Functional brain outcomes of L2 speech learning emerge during sensorimotor transformation

Daniel Carey[1,2,3], Marc E. Miquel[4,5], Bronwen G. Evans[6], Patti Adank[6], Carolyn McGettigan[1,2,7]

1: Department of Psychology, Royal Holloway, University of London, TW20 0EX, UK

2. Combined Universities Brain Imaging Centre, Royal Holloway, University of London, TW20 0EX, UK

3: The Irish Longitudinal Study on Ageing (TILDA), Dept. Medical Gerontology, TCD, Dublin, Irl.

4: William Harvey Research Institute, Queen Mary, University of London, EC1M 6BQ, UK

5: Clinical Physics, Barts Health NHS Trust, London, EC1A 7BE, UK

6: Department of Speech, Hearing & Phonetic Sciences, University College London, WC1E 6BT, UK

7: Institute of Cognitive Neuroscience, University College London, WC1N 3AR, UK

Corresponding author:

Dr. Carolyn McGettigan,
Department of Psychology, Royal Holloway, University of London,
Egham, TW20 0EX, UK
Email: Carolyn.McGettigan@rhul.ac.uk

Pages: 54

Figures: 4 **Tables:** 1

Supplemental figures: 3 **Supplemental tables:** 2

Keywords: Speech; Sensorimotor transformation (ST); fMRI; L2 Learning; Generalization.

The authors declare no competing financial interests. The work was funded by the Economic and Social Research Council (grant number ES/L01257X/1).

Abstract

Sensorimotor transformation (ST) may be a critical process in mapping perceived speech input onto non-native (L2) phonemes, in support of subsequent speech production. Yet, little is known concerning the role of ST with respect to L2 speech, particularly where learned L2 phones (e.g., vowels) must be produced in more complex lexical contexts (e.g., multi-syllabic words). Here, we charted the behavioral and neural outcomes of producing trained L2 vowels at word level, using a speech imitation paradigm and functional MRI. We asked whether participants would be able to faithfully imitate trained L2 vowels when they occurred in non-words of varying complexity (one or three syllables). Moreover, we related individual differences in imitation success during training to BOLD activation during ST (i.e., pre-imitation listening), and during later imitation. We predicted that superior temporal and peri-Sylvian speech regions would show increased activation as a function of item complexity and non-nativeness of vowels, during ST. We further anticipated that pre-scan acoustic learning performance would predict BOLD activation for non-native (vs. native) speech during ST and imitation. We found individual differences in imitation success for training on the non-native vowel tokens in isolation; these were preserved in a subsequent task, during imitation of mono- and trisyllabic words containing those vowels. fMRI data revealed a widespread network involved in ST, modulated by both vowel nativeness and utterance complexity: superior temporal activation increased monotonically with complexity, showing greater activation for non-native than native vowels when presented in isolation and in trisyllables, but not in monosyllables. Individual differences analyses showed that learning versus lack of improvement on the non-native vowel during pre-scan training predicted increased ST activation for non-native compared with native items, at insular cortex, pre-SMA/SMA, and cerebellum. Our results hold implications for the importance of ST as a process underlying successful imitation of non-native speech.

Highlights

1. We trained monolingual English speakers to imitate a non-native vowel (/y/);
2. Participants then imitated the trained vowel within novel non-words;
3. Substantial individual differences in imitation accuracy emerged;
4. Performance at pre-scan training predicted activation increase or decrease during ST;
5. Vowel nativeness and non-word complexity modulated superior temporal activity during ST.

Introduction

Producing speech in a non-native language requires phonemes to be deployed flexibly in a variety of lexical contexts (e.g., Flege & Hillenbrand, 1984; Levy & Law, 2010). Yet the complexity of sensory and articulatory behavior – particularly in the case of multisyllabic utterances – may pose considerable challenges to the faithful production of non-native (L2) phonemes within words (Segawa et al., 2015; Klein et al., 2006). In particular, L2 learners are faced with the non-trivial demands of perceiving and parsing the incoming speech acoustic signal, matching perceived phonemes to targets within phonological memory, transforming from phonological to motor targets, selecting and executing appropriate speech articulatory plans, and relaying auditory and somatosensory feedback in order to correct speech errors online (Hickok, 2012; see also Guenther, 2006; Bohland, Bullock & Guenther, 2010; Simmonds et al., 2014a; Cogan et al., 2014; Parker-Jones et al., 2013; Chang et al., 2013).

In light of these challenges, studies have begun to explore speech articulation and imitation, in tandem with earlier neural processes that reflect transformation from the perceived speech signal to the phonemic and motor representations that support speech articulation (i.e., sensorimotor transformation, ST; Cogan et al., 2014; Leonard et al., 2016; Parker-Jones et al., 2014; Carey et al., 2017). Recent electrocorticography evidence from native speech production points towards dissociable but related neural responses that show selectivity for the perceptual, phonological, memory and articulatory processes that underlie speech production (Leonard et al., 2016). In particular, regions of the superior temporal plane and peri-Sylvian cortex, in addition to sensorimotor cortex, have been implicated within networks that subserve ST (Cogan et al., 2014; Leonard et al., 2016; see also Cheung et al., 2016). Although advances have been made in understanding ST in native speech, comparably less work has explored ST processes with respect to L2, or changes in ST processes in the context of L2 learning.

While ST forms an important process that enables subsequent speech imitation behavior, relatively few studies have charted differences in the brain regions involved in ST when it precedes the production of non-native compared to native speech. Recent functional imaging studies of overt speech have nevertheless revealed differences in the modulatory roles of sensorimotor and superior temporal regions, when comparing native and non-native English speakers. In particular, sensorimotor suppression of posterior and anterior superior temporal regions occurs to a lesser degree in non-native than in native speakers when producing familiar English nouns (Parker-Jones et al., 2013). This strongly suggests that language experience can mediate the interplay between sensory and motor regions that subserve speech production (further to Simmonds et al., 2011a, 2011b; for covert articulation, see also Perani et al., 2003; for perception, see Callan et al., 2004). However, the role of L2 experience, particularly in the context of L2 learning, has yet to be investigated with respect to processes such as ST that likely support native and non-native speech imitation.

