (between-category versus within-category) (Zevin & McCandliss, 2005). There were 80 between-category and 80 within-category trials in each scan session. Each session also included 20 attentional catch trials; on these trials, a high-pitched token replaced one stimulus in the train. Participants were instructed to press a button whenever they heard a high-pitched stimulus. Two versions of the scanner task were created, with subjects receiving one version during their first

scan and the other during the second (with order counterbalanced). Participants also completed a set of 15 practice trials before each session.

Incidental learning sessions

The incidental learning task was programmed in MATLAB (The Mathworks Inc., Natick, MA, USA) using Psychtoolbox (Brainard & Vision, 1997). Each day's session took approximately 20 minutes to complete and consisted of 200 trials, presented in four blocks of 50 trials. On each trial, participants heard two pairs of tokens and were asked to identify whether the tokens in the first pair or the tokens in the second pair differed in volume. Unbeknownst to the subject, one pair was composed of two dental tokens and the other was composed of two retroflex tokens (Figure 1D).

Critically, the volume difference was always associated with the same non-native phoneme (counterbalanced across participants). The volume difference could be either 4 dB, 3 dB, 2 dB, 1 dB or 0.5 dB, with the precise difference determined by a 3-down-1-up adaptive staircase (Figure 1E), such that if a subject correctly identified the target pair on three consecutive trials, the volume difference was reduced by one step. The volume difference was made larger after any incorrect response (e.g., it was raised to a 4-dB difference if the incorrect response occurred on a trial with a 3-dB difference). In this way, task difficulty was modulated by the subject's performance on the task. The rationale for this design is that because the task-relevant dimension (the volume difference) was consistently associated with phonetic information, it would enhance learning of the phonetic category information, particularly as the volume difference became smaller.

The quieter token (always 70 dB) was presented first for half the target pairs, and the louder token was presented first for the remaining half. The amplitude of non-target stimuli was consistent with the subject's place on the adaptive staircase. For instance, if a subject was on a 3dB trial, the non-target stimuli were either both at 73dB or both at 70dB. Within each pair, the same dental or retroflex token was always presented, even if the amplitude differed; across all trials, all dental and all retroflex tokens were presented. There were an equal number of retroflex-first trials as dental-first trials, with order held constant across subjects. Tokens within a pair were separated by a mean ISI

of 250 ms (SE = 3 ms), and pairs of tokens were separated by a mean ISI of 500 ms (SE = 3 ms).

To assess whether participants were relying on phonetic cues, ten *probe trials* were scattered throughout the second half of each session (Figure 1D). For these trials, neither pair contained a volume difference; of interest was whether subjects would consistently choose the pair type on which they were being trained or whether they would be at chance in their responses. Unlike the results reported by Vlahou et al. (2012), no significant differences in probe trial performance were found, so detailed results are not reported here.

Behavioral posttest

All participants completed a behavioral posttest to assess their sensitivity to the dental-retroflex contrast; token amplitude was not manipulated during this posttest. The posttest began with an initial familiarization portion, during which participants heard each of the 10 tokens (5 from each category) paired with a category label (D1 or D2). Participants then completed a singleton identification task in which they heard one token at a time and were asked to categorize the stimulus as belonging to D1 or D2. Each participant received ten trials for each token, resulting in 100 total trials; the same random order was used for all participants. Finally, participants completed a pair discrimination task in which they heard two tokens and were asked whether they came from the same category or different categories. There were 100 total discrimination trials, allowing each possible trial combination to be presented once, and the same random order was used for all participants. No feedback was given during the identification or discrimination trials.

2.1.4. fMRI analyses

Neuroimaging data were analyzed in AFNI (Cox, 1996). Because we obtained oblique data for our functional runs, EPI data were first rotated to cardinal orientation to match the coordinates of the anatomical data. Preprocessing was done separately for each run. In particular, an afni_proc.py script was used to register functional volumes to the first volume of each run, to align EPI data to the anatomical data, and to align

functional data with AFNI's Colin27 template in Talairach & Tournoux (1988) space; these transformations were done in a single warp to minimize interpolation. Data were smoothed using a 4-mm full-width half-maximum Gaussian kernel and scaled to a mean of 100 for each run to represent percent signal change.

To conduct univariate analyses, idealized hemodynamic response functions (HRFs) for each condition (between-category, within-category, attentional catch trials) were created by convolving stimulus onset times with a gamma function; trial onsets were measured from the start of the first token in the stimulus train. For each subject, preprocessed BOLD data were submitted to a deconvolution analysis with the condition HRFs (between, within, and catch trials) and six rigid-body motion parameters as regressors; false alarm trials (in which participants had made a button response to a trial without a pitch change) were censored (i.e., those rows were removed from the regression matrix). Using the AFNI program 3dMVM (Chen, Adleman, Saad, Leibenluft, & Cox, 2014), beta coefficients were submitted to two group-level ANCOVAs with *Session* (Pre / Post) and *Phonetic Category* (Between / Within) as categorical factors. Individual *Threshold* scores from the final day of training were used as a continuous covariate in one ANCOVA (as these scores are assumed to reflect the degree of learning in the incidental learning task), and another ANCOVA looked for relationships with individual post-test identification performance.[1]

A group mask containing only voxels that were imaged in all the participants was applied at the ANCOVA stage. Subsequently, a small volume correction was applied, limiting analyses to regions known to be involved in language processing (bilateral IFG, MFG, insula, supramarginal gyri, AG, STG, middle temporal gyri, and transverse temporal gyri) as well as the striatum, a subcortical region thought to be involved in incidental learning of auditory categories (e.g., Lim et al., 2019). These regions were defined using the Talairach and Tournoux (1988) atlas built into AFNI; the set of voxels considered is shown in Figure 3A. To correct for multiple comparisons, we first

[1] We also conducted an analysis that used mean d' scores from the *Discrimination* task as a continuous covariate. We observed similar results for the effects of *Session* and *Phonetic Category*, as expected. However, no effects of *Discrimination* performance were observed, and so this analysis is not discussed further.

estimated the noise smoothness for each subject by applying the 3dFWHMx command to the residual time series; notably, we used a mixed autocorrelation function as suggested by Cox, Chen, Glen, Reynolds, and Taylor (2017) in order to address concerns about Type I error raised by Eklund, Nichols, and Knuttson (2016). Estimated smoothness values were averaged across subjects, and these averages were used in Monte Carlo simulations to estimate the likelihood of noise-only clusters. Simulations indicated that a cluster size of 218 voxels was needed at a voxel-level significance of $p < 0.05$ to yield cluster-level threshold of $\alpha < 0.05$.