A necessary challenge to ST and imitation processes that faces L2 learners is the production of L2 phonemes in lexical contexts – i.e., within single- or multi-syllable utterances. In such instances, learners must parse the individual phoneme(s) from the continuous speech signal (e.g., Tyler & Cutler, 2009; McNealy et al., 2006) while coping with speaker variability and intelligibility differences (e.g., Kreitewolf et al., 2014; Okada et al., 2010; see Obleser & Eisner, 2009 for review). Moreover, learners must also transform percepts to the requisite output motor program for producing the phoneme(s) in the context of other consonants needed to form the word. The processing demands of this latter transformation stage are non-trivial, and may be followed by effects of coarticulation during speech (i.e., where preceding speech impacts upon the articulation of subsequent speech) that can extend across phonemes in complex utterances (Magen, 1997; Cho, 2004). In these cases, the complexity of ST and articulatory execution likely pose major

obstacles to non-native speakers' accuracy in producing phonemes (e.g., vowels) within word-level utterances. Considering vowels as an example, we could expect challenges to ST and articulation to manifest behaviorally as reduced accuracy in the imitation of non-native vowels within word contexts, particularly where item complexity is at its greatest and where L2 vowels pose novel articulatory demands (further to Kartushina et al., 2015; Delvaux et al., 2014).

Although much is known concerning the neural networks involved in producing connected L2 speech (Reiterer et al., 2011; 2013; Simmonds et al., 2011a), relatively few functional MRI (fMRI) studies have investigated the neural outcomes of learning on L2 speech, where those learned L2 phones must later be produced within words. Some previous fMRI investigations have shown short-term, training-related adaptations of sub-regions within networks involved in speech production. For instance, across multiple productions of novel words comprising non-native consonant clusters, activation in basal ganglia and peri-Sylvian speech regions decreases, as compared to repeated production of words composed of familiar consonants (Moser et al., 2009). Similarly, greater accuracy by native English speakers in perceiving trained Hindi retroflex contrasts has been associated with reduced activation in left frontal operculum and anterior insula (Golestani & Zatorre, 2004; see also Wong et al., 2007). Nevertheless, it remains to be addressed how L2 phonemes learned in isolation may be imitated at word level, and importantly, the role played by both ST and articulatory processes at a neural level in this regard.

Here, we used a speech learning paradigm combined with fMRI to explore vocal imitation and ST, by charting the neural outcomes of imitating learned L2 phonemes within novel non-words. Monolingual English speakers were trained to imitate a (native) front vowel and (non-native) front rounded vowel prior to scanning. Their performance on imitation of these vowels in isolation and within novel mono- or trisyllabic word contexts was later measured during ~1 hour of task performance, as we acquired BOLD fMRI. We

examined the acoustic outcomes of pre-scan training across the session, probing the distance-to-target in formant space between each of the stimuli and the corresponding imitations participants provided. We predicted that individual differences would emerge in the degree of learning success, specifically as varying change in distance-to-target over time, across subjects. We then explored the ability of subjects to imitate the trained (non-native/native) vowels within novel non-word utterances that varied in complexity (i.e., 1 or 3 syllables). Here, we predicted that performance in imitating the vowels would be less successful for the most complex items, and particularly so for the less familiar non-native vowel. We followed behavioral training with a rapid-sparse event-related fMRI protocol, where we imaged the BOLD response as participants listened to, or *listened to and then imitated*, the trained vowels (native and non-native), or their corresponding non-words. In particular, we predicted that regions of speech networks implicated within audio-motor processing (superior temporal and peri-Sylvian cortex) would show increased activation as a function of item complexity and non-nativeness of vowels, during sensorimotor transformation (ST). We further anticipated that increased item complexity effects would manifest within regions of speech motor networks (cerebellum, somatomotor cortex) during overt imitation (further to Riecker et al., 2008; Sörös et al., 2006). Finally, to explore individual differences in learning outcomes with respect to ST and imitation processes at a neural level, we conducted individual differences analyses, anticipating that acoustic performance during pre-scan imitation would predict BOLD activation for non-native (vs. native) speech during ST and during later imitation.

Materials and Methods

2.1 Participants

Participants were 28 right-handed female volunteers (mean age ± SD: 23.3 ± 4.4; range: 18-33), with no history of hearing difficulties or neurological insult. All were native British English speakers; none had studied a non-native language beyond UK GCSE level

or equivalent (see footnotes 1 & 2; supplemental table 1). In particular, fifteen participants had studied French to GCSE (front rounded vowel /y/ is native to French, see 2.2), but none had completed any further study or had immersive experience with French thereafter. Given that a female talker provided our stimuli (see 2.2), we tested female participants only, in order to avoid potential gender confounds in imitation accuracy. Participants were recruited from local subject pools; all had completed or were completing an undergraduate degree at the time of the study. All provided written informed consent in line with local ethics and MRI protocols. The study was approved by the Ethics Committee at the Department of Psychology, Royal Holloway, University of London.

2.2. Stimuli

Stimuli were sustained front vowels (mean duration [ms] ± SD: 974 ± 141), monosyllabic non-words (692 ± 69), and trisyllabic non-words (919 ± 54), produced by a phonetician of the same gender and language background as the participants. Vowels belonged to two categories: one native (/i/) and one non-native (/y/) to British English. These specific vowels were chosen since the native/non-native distinction maps onto the articulatory feature of lip rounding; i.e., rounding of front vowels is non-native to British English (Wells, 1982). Both monosyllabic and trisyllabic non-words were produced with a falling intonation contour. For the trisyllabic words, a falling nuclear tone was produced on the 2nd (i.e., middle) syllable which meant that the nucleus (i.e., vowel) of the trisyllable non-word corresponded with that of the monosyllabic word. Alveolar stops (/t/) were used as the onset of the 1st and 2nd syllables, to reduce involvement of the lips during production of the consonant that preceded the vowels, since we wished to measure lip dynamics for non-native /y/ vowels within the word contexts (data not presented).