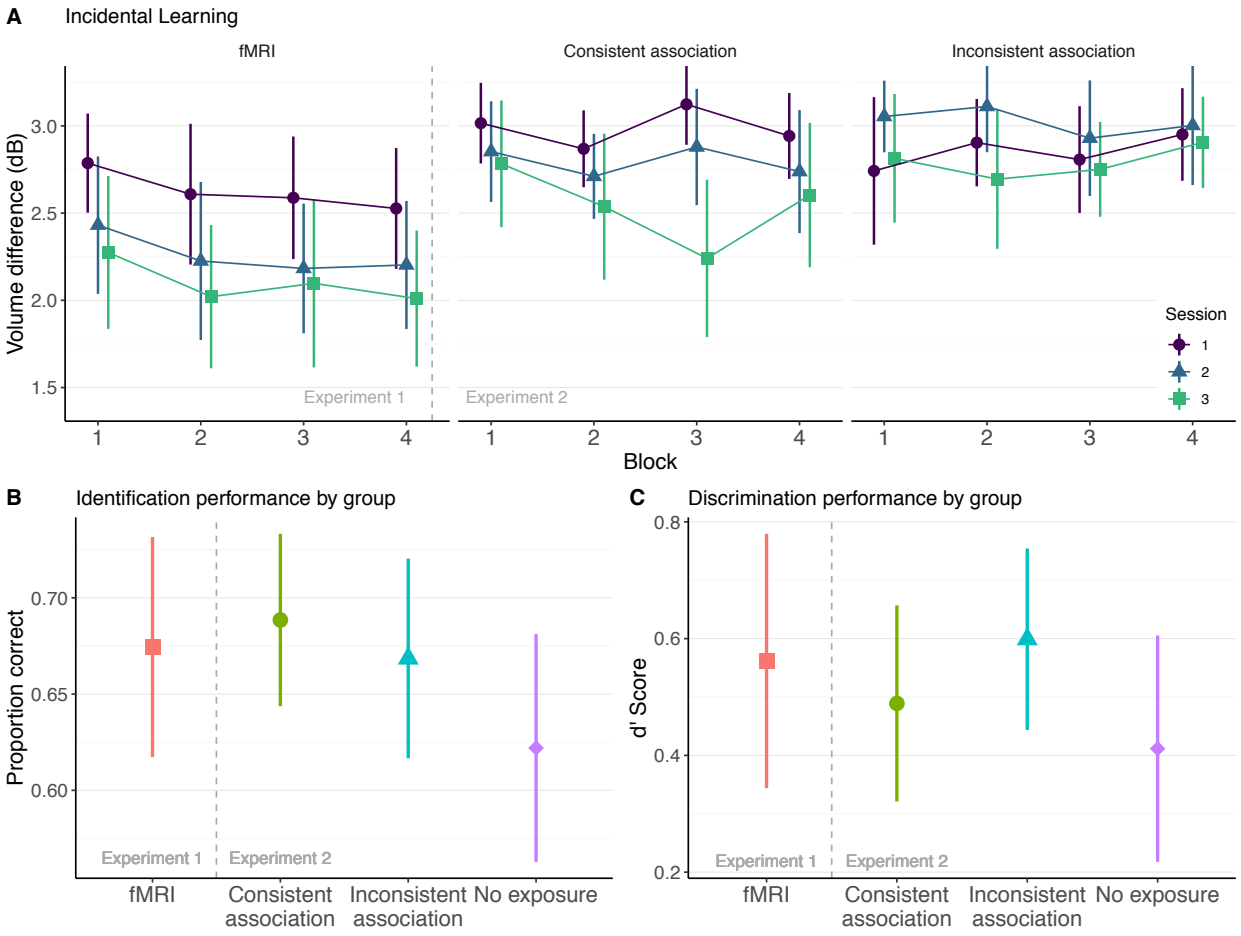

Figure 2. Behavioral results from Experiment 1 (left of dashed lines; fMRI group) and Experiment 2 (right of dashed lines; *consistent association*, *inconsistent association*, and *no exposure* groups). In all plots, error bars indicate 95% confidence intervals around the mean. **(A)** Performance on the incidental learning task was assessed by computing a threshold for each block of the volume task and examining how volume thresholds changed within and across sessions for each group; each session is shown

in a different color. The *no exposure* group in Experiment 2 did not complete this task and so no data are shown for this group. **(Lower panels)** Plots showing group-level performance on the identification **(B)** and discrimination **(C)** tasks for each group, with each group shown in a different color.

2.2. Results

2.2.1. Behavioral results

To examine the trajectory of learning during the incidental learning sessions, we calculated a dB threshold measure for each block of 50 trials by taking the average of the inflection points (dB level at a change in direction) on the adaptive staircase within that block (Figure 1E). If there were no inflection points within a block (i.e., if the participant never advanced beyond the initial 4 dB difficulty level), the modal difficulty level was taken as the threshold. Conceptually, this threshold measure estimates the smallest volume difference at which participants can reliably make a correct response. The threshold scores for each block are visualized in Figure 2A.[2] We then used a linear mixed effects model to estimate how threshold levels changed over time, implementing the model in R using the *mixed* function of the "afex" package (Singmann, Bolker, Westfall, & Aust, 2018). This model included a fixed factor of Block (mean-centered) nested within a fixed factor of Session (mean-centered); we also included random intercepts for each subject.[3] A likelihood ratio test yielded only a significant main effect of Session, $\chi^2(2) = 39.74$, $p < 0.0001$, suggesting that participants in Experiment 1 improved at the volume detection task from day to day.

[2] We opted to analyze threshold scores for each block rather than analyzing trial-by-trial data because of the considerable interdependence between consecutive trials. That is, due to the 3-down-1-up adaptive staircase, the dB level for a given trial depends on the dB level of several previous trials as well as a participant's accuracy on previous trials, therefore adding tremendous complexity to any model that attempts to estimate effects on subject performance. By using a threshold measure instead, we were able to analyze changes in performance over the course of the incidental learning sessions while respecting the structure of the data and facilitating interpretation of model results.

[3] Given our decision to compute a threshold score for each block, we were only justified in using random intercepts, as there would be insufficient data per cell (only four observations per session) to estimate random slopes.