We included five tokens for each of the six stimulus classes (stimulus classes: native vowel, non-native vowel, native monosyllable, non-native monosyllable, native

trisyllable, non-native trisyllable), giving 30 stimuli in total (see Fig. 1, lower left). For vowels, first (F1) and second (F2) formants were measured across the full, steady duration of the stimulus. For monosyllables, formants were measured from the steady state portion of the vowel, in the interval between the alveolar burst and the bilabial stop. For trisyllables, formants were measured from the steady state portion of the middle vowel, in the interval between the alveolar burst and the bilabial stop. All measurements were made using LPC analysis in Praat (Boersma & Weenink, 2016), calculated per vowel as the mean across the analysis window. The formant tracking procedure (see 2.5.1) excluded non-steady regions of the speech signal (e.g., formant transitions) using an intensity threshold (>77dB) that restricted the analysis window to the steady state portion.

For each level of complexity (vowel, monosyllable, trisyllable), raw recordings comprised ~40 exemplars (20 per vowel); we converted the F1 and F2 formant measurements to mels (O'Shaughnessy, 1987), and selected the five tokens per class as follows. First, we calculated the median of F1 and F2 values across five potential tokens (formant frequencies measured per class as described above). Second, we calculated the 2D Euclidean distance between the particular vowel category median (F1 and F2) and each token (F1 and F2). Third, we calculated the standard deviation of the 2D Euclidean distances for that category. Finally, we matched each of the stimulus classes as closely as possible for the SD of token distances to the respective category median (sometimes replacing tokens with other exemplars to achieve more similar SDs). Across all levels of complexity (vowel, monosyllable and trisyllable), native vowels differed from non-native vowels primarily along F2, with F1 showing overlap (see Fig. 1 & supplemental table 2), as expected. Stimuli were selected in this systematic fashion to ensure that variability of tokens within each class was controlled as carefully as possible across levels of complexity (see Carey et al., 2017). Stimuli were scaled to equal total RMS amplitude in Adobe Audition CS 5.5 (Adobe Systems Inc., San Jose, CA).

Stimuli were used for behavioral training and in-scanner protocols. Scanner stimuli were parametrically equalised in Adobe Audition (filter CF: 3.5 kHz; 10 dB gain; Q factor = 2), filtered with earbud-specific parameters for use with Sensimetrics earbuds, and amplified by +6dB in Adobe Audition. Additional parametric EQ and amplification were applied to improve audibility of stimuli against continuous rtMRI background noise.

2.3 Behavioral Procedure

Participants completed a language background questionnaire including proficiency estimates for any languages they had learned (see 2.1). All testing took place in a sound attenuated booth at the Department of Psychology, Royal Holloway, University of London.

The training procedure comprised a speech production training task, practice for the fMRI speech task, and a two-alternative forced choice (2AFC) perceptual discrimination task. To ensure that participants' perceptual discrimination of the vowel categories was indexed prior to and after training, the 2AFC task was completed twice; once at the beginning and once at the end of the testing session. All experiments were presented using Psychophysics toolbox (Kleiner, Brainard & Pelli, 2007) via 64-bit Matlab (Version 2015a). Audio stimuli were presented through Sennheiser HD 201 headphones.

2.3.1 Speech production training. Participants trained on producing native unrounded /i/, and its non-native rounded partner /y/. Participants watched a two-minute video, in which the same phonetician as heard in the stimuli instructed them on producing the rounded and unrounded vowels in isolation. The video included: multiple repetitions of the rounded and unrounded vowels; instructions on lip rounding and achieving it for the non-native vowel; multiple camera angles, with close-up front and profile views of the rounded and unrounded dynamics (with and without phonation).

Participants then completed a production training task requiring them to imitate the tokens from each category as accurately as possible. The task was presented over 16

blocks of 10 trials; in a given block, participants imitated all five tokens from within a single category twice. Participants sat at a fixed distance from a Røde NT1-A condenser microphone, and an LCD monitor. Each trial began with a visual prompt ('Listen') at the upper left of the screen, and presentation of one token from the category for that block. At stimulus offset, the upper left visual prompt was replaced ('Pause..') for 1.7s; this was followed by a 2s repeat window ('Repeat' instead of 'Pause..'), during which participants imitated the vowel. The next trial began after 2s had elapsed. Within-block token order was pseudorandomised separately for the initial and latter 5 trials per block, such that participants imitated all five tokens from the category in each half of the block, with the condition that identical tokens never occurred on consecutive trials. Block order for vowel category was randomised, constrained such that the same vowel category could not repeat more than once on adjacent blocks. Imitations were recorded with a condenser microphone (Røde NT1-A; Sydney, Australia), digitized in Matlab, and saved as separate .wav files per trial.

2.3.2 fMRI task practice. Following the speech imitation training task, participants completed a practice run of the fMRI in-scanner task (described in detail below - see 2.4 MRI procedure), within the sound attenuated booth. The task comprised the same vowel stimuli as during the initial training, and in addition, the monosyllabic and trisyllabic words containing the native/non-native vowels. Participants were informed that during the task, words containing the native and non-native vowels they had practiced would occur, in addition to the isolated vowels. On trials that required imitation, participants were instructed to mimic the stimulus as accurately as possible, paying particular attention to the vowel(s) in the word in the case of mono- or trisyllables. Recordings were digitised with the same procedure as for the training task above. Task timing and procedure was identical to the in-scanner protocol, with the exception that recorded scanner noise was

delivered via headphones to simulate the rapid-sparse fMRI protocol (see 2.4, below); additionally, participants completed the task while seated (supine in the scanner).

2.3.3 Perceptual discrimination. Participants made 'same or different' 2AFC judgements on exemplars from the vowel categories. Each trial presented a pair of vowels, with a 1s interval between offset of the first stimulus and onset of the second. Order of the vowels (first or second) was counterbalanced across the two intervals on the 'different' trials. 'Same' trials always consisted of two different tokens from within a single category. A visual prompt at the upper left of the screen ('Listen') appeared during audio playback, replaced with response instructions after offset of the second stimulus; participants indicated a 'same' or 'different' response with left or right arrow key press, respectively. Participants completed 4 blocks of 10 trials (20 same and 20 different), with trial order randomised. The 2AFC task was completed at the beginning and again at the end of the session. 2AFC responses were scored offline and performance measures that account for response bias (d') were calculated for each comparison (cell values of 0 or 1 were corrected for by adding 0.5 to all cells; Hautus, 1995). D prime results showed participants were readily able to perceive the distinction between the vowels (pre- and post-training mean d' > 2.5).

2.4 MRI procedure

Data were acquired on a 3T Siemens Tim Trio with a 12-element headcoil (fMRI & rtMRI) and 3–element neck array (real-time MRI of the vocal tract; rtMRI) (Siemens, Erlangen, Germany). All stimuli were delivered through MR-compatible earbuds (Sensimetrics Corp., Malden MA, USA); speech was recorded per run with a fibre-optic microphone (FOMRI-III; OptoAcoustics Ltd., Moshav Mazor, Israel). All stimuli were presented via the Psychophysics toolbox running in Matlab, with back projection for presentation of visual stimuli.

We presented participants with the native and non-native vowels, monosyllables and trisyllables during fMRI and rtMRI, under a 2 (nativeness: native/non-native) x 3 (complexity: vowel, monosyllable, trisyllable) design. This enabled us to probe whether effects of vowel nativeness might extend to the imitation of words that required some of the same novel articulatory demands as were instructed during pre-scan training, and that were practiced during the mock fMRI block prior to scanning. A trio of rtMRI runs (40s each) was presented before each of the three fMRI runs (~13 min each; total scanning time ~65 mins; Fig. 1). During rtMRI, fast gradient echo images of the vocal tract were acquired at 10 frames per second, as participants articulated each of the vowels and non-words. This enabled articulatory gestures associated with native/non-native vowel production to be imaged in isolation, and in non-word contexts. Results of rtMRI data analyses are beyond the scope of the present report and will be presented elsewhere.

fMRI acquisition entailed a rapid-sparse, event-related protocol, where auditory stimuli and speech production events were timed to occur during short silent periods between acquisition of whole-brain volumes. Each listen-then-imitate trial occurred over two acquisition + silent gap periods; participants listened to a particular vowel, and imitated it when cued after the next acquisition (Fig. 1, right). This enabled us to capture BOLD activation reflecting sensorimotor transformation and the subsequent vowel imitation. Listen only and rest trials occurred in a single acquisition + silent gap period (see Fig. 1). In the following, we distinguish listening that entailed sensorimotor transformation from passive listening, as ‘listen pre-imitate’ and ‘listen only’, respectively. Four event types were thus presented during fMRI: 1) listen pre-imitate; 2) imitation; 3) listen only; 4) rest.

fMRI trials for listen-then-imitate were cued as follows. At the onset of the first acquisition, a blue fixation cross cued that the trial would require vowel imitation. Stimuli were presented in the silent period after the first acquisition. Stimulus onsets were jittered variably from the start of the silent gap, according to a distribution of jitter onsets ranging

from 50-250 ms, with the jitter selected randomly from within that range on a per trial basis. At the offset of the next acquisition, the blue fixation cross changed to green, cueing the participant to imitate the stimulus (Fig. 1, right). Listen only trials and rest trials were cued at acquisition onset with a yellow fixation cross that remained for the trial duration; stimuli were delivered with onsets jittered variably as above. Participants were instructed to remain alert during listen trials and to not produce any speech. Five mini-blocks of 32 trials (18 listen then imitate, 12 listen only, 2 rest) were presented per fMRI run (160 trials total: 90 listen then imitate; 60 listen only; 10 rest). Trial order was randomised separately for each mini-block.

fMRI data were 3D echo-planar images (EPI) collected with rapid-sparse acquisition: voxel size 3mm isotropic; flip angle 78°; slice gap 25%; echo time (TE) 30ms; vol. acquisition time 1.7s; inter-scan silent period 1.5s. A 3D T_1-weighted MP-RAGE scan was acquired for EPI image alignment and spatial normalisation: voxel size 1 mm isotropic; flip angle 11°; TE 3.03ms; TR 1830ms; image matrix - 256 x 256 x 160.

rtMRI blocks comprised three 40s runs, completed back-to-back with a brief (~5s) pause between runs. Acquisition of rtMRI data and results of real-time MRI analyses will be presented elsewhere and are not discussed further here.

2.5 Data processing and analyses

2.5.1 Behavioral data. All audio data were screened for artefacts prior to analysis; trials containing non-speech noise (e.g., subject movement) during the vowel were not analysed. Formant extraction was performed in Praat using an automated LPC procedure that blinded the experimenter to vowel identity and allowed rapid trial-by-trial inspection of formant tracks. A minimum intensity threshold was used to isolate the steady state portion of the speech waveform as the analysis window in each recording; thresholds typically ranged from 45-60dB, modified per participant in line with the properties of their recorded

signal. Formants were measured and saved per trial as the mean of each formant across the vowel steady state duration that was identified by the analysis window (using Praat default parameters: no. of formants - 5.0; window length - 0.025s; max frequency - 5.5 kHz). Where visual inspection showed that tracks deviated from the true formant(s), we modified the number of formants and re-analysed manually. For audio data from the fMRI practice block, vowels were excised from the recordings of whole words ahead of formant analysis, to remove bursts and aspiration due to alveolar stops. The middle vowel was excised from all trisyllables. All audio editing was performed manually in Adobe Audition; cuts were made after the aspiration of the alveolar stop prior to vowel onset, but before onset of the bilabial stop (i.e., formant transitions were included in the waveform, but were excluded from the analysis window by the intensity threshold, and not analysed).