Next, we analyzed participants' ability to identify the correct phonetic category for each speech sound during the behavioral posttest. To account for participants who may have been able to distinguish the speech sounds but confused the labels in the categorization phase, response labels were reversed if a subject's mean accuracy was at or below 0.41; this affected 3 participants. This criterion was selected because the binomial probability of obtaining a score of 0.41 or lower by chance was less than 5%. Performance on the identification task is visualized in Figure 2B. Participants in Experiment 1 had a mean accuracy of 0.67 (SE: 0.03). A one-sample t-test indicated that this was significantly above chance, $t(17) = 24.89$, $p < 0.001$.

We next considered participants' explicit discrimination of the non-native categories, as assessed on the behavioral posttest. To account for potential effects of response bias, percent accuracy scores were converted to d' scores (MacMillan & Creelman, 2004). Discrimination data are displayed in Figure 2C. Participants in Experiment 1 had an average d' of 0.56 (SE: 0.10), significantly above what would be expected by chance, $t(17) = 5.44$, $p < 0.001$.

Finally, we examined the relationship between subjects' performance on the various behavioral tasks. Correlation tests revealed that participants who performed well on the identification task also performed well on the discrimination task, $r = 0.72$, $t(16) = 4.21$, $p < 0.001$, and that participants who succeeded on the volume task (as measured by lower mean threshold scores on the third day) did better on the discrimination posttest, $r = -0.47$, $t(16) = -2.13$, $p = 0.05$. There was no significant correlation between success on the volume task and performance on the identification task, though the relationship was in the expected direction, $r = -0.35$, $t(16) = -1.50$, $p = 0.15$.

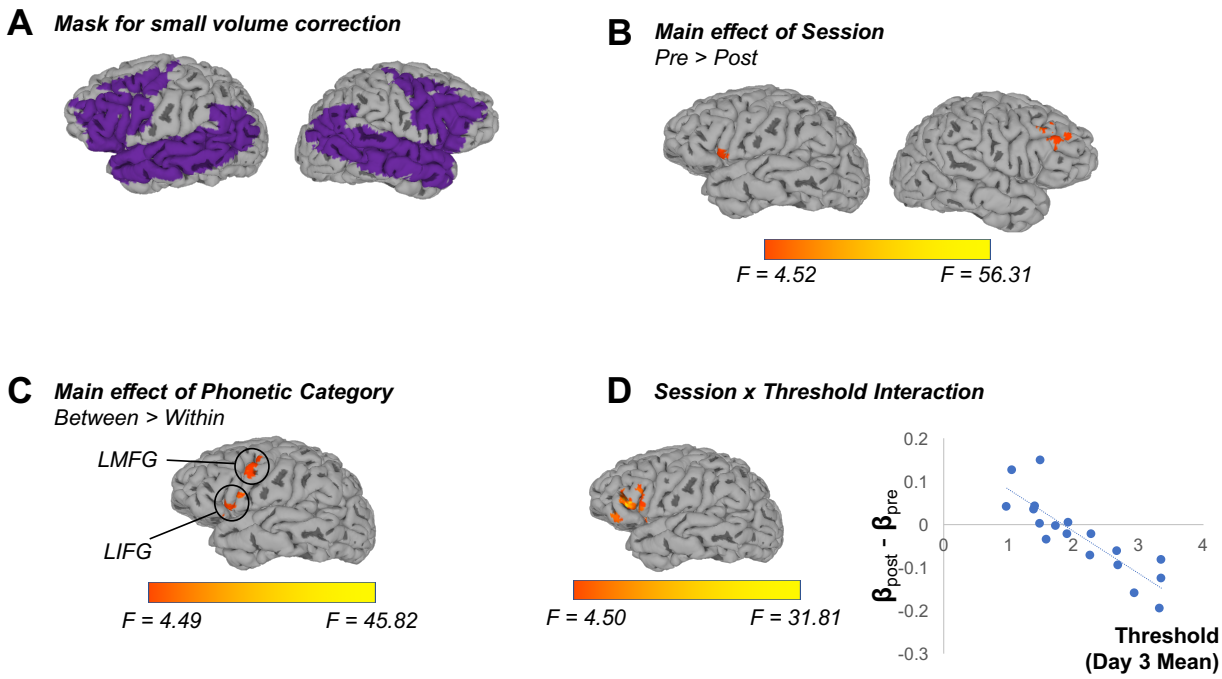

Figure 3. Results of fMRI analysis considering effects of Session, Phonetic Category and Threshold (our behavioral measure of performance during incidental learning). While a volumetric approach was used for statistical analyses, results are visualized on the anatomical surface of a single subject. FreeSurfer (Fischl, 2002) was used for surface reconstruction, and SUMA (Saad & Reynolds, 2012) was used to map volume-based statistical maps to the surface reconstruction. **(A)** Analyses were limited to a set of cortical regions known to be involved in language processing and the striatum, which has been implicated in the learning of auditory categories. **(B)** Frontal regions bilaterally showed greater activation on the first scan session than the second. **(C)** Between-category trials elicited greater activation in left frontal regions than did within-category trials. **(D)** An interaction between Session, Phonetic Category and Threshold emerged in left inferior frontal gyrus. This interaction is visualized in the associated scatterplot, with each data point indexing an individual subject; a trend line shows the general relationship. Threshold scores are plotted on the x-axis, while the y-axis indicates the change in activation from the first scan to the second one.

2.2.2. Univariate fMRI results

A univariate analysis examined potential effects of *Session* (pre-training vs post-training) and *Phonetic Category* (between-category trials vs within-category trials). To examine potential differences in activation due to individual differences in learning, mean *Threshold* scores from the final day of incidental learning were included as a

continuous covariate. Results are summarized in Table 1 and visualized in Figure 3; the full mask of voxels considered in analyses is shown in Figure 3A.

The MFG and IFG bilaterally were less active after training (main effect of *Session*, Figure 3B). We also observed that even though participants did not have explicit labels for the two categories when completing the scanner task, frontal regions – specifically, the LIFG and LMFG – were sensitive to phonetic category-level information, showing relatively more activation for between-category trials than within-category trials even before training (main effect of *Phonetic Category*, Figure 3C). Finally, we observed an interaction between *Session* and *Threshold*, whereby participants with the lowest final thresholds (i.e., those who succeeded most in the learning task, as measured on the third session) showed relatively stronger recruitment of left IFG/MFG after learning compared to before; by contrast, those who performed worst on the incidental learning task showed reduced activation of LIFG on the second scan compared to the first (Figure 3D).