For each trial, we calculated the 2D Euclidean distance (in Mels) between the first and second formant (F1 & F2) of the stimulus and the F1 and F2 of the participant's imitation. For each participant, we calculated the mean of these 2D Euclidean distances, averaging the 2D distances across the initial four and final four blocks completed per vowel. We predicted that if learning occurred, the mean 2D Euclidean distance would reduce from the initial four to the latter four blocks (i.e., reduced distance to target would indicate improved imitation performance). For each subject, we calculated difference scores to express this: we subtracted the mean 2D Euclidean distance to target for the latter four blocks from that of the initial four blocks, per vowel. Positive difference scores therefore indicated that learning had occurred, while negative difference scores indicated poorer performance over time (see Fig. 2a). During pilot testing, we observed a clear spread in outcomes, such that approximately half of participants showed evidence of learning following imitation training (i.e., positive 2D Euclidean distance difference scores). We therefore elected *a priori* to analyse 2D Euclidean distance measures by including a binary group term for the split of the cohort into those with positive versus negative 2D

distance difference scores. Analysing the data with the group term allowed us to model performance with respect to anticipated differences in learning outcomes for the non-native vowel, where interaction terms with the grouping variable were of central interest. Analysing the data as such therefore allowed us to probe quantitatively any differences in the profiles of outcomes between learners and non-learners, where differences in the direction and nature of effects were strongly expected (see 3.1.1).

2.5.2 fMRI audio data. fMRI audio recordings for each run were processed in Praat using an automated intensity-based procedure. Detected sound boundaries were used to excise separate audio files from the full recording of all produced speech. fMRI task performance was verified by screening all excised audio and comparing to the saved stimulus logs (see 2.5.3).

2.5.3 fMRI analyses. fMRI data were pre-processed and analysed in SPM8 (Wellcome Trust Centre for Neuroimaging, UK). Functional images for each run were realigned, and the mean functional image co-registered with the anatomical scan. Location of the anterior commissure (AC) was determined manually from the anatomical scan; structural and functional images were then re-oriented so the origin of each image matched the AC. Functional images were spatially normalised using parameters derived from the segmented anatomical image, and were smoothed with an 8 mm FWHM Gaussian kernel. For each run, we set a motion criterion such that all acquisitions had maximum translations that were less than a single dimension of one voxel (i.e., for any single acquisition, the total translation over the 3 axes was <3.0 mm, relative to the mean functional image). Four participants exceeded this criterion on two or more runs, and were excluded from further analyses. In practice, we found that translations about the z-axis were most common, and of 1–2mm magnitude; rotations about the axes were generally small and did not exceed 3 degrees for any participant in any run.

At first level analysis, each condition was modelled using a separate regressor of events in a general linear model with canonical haemodynamic response function (HRF), with rest modelled implicitly. Event onsets for listen only trials and listen pre-imitate trials were modelled using the onset time of the audio stimulus. Event onsets for speech imitation were modelled using the onset of the cue to imitate (i.e., crosshair colour change at acquisition offset). Although stimuli and speech responses occurred sequentially on imitation trials, independence of regressors was assured by our design. Imitation trials could be followed by listen or rest trials; thus, subjects could not accurately predict the next trial type. Further, we jittered stimulus onsets across imitation trials; temporal onsets of participants' imitations also varied trial-by-trial, reflecting natural speech onset jitter. Error trials (e.g., speech on listen only trials, no speech on production trials) were flagged from scanner audio and onsets for any such events were included as a regressor of no interest per run in first-level models (cohort mean task accuracy >96% per block). The six motion parameters from realignment (translations & rotations about the x, y & z axes) and the run's mean EPI image were also included as separate regressors of no interest per run. First-level t-contrasts were specified for each main effect of listen only, listen pre-imitate and speech imitation (each versus rest). Contrasts of interest modelling effects of each level of nativeness and complexity (vs. rest) were specified separately for listen pre-imitate and imitation; these were later taken up to the second level and entered into 2 (nativeness) x 3 (complexity) flexible factorial random effects ANOVA models, as F tests specifying all possible directions of orthogonal effects for the three-level main effects and for the interaction terms. Additional t-contrasts were run to test for specific directional effects, and to break down significant interactions. All t-contrasts were specified as one-sample tests outside the flexible factorial ANOVA models. Results were thresholded at $p < 0.0015$, k = 50 (achieving cluster-level FDR, $q < 0.05$, unless otherwise specified). Cluster locations were determined by comparing functional activation maps overlaid onto a standard atlas

(AAL) using the xjView toolbox in SPM (http:// alivelearn.net/xjview), verified by visual inspection of activation overlaid onto standard brain volumes in MNI space.

Whole brain regression analyses were additionally conducted using SPM. We used subject-wise first-level t-contrasts (non-native > native, collapsing across complexity) separately for both listen pre-imitate and imitation, and regressed pre-scan imitation training acoustic performance (i.e., difference score for learning or lack of improvement across blocks, as a continuous variable) against voxel-wise parameter estimates. Additional analyses of listen pre-imitate ST data were conducted using separate first-level t-contrasts of non-native > native, at each level of complexity (vowel, monosyllable, trisyllable); the same pre-scan imitation training regressor was used in each of the three analyses. Participant age was entered as a covariate of no interest in all models. To assure that regression analyses were not leveraged by outlying data points, we conducted 'leave-one-out' jackknife analyses of reported regression results. We re-ran the significant models, iteratively omitting one subject from each analysis, yielding a set of 24 partial estimates of the significant fit. We then subtracted the mean of these partial estimates from the full model estimate, accounting for the subject Ns for the full model and partial estimates mean in so doing (see Abdi & Williams, 2010). The resulting jackknife estimates of the fitted models were overlaid onto a standard MNI brain template in MRICroGL (https://nitrc.org/frs/?group_id=889) for display purposes. Additionally, we extracted subject-wise mean parameter estimates from within spherical regions-of-interest (5mm radii) that were centred on the peak co-ordinates of the full regression model results, using the Marsbar toolbox in SPM. We plotted pre-scan acoustic imitation performance (i.e., 2D Euclidean distance difference scores for initial 4 training blocks minus latter 4 blocks) against these mean parameter estimates for each significant cluster (whole-brain $q < 0.05$, cluster-level FDR-corrected, unless otherwise specified).

Results

We explored the effects of imitation training on the acoustic accuracy of vowel production, for non-native and native vowels in isolation; we later tested imitation performance within syllabic contexts. We predicted that vowel behavioral training would be associated with individual differences in the acoustic accuracy of imitations; we further expected that item complexity would impact imitation success, with reduced accuracy predicted in the most complex conditions (i.e., multi-syllabic utterances containing non-native vowels). Using fMRI, we indexed changes in brain activation related to vowel nativeness and item complexity; we predicted increased activation in fronto-temporal and peri-Sylvian speech regions for non-native compared to native conditions, as well as increased activation in those regions as a function of complexity. We also expected that effects of item complexity would modulate speech articulatory networks during imitation, emerging as increased cerebellar and somatomotor activation. Finally, to explore the role of sensorimotor transformation (ST) and imitation with respect to learning success, we probed whether individual differences in pre-scan acoustic training performance could account for subsequent variation in ST or imitation activation for non-native versus native speech.

3.1 Imitation Training and generalization to multi-syllabic non-words

3.1.1 Imitation training. We trained participants to imitate a native close front vowel (/i/) and its non-native front rounded counterpart (/y/). In exploring learning performance, we defined those who had learned during training ('learners') as participants with positive difference scores for the non-native vowel when comparing the first and second half of the training session (11/24 participants met this criterion; Fig. 2a, left). Participants who did not learn ('non-learners') were defined as those with negative training difference scores for the non-native vowel (13/24 met this criterion; Fig. 2a, left). We

included learner/non-learner status as a binary between-subject factor in a mixed model ANOVA of the 2D Euclidean distance data, modelling block group (initial vs. latter four) and nativeness as within-subject factors. We predicted that learners would improve on the non-native but not the native vowel across blocks (i.e., a two-way interaction), whereas we expected no significant interaction for non-learners.

We found a significant three-way interaction of these factors, $F(1,22) = 10.1$, $p = 0.004$, $\eta_p^2 = 0.32$. Critically, splitting the interaction term by the learner/non-learner between-subject factor revealed a significant two-way block group x nativeness interaction for learners [$F(1,10) = 23.68$, $p = 0.001$, $\eta_p^2 = 0.7$]. Post-hoc tests (False Discovery Rate [FDR] corrected; Benjamini & Hochberg, 1995) showed that learners manifested no significant change over blocks in the 2D distance to target for the native vowel [$t(10) = 0.81$, $p > 0.4$], whereas by definition, they showed significant learning on the non-native vowel [$t(10) = 4.73$, $p < 0.01$] (Fig. 2b). In contrast, no significant two-way interaction emerged for non-learners [$F(1,12) = 1.34$, $p = 0.27$, $\eta_p^2 = 0.1$], and instead main effects of block group [$F(1,12) = 24.948$, $p < 0.0001$, $\eta_p^2 = 0.675$] and nativeness [$F(1,12) = 9.917$, $p = 0.008$, $\eta_p^2 = 0.452$] were significant (Fig. 2b). Thus, while learners showed selective improvement in imitation performance for the non-native vowel across training, the non-learners showed worse imitation performance for the non-native than the native vowel overall, and significantly *worsened* in their imitation of both vowels over time.

Probing the source of the 2D Euclidean distance effects in the training data, we also tested whether F1 or F2 values (in Mels) differed across block groupings per vowel, for learners and non-learners. Supplemental figure 1 presents F1-F2 plots of each vowel, for learners and non-learners across the initial four and latter four blocks. We found that for both the native and non-native vowels, F1 values increased significantly for non-learners for both vowels, from the initial four to latter four blocks [/i/: $t(12) = 4.0$, $p < 0.02$; /y/: $t(12) = 3.73$, $p < 0.03$] (all tests FDR-corrected) (Suppl. fig. 1). However, learners did not show

any robust changes in F1 or F2 alone for either vowel (all $p > 0.1$). We note however that the 4-way interaction term when testing these effects was non-robust [2 (formant) x 2 (native/non-native) x 2 (initial/latter 4 blocks) x 2 (learner/non-learner): $F(1,22) = 1.85$, $p > 0.18$].

Finally, considering those in our sample that reported second language (L2) experience (n=21), we tested whether French experience influenced status as a learner or non-learner (given that /y/ is native to French, and that participants frequently reported having learned elementary French). Almost the same number of learners (7) as non-learners (8) reported experience with French; the difference was not significant (χ^2 [df=1] = 0.1, $p > 0.7$).

3.1.2 Imitation of multi-syllabic non-words. To probe the extent to which vowel imitation would be impacted in the context of mono- and trisyllabic non-words that contained the same vowels, we assessed imitation performance in the fMRI practice task (based on 2D Euclidean distance) across miniblock, nativeness, levels of complexity (isolated vowel, monosyllable, and trisyllable) and learner/non-learner status from the training. Supplemental figure 2 presents F1-F2 plots for each vowel across levels of complexity, for learners and non-learners (based on the preceding training).

The 5 (miniblock) x 2 (nativeness) x 3 (complexity) x 2 (learner/non-learner) ANOVA showed significant main effects of nativeness [$F(1,22) = 31.92$, $p < 0.0001$, $\eta_p^2 = 0.592$] and complexity [$F(2,44) = 30.978$, $p < 0.0001$, $\eta_p^2 = 0.585$], along with a significant nativeness x complexity interaction [$F(2,44) = 15.351$, $p < 0.0001$, $\eta_p^2 = 0.411$] and a significant nativeness x learner/non-learner interaction [$F(1,22) = 7.759$, $p = 0.011$, $\eta_p^2 = 0.261$].