In a parallel analysis, we considered effects of *Session* and *Phonetic Category*, using accuracy on the *ID* task as a continuous covariate. Results are summarized in Table 2 and visualized in Figure 4. As expected, comparable results were obtained for the main effects of *Session* and of *Phonetic Category*. This analysis also revealed a cluster in left temporoparietal cortex that was sensitive to overall accuracy on the identification task; however, this effect was driven by three outlier participants, and so we do not consider it further (Figure 4A). Finally, this analysis identified several clusters showing an interaction between *Session* and *ID*, such that participants who were most accurate on the identification task showed the most pronounced increases in recruitment of bilateral IFG and left insula after training as compared to before (Figure 4B).

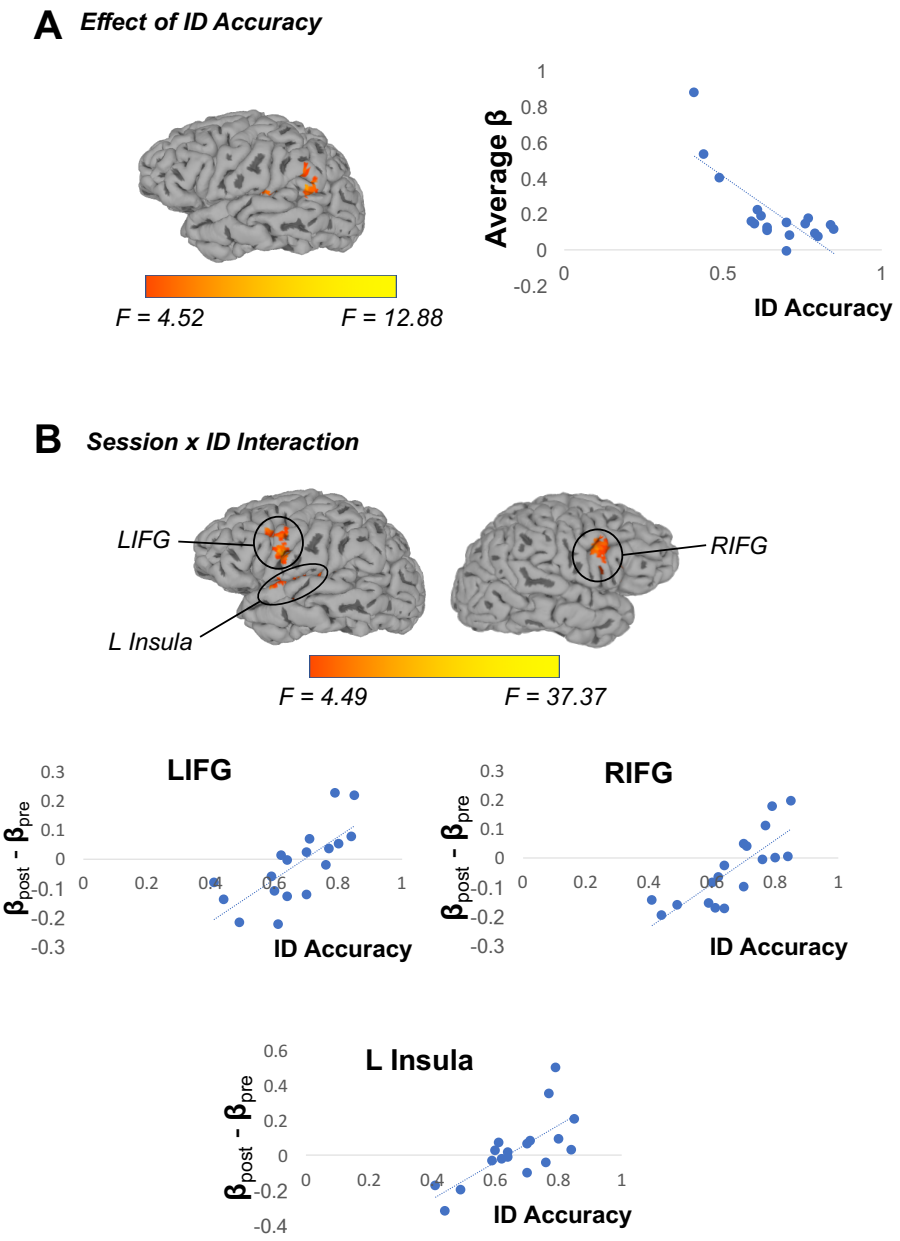

Figure 4. Results of fMRI analysis considering potential effects of Session, Phonetic Category and accuracy on the Identification task posttest (which was completed after the final scan). **(A)** Analyses revealed a left temporoparietal cluster where activation was negatively associated with overall performance on the identification task. However, this effect is driven by three outlier participants who showed below-chance accuracy on the task, and so we do not interpret this finding further. **(B)** Clusters in bilateral IFG and the left insula showed a Session x ID interaction, with participants who were relatively accurate on the identification task showing greater recruitment of these areas at the second scan compared to the first.

2.3. Discussion

2.3.1. Frontal contributions to novel perceptual category learning

The current results support other findings that frontal structures – particularly in the left inferior frontal gyrus and middle frontal gyrus – detect phonetic changes in passive, oddball-type paradigms (Myers & Swan, 2012; Myers et al., 2009). Of interest, this sensitivity to the category status of the tokens was present across sessions, as participants showed differential activation in response to a change in phonological category (for *between-category* dental-retroflex trials compared to *within-category* trials). Previous studies showing category-level sensitivity in the left frontal lobe have used well-established native language phonological categories (Chevillet, et al., 2013; Lee et al., 2012; Myers et al., 2009) or explicitly-trained non-native categories (Myers & Swan, 2012). This frontal sensitivity to phonetic category changes has been interpreted as reflecting access to information about the category status of tokens at a level abstracted from the acoustic input. Yet in the current study, the dental and retroflex stops are presumably both heard as variations on the alveolar /d/ for English-speaking listeners and should therefore activate the same phonological category, at least prior to incidental learning (e.g., the perceptual assimilation model, Best & Tyler, 2007). Put differently, prior to any learning, between-category tokens did not yet have any category status (implicit or otherwise) to map on to. Nonetheless, dental and retroflex tokens do differ from one another acoustically, and the acoustic differences encountered in between-category trials (i.e., the difference between a retroflex token and a dental token) are necessarily greater than the acoustic differences encountered in within-category trials (e.g., the difference between two dental tokens). The frontal response to phonetic category change may thus reflect a passive detection of auditory change (see Zevin, Yang, Skipper, & McCandliss, 2010), rather than necessarily activation of established phonological categories. We suggest that this response can still be modulated by experience, since lifetime exposure to native sounds or laboratory training on non-native sounds may change the salience of certain relevant (between-category) phonetic dimensions (see also Holt & Lotto, 2006). That is, frontal regions may be recruited in response to *any* auditory change, though the specific degree of recruitment

may depend on the *relevance* of changes along that particular auditory dimension, which may in turn vary among individuals depending on their individual language experiences.