Exploring the nativeness x complexity interaction, post-hoc tests showed marginally greater distance to target for non-native than native vowels in isolation and in monosyllabic context ($p < 0.065$, Fig. 2c, left panel), but significantly greater distance to target for non-

native vowels than native vowels in trisyllabic context ($p < 0.01$, Fig. 2c, left panel) (all tests FDR-corrected). Moreover, post-hoc comparisons further showed that native vowels in syllabic contexts (both mono- and tri-) were imitated significantly more accurately than the native vowel in isolation (both vs. isolation $p < 0.01$, FDR-corrected; no sig. difference between mono- and trisyllable 2D distance, $p > 0.7$). In contrast, the non-native vowel was imitated most accurately within the monosyllabic context, with significantly reduced 2D distance to target compared to both the isolated non-native vowel and non-native trisyllable (both $p < 0.01$, FDR-corrected) (Fig. 2c, left; suppl. fig. 2).

The nativeness x learner/non-learner interaction revealed that for learners, the native and non-native items did not differ significantly in 2D distance to target, regardless of item complexity ($p > 0.1$) (Fig. 2c, right). This agrees with the learning found during the latter stages of training (see 3.1.1). In contrast, the non-learners maintained a significantly increased distance to target for non-native items compared to native, regardless of item complexity [$t(12) = 7.64$, $p < 0.01$] (all tests FDR-corrected) (Fig. 2c, right; suppl. fig. 2).

3.2 fMRI results - sensorimotor transformation, imitation, and 2 x 3 analyses

To probe activation reflecting speech sensorimotor transformation (ST) and imitation overall, we contrasted separately the listen pre-imitate (i.e., ST) and imitation stages of the task with rest, collapsing across all conditions. Exploring activation with respect to the conditions of nativeness and complexity, we used 2 (nativeness) x 3 (complexity) flexible factorial ANOVAs in SPM to model the ST and imitation data, specifying separate models for ST and imitation. Table 1 provides peak co-ordinates for ST and imitation activation from the 2 x 3 analyses; supplemental table 3 presents activation from t-contrasts for all ST versus rest, and all imitation versus rest.

3.2.1 Main effects. Modelling listen pre-imitate using t-contrasts, we found evidence of extensive activation for ST when collapsing across levels of nativeness and

complexity (Fig. 3a, left). Regions activated bilaterally included somatomotor cortex, insular cortex, superior and middle temporal cortex, supplementary motor area (SMA), cerebellum (including lobules V/VI), and medial temporal lobe (hippocampus, parahippocampal gyrus, entorhinal cortex). Additionally, left inferior frontal gyrus (IFG) and left middle frontal gyrus (MFG) showed significant activation. Modelling imitation (collapsing across nativeness and complexity), we found evidence of activation within sensorimotor regions implicated within speech networks, including bilateral ventral somatomotor cortex, bilateral anterior cerebellum (lobules V/VI), and left insular cortex (Fig. 3a, right).

3.2.2 2 x 3 results - imitation. The flexible factorial model of imitation data revealed significant main effects of nativeness and complexity, and suggested further evidence of significant effects for the interaction term across temporal and occipital regions. Planned one-sample t-contrasts (conducted outside of the flexible factorial ANOVA) revealed that the main effects were robust, but showed the effects related to the interaction term were non-robust (all cluster-level FDR $q > 0.05$) when testing across levels of nativeness at each level of complexity. We therefore restrict report to the main effects.

The main effect of nativeness emerged as significantly greater activation at left IFG and anterior insula for non-native than native items (t-contrast, Fig. 3b, upper left). In addition, the reverse t-contrast (native > non-native) showed significant activation at right medial pre-frontal cortex (mPFC) (Fig. 3b, lower left). In line with previous results found by our group, this may reflect differential modulation of default mode network (DMN) regions

Table 1: cluster significance and peak co-ordinates for listen pre-imitate and imitation t-contrasts

Contrast	Cluster FDR	Cluster size (voxels)	t-stat (peak)	x	y	z	Location
Imitation: Native > Non-Native							

Contrast	Cluster FDR	Cluster size (voxels)	t-stat (peak)	x	y	z	Location
	< 0.00001	720	6.16	8	50	0	RH med. sup. frontal
			4.14	-12	46	-6	LH med. orbito-frontal
			4.1	14	32	-16	RH gyrus rectus
Imitation: Non-Native > Native							
	0.003	428	4.73	-32	20	2	LH insula
			4.42	-48	2	18	LH IFG (pars opercularis)
			4.06	-42	16	2	LH front. operculum/ant. insula
Imitation: Trisyllable > Vowel							
	< 0.00001	900	8.67	22	-60	-18	RH Cerebellum (lobule VI)
	0.004	564	5.94	-60	-4	-8	LH STG
			5.55	-64	-14	8	LH STG
			5.37	-66	-14	16	LH post-central
	0.027	328	4.92	68	-30	-4	RH MTG
			4.55	68	-22	-2	RH STG/STS
			4.24	66	-36	-12	RH MTG
Listen pre-imitate: Interaction split by Complexity							
Vowel - Non-Native > Native							
	< 0.00001	2318	7.4	62	-22	0	RH STG
			5.86	58	-10	-2	RH STG
			5.75	66	-30	4	RH STG
	0.001	658	6.69	-4	14	44	LH SMA
			5.4	2	6	32	LH mid. cingulate
			4.64	0	12	26	Ant. cingulate
	< 0.00001	2159	5.93	-60	-26	2	LH MTG
			5.59	-42	-8	-6	LH insula