We also observed a general reduction in frontal recruitment from the first scan session to the second one. However, the fact that there is often considerable variability in behavioral success makes it challenging to interpret group-level changes in activation over time without also accounting for behavioral performance. Indeed, recruitment of frontal brain regions has been shown to differ as a function of learning, with better learners tending to rely less on frontal brain regions over time (Golestani & Zatorre, 2004; Myers & Swan, 2012). We thus examined how frontal involvement was modulated by subjects' out-of-scanner behavioral performance. Critically, we observed increased recruitment of left frontal structures in those participants who consistently reached the hardest difficulty levels on the task used in training (the Session x Threshold interaction). The training paradigm was structured so that phonetic information (i.e., whether the pair was dental or retroflex) served as a redundant cue to the volume difference, so we infer that successful performance at the hardest levels of the volume task could result from implicit detection of the phonetic category differences. One interpretation of this finding is that listeners who capitalized on the phonological structure of the training developed emerging category sensitivity in frontal regions, consistent with results from more explicit training paradigms (Myers & Swan, 2012). Greater categorization success at post-test was associated with a similar pattern. Namely, we observed increased recruitment of bilateral frontal structures in the participants who performed best on the explicit identification task that took place after the final scan session (the Session x ID interactions). Taken together, the results suggest that participants who succeeded most on the behavioral tasks were also those who showed the largest increases in reliance on frontal brain structures.

These findings can be explained in the context of reverse hierarchy theory (Ahissar, Nahum, Nelken, & Hochstein, 2009). This theory proposes that rapid perception is based primarily on higher-level representations and that changes in perceptual encoding of fine-grained low-level detail emerge only over time. For speech sound learning, the anatomical correlate of this hypothesis is that sensitivity to novel

phonetic category distinctions is predicted to emerge first in more domain-general, non-sensory neural systems (i.e., frontal systems), and only over time does experience retune perceptually-sensitive regions (i.e., temporal regions) (see also Myers, 2014; Reetzke et al., 2018). Notably, we did not observe sensitivity to phonetic category structure in temporal regions. In light of previous work showing that temporal recruitment is tied to participants' degree of behavioral success (Leech et al., 2009), it may be the case that participants in the current study did not show temporal recruitment because they did not progress past relatively early stages of learning; this notion is also supported by the generally weak identification and discrimination abilities of our participants (Section 2.3.3).

2.3.2. Potential striatal contributions to category learning

Recent work has posited a role for the basal ganglia, and more specifically the striatum, in the acquisition of novel auditory categories (Lim et al., 2014; 2019; Yi et al., 2016). In particular, activity of the prefrontal cortex and the anterior dorsal striatum (the head of the caudate) have been linked to the use of a reflective, rule-based system for category learning, whereas engagement of the posterior dorsal striatum (the tail and body of the caudate as well as the putamen) has been linked to the use of a more procedural reflexive system, the latter which has been argued to be better suited to learning speech sound categories (Chandrasekaran, Yi, & Maddox, 2014; Lim et al., 2014; Yi et al., 2016). In this way, the striatum has been theorized to support the coordination between frontal and temporal regions in auditory category learning (Lim et al., 2014).

Notably, no significant effects were observed in the striatum in the present study. We suspect that this is attributable to the fact that incidental learning sessions were conducted outside the scanner, in contrast to studies where the learning task and scanning occur concurrently (e.g. Lim et al., 2019). Other studies have suggested that the activity in specific sub-regions of the striatum may depend on trial-by-trial performance feedback (Lim et al., 2014; Tricomi, Delgado, McCandliss, McClelland, & Fiez, 2006), which was not provided in the present study.

2.3.3. Behavioral success in incidental learning

Behavioral performance during the training task suggests that participants may have begun to learn the relevant dimensions for non-native category discrimination as they reached progressively harder difficulty levels (lower dB thresholds) on the volume change task. However, behavioral gains did not consistently generalize to behavioral success on the posttests, where there was considerable variability in subjects' behavioral performance. Indeed, many subjects performed at near-chance levels on the identification and discrimination tasks. The relatively inconsistent posttest performance of participants in Experiment 1 may be partly attributable to fatigue, as these participants completed the posttest assessment immediately after spending an hour in the MRI scanner. Furthermore, the unstructured exposure to the sounds that subjects encountered during the in-scanner sessions may have attenuated the overall amount of non-native phonetic learning in this group (Fuhrmeister & Myers, 2017). We suggest that future work examines how frontal recruitment relates to the ultimate level of behavioral success on non-native phonetic learning tasks, given previous work showing reduced reliance on frontal regions in better learners (in contrast to the current findings; Golestani & Zatorre, 2004; Myers & Swan, 2012) as well as theoretical accounts positing relatively greater reliance on temporoparietal regions as individuals' perceptual performance improves (e.g., Myers, 2014).

3. Experiment 2

While Experiment 1 supports a role for frontal brain regions in the development of sensitivity to non-native phonetic category structure, it is unclear how much of this is attributable to learning *per se*. The incidental learning paradigm used in Experiment 1 was adapted from a study conducted by Vlahou et al. (2012), who demonstrated that subjects who had completed incidental learning sessions were more sensitive to phonetic category structure than a group of naïve participants. Learning observed in the incidentally trained participants could be attributable to the structure of the incidental learning task, as Vlahou et al. suggested. However, it is also possible that the results reflect the fact that incidentally trained participants had more exposure to the stimuli than did the naïve group of listeners. That is, it is unclear whether behavioral gains

brought about by this incidental learning task are merely a result of exposure to the stimuli, or whether a consistent pairing of the volume discrimination with one of the sound categories is necessary for learning. Experiment 2 examines this issue directly.

3.1. Experiment 2: Methods

3.1.1. Participants.

Sixty adults (38 female) were recruited from the University of Connecticut. All subjects were monolingual native speakers of American English with no history of neurological, speech, hearing or language impairments. All participants received course credit or monetary compensation for their participation, and all provided informed consent prior to participating.

Participants were assigned to one of three groups (20 subjects per group). One group of participants completed the incidental learning task used in Experiment 1, in which phonetic category information was consistently associated with the task-relevant volume change. A second group completed the same protocol, but for these participants, the phonetic category information was inconsistently associated with the volume difference. Finally, a third group completed only the posttest assessments without any prior exposure to the stimuli.

3.1.2. Stimuli.

The same stimuli were used as in Experiment 1 apart from the pitch-shifted tokens, since participants in Experiment 2 did not complete the scanner task.

3.1.3. Procedure

Participants in the *consistent association* and *inconsistent association* groups participated in three lab training sessions and completed a behavioral posttest on the fourth session, with each session occurring on consecutive days. Subjects in the *no exposure* group completed only the behavioral posttest and were used as a baseline against which to compare the other two groups.

Participants in the *consistent association* group completed the same protocol as in Experiment 1, apart from completing the fMRI sessions. For participants in the

inconsistent association group, the protocol was identical with one key difference: The volume difference occurred within the retroflex pair for half of a subject's trials and within the dental pair for the other half. In this way, participants in the *consistent association* and *inconsistent association* groups both received equal exposure to the auditory tokens over the three incidental learning sessions, and both groups engaged in a challenging volume-discrimination task. However, individuals in the *inconsistent association* group were not able to take advantage of a systematic association between phonetic cues and volume to succeed on the training task. As such, the comparison between the *consistent association* and *inconsistent association* groups allows us to evaluate the extent to which consistent associations between the phonetic category distinction and the volume difference support incidental learning.

3.2. Experiment 2: Results

3.2.1. Incidental learning sessions

The threshold data from Experiment 2 are displayed in Figure 2A. Following our approach for analyzing threshold data in Experiment 1, a linear mixed effects model was used to assess group differences in threshold over time, allowing us to model both the fixed effects of interest and random variation between subjects.[4] In particular, we modeled fixed effects of Group (*consistent association*, *inconsistent association*) as well as fixed effects of Block (mean-centered) nested within Session (mean-centered). As before, we also modeled random intercepts for each subject. A likelihood ratio test revealed a main effect of Session, $\chi^2(2) = 21.71$, $p < 0.0001$, and a Group by Session interaction, $\chi^2(2) = 9.23$, $p = 0.002$. No other main effects or interactions were significant.

To unpack this Group by Session interaction, we ran separate mixed effects models for each group, dropping the fixed effect of Group but otherwise keeping the

[4] Here, we benefited from an additional advantage of mixed effects models, which is in how they handle missing data (Baayen, Davidson, & Bates, 2008). Due to a programming error on one of the testing computers, some data collected in the first few days were not saved for the first few subjects. This resulted in the loss of 8 sessions' worth of data from the *inconsistent association* group, distributed across 6 participants (13.3% of the all the *inconsistent association* data).

same model structure as in the omnibus model. For the *inconsistent association* group, a likelihood ratio test revealed no significant effects. In contrast, we obtained a main effect of Session for the *consistent association* group, $\chi^2(2) = 28.46$, $p < 0.0001$. The effect of Session indicates that participants in the *consistent association* group improved in their performance on the volume detection task from day to day, suggesting that these participants may have been implicitly relying on the phonetic category information in order to perform the task. By contrast, the *inconsistent association* group did not receive a systematic pairing between the volume difference and the phonetic category distinction and therefore could not rely on this information to improve performance on the volume detection task.

3.2.2. Behavioral posttests

Data from the identification task are visualized in Figure 2B; note that we applied the same label correction as in Experiment 1. Participants who received a *consistent association* between the volume difference and phonetic category information had a mean accuracy of 0.69 (SE: 0.01), those who received an *inconsistent association* had a mean accuracy of 0.67 (SE: 0.01), and those who had received *no exposure* prior to testing had a mean accuracy of 0.62 (SE: 0.01).

Label-corrected trial-by-trial data were submitted to a logistic mixed effects model using the *glmer* function of the "lme4" package (Bates, Maechler, Bolker, & Walker, 2015) to test for group differences. The fixed factor of Group (*consistent association*, *inconsistent association*, and *no exposure*) was dummy-coded with the *consistent association* set as the reference level; the model also included random by-subject intercepts. The model found that participants who received a *consistent association* between phonetic category information and volume differences performed marginally better than participants who received *no exposure* to the stimuli prior to the behavioral posttest, ß = -0.33, SE = 0.19, $z = -1.81$, $p = 0.07$. Further, the model found no significant difference between the performance of participants who received *consistent associations* and participants who received *inconsistent associations*, *ß* = -0.12, SE = 0.18, $z = -0.63$, $p = 0.53$.

Discrimination data are displayed in Figure 2C. Participants in the *consistent association* group had a mean d' of 0.48 (SE: 0.08), those in the *inconsistent association* group had a mean d' of 0.60 (SE: 0.07), and those in the *no exposure* group had a mean d' of 0.41 (SE: 0.09). To assess the baseline ability of naïve participants to discriminate the two categories, we conducted a one-sample *t* test on the d' scores of the *no exposure* group; results indicated that participants were able to discriminate the two categories at above-chance levels without any training, $t(19) = 4.44$, $p < 0.001$. We then submitted d' scores from all three groups to a linear regression with Group (*consistent association*, *inconsistent association*, and *no exposure*) as a between-subjects factor. As before, the factor of Group was dummy-coded with the *consistent association* group used as a reference level. There were no significant differences in discrimination ability between participants who received *consistent associations* during incidental learning and those who received *inconsistent associations*, ß = 0.11, SE = 0.12, $z = 0.94$, $p = 0.351$. Furthermore, there was no significant difference in the discrimination abilities of subjects who received *consistent associations* and those who received *no exposure* to the stimuli prior to the behavioral assessments, ß = -0.08, SE = 0.12, $z = -0.66$, $p = 0.510$.

Finally, we examined potential correlations between the behavioral measures in each group. For the *consistent association* group, there was a significant positive correlation between performance on the identification task and performance on the discrimination task, $r = 0.51$, $t(18) = 2.54$, $p = 0.02$. There was also a significant correlation in the expected direction between performance on the volume task (measured by mean thresholds on the third session) and performance on the discrimination task, $r = -0.54$, $t(18) = -2.74$, $p = 0.01$, but no significant correlation between performance on the volume task and performance on the identification task, $r = -0.04$, $t(18) = -0.17$, $p = 0.87$. For the *inconsistent association* group, there was a significant correlation between discrimination and identification scores, $r = 0.50$, $t(18) = 2.46$, $p = 0.02$. However, there was no significant correlation between performance on the volume task and performance on the discrimination task, $r = -0.13$, $t(16) = -0.53$, $p = 0.60$, and no significant correlation between performance on the volume task and performance on the identification task, $r = 0.03$, $t(16) = 0.11$, $p = 0.92$.

3.3. Experiment 2: Discussion

Experiment 1 leveraged an incidental learning task used by Vlahou et al. (2012) to induce sensitivity to a non-native phonetic category distinction. Vlahou et al. found that participants who had completed this incidental learning task later showed better discrimination and identification of these non-native speech sounds than did a group of naïve participants. In Experiment 2, we showed weak evidence in support of their finding, as participants who completed an incidental learning modeled after the one used by Vlahou et al. (i.e., one in which there were consistent associations between phonetic and volume information) performed marginally better than naïve participants on an identification task though not significantly better on a discrimination task. We note that this may be in part attributable to the specific stimulus set we used or our particular sample of participants, and additional work is needed to assess the utility of incidental learning paradigms for non-native phonetic learning. It may be that this paradigm would be more effective in conjunction with training tasks where participants can explicitly practice on the tasks that will ultimately be used to assess learning. Indeed, in a recent study of non-native phonetic learning, Wright, Baese-Berk, Marrone & Bradlow (2015) found that alternating between periods of stimulus exposure and periods of explicit practice with posttest tasks yielded better learning than did explicit practice alone.

Critically, it is unclear both from the original study by Vlahou et al. (2012) and from Experiment 1 how much behavioral gains in this paradigm are attributable to the consistent associations between phonetic category information and the task-relevant dimension (i.e., the volume difference). In Experiment 2, we therefore also examined the degree of learning in a group of participants who received *inconsistent associations* between the phonetic category information and the volume difference. We found that participants who received consistent associations between phonetic information and the volume difference performed better on the loudness judgment task, as measured by lower dB thresholds across sessions. Since the task demands were the same across groups, the difference in thresholds is not likely due to the amplitude difference itself; rather, participants who received consistent associations appear to be able to capitalize on the consistent phonetic category information in order to succeed on the volume task.

However, participants' success on the volume task was not predictive of their posttest discrimination abilities. Indeed, these two groups performed equally well on posttest assessments of identification and discrimination, suggesting that group differences observed by Vlahou et al. (2012) may be attributable to differences in overall exposure to the stimuli rather than to the development of fully-fledged phonetic categories.

4. Conclusions

Non-native phonetic category learning offers a model system for auditory category learning in general. Recent attention to the learning systems underlying this process suggests that multiple learning systems can be recruited for novel speech sound learning (Chandrasekaran, Yi, & Maddox, 2014), and incidental paradigms that allow listeners to discover the nature of the phonetic category without explicit feedback have shown promise, especially insofar as these paradigms may recruit systems that more closely resemble those used during category acquisition in nature. In many of these paradigms, as in our study, phonetic differences are linked probabilistically to a response type, and it is this pairing that is thought to increase the perceptual distance between similar-sounding phonetic categories. However, we suggest that further investigation is needed to verify this assumption, as our data showed that consistent stimulus-response pairings were not necessary for success at posttest (Experiment 2). The specific ingredients that afford best speech sound learning in incidental paradigms is a subject of active study — these may include the degree of attention to the stimuli (Francis & Nusbaum, 2002), the statistical structure of the input (Roark & Holt, 2018), and the timing and consistency of reward signals (Chandrasekaran et al., 2014; Seitz & Watanabe, 2005) among others.

A surprising result in the current study is that frontal regions differentiate within-category and between-category trials for naïve participants and without any need for category labels. Frontal recruitment for speech has often been attributed to difficult perceptual decisions (Binder, Liebenthal, Possing, Melder & Ward, 2004) or to accessing category-level codes (Myers et al., 2009) or articulatory codes (e.g., Wilson, Saygin, Sereno, & Iacoboni, 2004). However, the frontal activation in the current study cannot be attributed to these factors, since participants completed passive tasks during

training and scanning (only completing the explicit phonetic categorization tasks after their final fMRI session); furthermore, participants had neither knowledge of how to produce the dental and retroflex tokens nor any knowledge or their differing category status. A necessary caveat in interpreting these results is that because the incidental paradigm did not result in strong perceptual performance at the group level, participants may not have developed clear dental and retroflex categories. As such, the evolution of the frontal response to categorical differences, and the degree to which the processing burden begins to include temporoparietal areas, may differ substantially when learners acquire sounds in a more elaborated, naturalistic fashion. We suggest that future investigations consider using increasingly naturalistic paradigms to differentiate the core neural systems involved in phonetic category acquisition from task-specific effects.

Acknowledgments

This research was supported by NIH grant R01 DC013064 to EBM and NIH NIDCD Grant R01 DC006220 to SEB. The authors thank F. Sayako Earle for assistance with stimulus development; members of the Language and Brain lab for help with data collection and their feedback throughout the project; Elisa Medeiros for assistance with collection of fMRI data; Paul Taylor for assistance with neuroimaging analyses; and attendees of the 2016 Meeting of the Psychonomic Society and the 2017 Meeting of the Society for Neurobiology of Language for helpful feedback on this project. We also extend thanks to two anonymous reviewers for helpful feedback on a previous version of this manuscript.

References

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1515), 285-299.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Bates, D., Maechler, M., Bolker, B, & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, *1334*, 1-47.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, *7*(3), 295-301.

Blumstein, S. E., Baker, E., & Goodglass, H. (1977). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia*, *15*(1), 19-30.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.21). Available from http://www.praat.org/.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory / orosensory internal models. *NeuroImage*, *22*(3), 1182-1194.

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488-495.

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., & Cox, R. W. (2014). Applications of multivariate modeling to neuroimaging group analysis: A comprehensive alternative to univariate general linear model. *NeuroImage*, *99*, 571-588.

Chevillet, M. A., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience*, *33*(12), 5208-5215.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*, 162–173.

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017). fMRI clustering and false-positive rates. *Proceedings of the National Academy of Sciences*, *114*(17), E3370-E3371.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174-1188.

Earle, F. S., & Myers, E. B. (2015). Overnight consolidation promotes generalization across talkers in the identification of non-native speech sounds. *The Journal of the Acoustical Society of America*, *137*(1), EL91-EL97.

Edmister, W. B., Talavage, T. M., Ledden, P. J., & Weisskoff, R. M. (1999). Improved auditory cortex imaging using clustered volume acquisitions. *Human Brain Mapping*, *7*(2), 89-97.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900-7905.

Fischl, B. (2012). Freesurfer. *NeuroImage, 62*(2), 774–781.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349-366.

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448-EL454.

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex*, *73*, 131-143.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage*, *21*(2), 494-506.

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2), 55-67.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67-99.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059-3071.

Lee, Y. S., Turkeltaub, P., Granger, R., & Raizada, R. D. (2012). Categorical speech processing in Broca's area: An fMRI study using multivariate pattern-based analysis. *Journal of Neuroscience*, *32*(11), 3942-3948.

Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *Journal of Neuroscience*, *29*(16), 5234-5239.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*(10), 1621-1631.

Lim, S. J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience*, *8*, 230.

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, *116*(10), 4671-4680.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390-1405.

Macmillan, N., & Creelman, C. (2004). *Detection Theory: A User's Guide*. New York, NY: Psychology Press.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006-1010.

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia*, *45*(7), 1463-1473.

Myers, E. B. (2014). Emergence of category-level sensitivities in non-native speech sound learning. *Frontiers in Neuroscience*, *8*, 238.

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, *20*(7), 895-903.

Myers, E. B., & Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *Journal of Cognitive Neuroscience*, *24*(8), 1695-1708.

Pruitt, J. S., Strange, W., Polka, L., & Aguilar, M. C. (1990). Effects of category knowledge and syllable truncation during auditory training on Americans' discrimination of Hindi retroflex-dental contrasts. *The Journal of the Acoustical Society of America*, *87*(S1), S72-S72.

Reetzke, R., Xie, Z., Llanos, F., & Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, *28*(9), 1419-1427.

Roark, C. L., & Holt, L. L. (2018). Task and distribution sampling affect auditory category learning. *Attention, Perception, & Psychophysics*, *80*(7), 1804-1822.

Rogers, J. C., & Davis, M. H. (2017). Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. *Journal of Cognitive Neuroscience*.

Saad, Z. S., & Reynolds, R. C. (2012). Suma. *Neuroimage, 62*(2), 768–773.

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

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of Factorial Experiments. R package version 0.21-2. https://CRAN.R-project.org/package=afex

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131-145.

Talairach, J., & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: An approach to cerebral imaging.

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*(6), 1029-1043.

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of non-native speech stimuli. *Journal of Experimental Psychology: General*, *141*(2), 363-381.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701-702.

Wright, B. A., Baese-Berk, M. M., Marrone, N., & Bradlow, A. R. (2015). Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. *The Journal of the Acoustical Society of America*, *138*(2), 928-937.

Xie, X., & Myers, E. (2018). Left inferior frontal gyrus sensitivity to phonetic competition in receptive language processing: A comparison of clear and conversational speech. *Journal of Cognitive Neuroscience*, *30*(3), 267-280.

Yi, H. G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, *26*(4), 1409-1420.

Zevin, J. D., & McCandliss, B. D. (2005). Dishabituation of the BOLD response to speech sounds. *Behavioral and Brain Functions*, *1*, 4.

Zevin, J. D., Yang, J., Skipper, J. I., & McCandliss, B. D. (2010). Domain general change detection accounts for “dishabituation” effects in temporal–parietal regions in functional magnetic resonance imaging studies of speech perception. *Journal of Neuroscience, 30*(3), 1110-1117.

Table 1

Results of analysis considering Session, Phonetic Category and Threshold. Coordinates and F-value correspond to peak activation in cluster. Approximate Brodmann areas are given in parentheses.

Anatomical region	Maximum intensity coordinates			Number of activated voxels	*F* value
	x	y	z		
Session (Post-Pre)					
1. Left inferior frontal gyrus (BA 45) / Left insula (BA 13)	-37	17	8	257	56.31
2. Right middle frontal gyrus (BA 9)	45	27	34	239	25.80
Phonetic Category (Between-Within)					
1. Left inferior frontal gyrus (BA 47)	-43	39	-12	356	45.82
2. Left middle frontal gyrus (BA 9)	-45	11	24	226	12.65
Session x Phonetic Category					
No significant clusters					
Threshold					
No significant clusters					
Session x Threshold					
1. Left inferior frontal gyrus (BA 45) / Left middle frontal gyrus (BA 46)	-43	33	8	430	31.81
Phonetic Category x Threshold					
No significant clusters					
Session x Phonetic Category x Threshold					
No significant clusters					

Table 2

Results of analysis considering Session, Phonetic Category and ID Accuracy. Coordinates and F-value correspond to peak activation in cluster. Approximate Brodmann areas are given in parentheses.

Anatomical region	Maximum intensity coordinates			Number of activated voxels	*F*-value
	x	y	z		
Session (Post-Pre)					
1. Left middle frontal gyrus (BA 46)	-31	35	30	291	21.31
2. Left inferior frontal gyrus (BA 45) / Left insula (BA 13)	-37	17	8	280	44.06
3. Right middle frontal gyrus (BA 9)	43	27	34	242	27.66
Phonetic Category					
1. Left inferior frontal gyrus (BA 47)	-43	39	-12	368	36.00
2. Left middle temporal gyrus (BA 21)	-55	-13	-14	288	22.38
3. Left middle frontal gyrus (BA 9)	-51	13	44	241	18.77
ID					
1. Left superior temporal gyrus (BA 42) / Left supramarginal gyrus (BA 40) / Left transverse temporal gyrus (BA 41)	-55	-41	22	245	12.88
Session x ID					
1. Left inferior frontal gyrus (BA 44) / Left middle frontal gyrus (BA 9)	-49	11	20	348	37.37
2. Right inferior frontal gyrus (BA 45) / Right middle frontal gyrus (BA 9)	49	11	26	327	19.64
3. Left insula (BA 13)	-45	-7	10	319	24.33
Session x Phonetic Category					
No significant clusters					
Phonetic Category x ID					
No significant clusters					
Session x Phonetic Category x ID					
No significant clusters					