Contrast	Cluster FDR	Cluster size (voxels)	t-stat (peak)	x	y	z	Location
			5.56	-50	-14	-4	RH STG
	0.018	279	5.57	-48	-52	-8	LH ITG
	0.018	271	5.44	10	-18	0	RH thalamus
	0.013	321	5.16	-62	-18	32	LH post-central
			4.2	-58	-28	26	LH supramarginal
			3.73	-60	-36	38	LH supramarginal
	0.013	337	5.16	-12	-18	-2	LH Thalamus
			4.34	-8	-12	-10	LH Thalamus/VTA
			4.02	-12	-6	-4	LH Globus pallidus
	0.013	321	4.81	2	-28	32	RH middle cingulate
			4.64	4	-18	30	RH middle cingulate
Trisyllable - Non-Native > Native							
	< 0.00001	1887	6.3	-46	-12	14	LH Rolandic operculum
			5.51	-64	-22	8	LH STG
			5.5	-48	-44	10	LH MTG
	< 0.00001	963	5.21	48	-36	6	RH STG
			5.17	68	-22	0	RH STG
			5.02	44	-30	0	RH STS
	< 0.00001	1246	4.89	-14	-30	-32	LH Cerebellar peduncle
			4.78	0	-34	-38	Pons
			4.49	-4	-28	-24	LH superior Pons
	0.02	255	4.71	-54	6	28	LH pre-central
			4.08	-52	20	22	LH IFG (pars triangularis)
			4.01	-58	-2	30	LH pre-central

(S/M/I)TG – superior/middle/inferior temporal gyrus; STS - superior temporal sulcus; IFG – inferior frontal gyrus; SMA – supplementary motor area.

as a function of task demands (i.e., less suppression of mPFC in the less demanding native conditions; Carey et al., 2017; see also Geranmayeh et al., 2014).

The main effect of complexity was revealed as significantly greater activation at right anterior cerebellum (lobules V/VI), and bilateral superior temporal gyrus (STG) for

trisyllabic items compared to isolated vowels (t-contrast, Fig. 3b, right). The reverse contrast did not reveal any evidence of robust activation.

3.2.3 2 x 3 results - listen pre-imitate (ST). The listen pre-imitate flexible factorial model showed significant main effects of nativeness and complexity, in addition to significant interaction effects, which were focal to bilateral STG. Exploring these effects further, we conducted t-contrasts of non-native > native conditions (and the reverse) outside the flexible factorial ANOVA, across each level of complexity (Fig. 3d). In addition, we used the Marsbar toolbox in SPM to define spherical regions-of-interest (ROIs) of 5 mm radius at the peaks of the interaction effects at left and right STG; we then calculated and plotted mean parameter estimates across subjects within these ROIs, for each of the conditions in the 2 x 3 design (i.e., for each condition versus rest; Fig. 3c, inset).

Inspection of the parameter estimate plots revealed a monotonic increase in bilateral STG activation as utterance complexity increased, during ST (Fig. 3c). Importantly, the plots suggested greater activation for non-native than native items within the isolated vowel and the trisyllable conditions, but not in the monosyllable conditions (Fig. 3c). This was supported by the significant interaction term and whole-brain contrasts of non-native > native ST, across each level of complexity (t-contrasts, Fig. 3d; the reverse t-contrasts revealed no significant activation across any of the levels of complexity). Specifically, whole brain non-native > native contrasts revealed significantly greater activation in bilateral STG (as well as left inferior frontal and left somatomotor cortex) in the vowel and trisyllable conditions (Fig. 3d, top and bottom rows, respectively); however, no significant differences in activation were found across the superior temporal plane for monosyllables, for the non-native > native contrast (Fig. 3d, middle row).

3.3 fMRI results - individual differences analyses

A core motivation for our present study was to explore the extent to which individual differences in imitation performance for non-native vowels might account for differences in recruitment of neural resources during ST and imitation. To probe this account, we conducted whole-brain regression analyses, using participants' acoustic performance during pre-scan imitation training (i.e., either learning or reduced accuracy on /y/) as a continuous regressor for voxel-wise activation. We tested both positive and negative directions of effects, anticipating that increasing activation and/or decreasing activation might emerge within different regions of speech networks as a function of pre-scan imitation performance (further to Simmonds et al., 2014b; Moser et al., 2009). We assured robustness of regression results using a 'leave-one-out' jackknife procedure (see Materials and Methods, 2.5.3). We found evidence of significant (whole-brain cluster level FDR $q < 0.05$) positive effects of pre-scan imitation acoustic performance in predicting activation during ST (i.e., listen pre-imitate), across left insular cortex, left pre-SMA (and proximal Brodmann Area 8), and left anterior cingulate gyrus (Fig. 4, top row). Positive effects were also found at right middle cingulate sulcus and right lateral cerebellum (Crus I/II) (Fig. 4, bottom row; we note however that the cerebellar cluster only survived at $q < 0.065$ whole-brain cluster-level FDR correction). At each cluster, analyses indicated that learners (square symbols, Fig. 4) showed greater activation with improved pre-scan acoustic performance on the non-native /y/ vowel (i.e., learning during training) while non-learners (x symbols, Fig. 4) showed reduced activation as a function of less successful pre-scan acoustic performance for /y/ (i.e., reduced accuracy during training).

Extending these results, we asked whether the observed positive effects would differ across item complexity. We therefore repeated the analysis, specifying the non-native > native contrast separately for ST at each level of complexity – vowel, monosyllable and trisyllable. We again used acoustic performance for /y/ (i.e., pre-scan training 2D Euclidean distance difference score) as a continuous regressor for parameter

estimates, with age as a covariate of no interest in each model. Across the levels of complexity, our results showed differing locations of clusters where positive effects of acoustic performance on activation emerged (Suppl. fig. 3). For vowels, positive effects of acoustic performance were observed across much of left insular cortex; for monosyllables, positive effects emerged at left medial superior frontal cortex, proximal to cingulate gyrus. For trisyllables, clusters were found over more distal territories, including medial and lateral regions of right cerebellum (Crus II), right temporal pole, right medial superior frontal cortex and left SMA (Suppl. fig. 3).

We found further evidence of negative relationships between pre-scan acoustic imitation performance and listen pre-imitate activation within the head of the caudate nucleus bilaterally; however, jackknife analysis showed these relationships to be non-robust (data not presented).

We found no evidence of robust positive or negative relationships when using pre-scan acoustic performance as a predictor of non-native > native activation for imitation.

Discussion

Here, we used a speech imitation training paradigm combined with fMRI to explore L2 vocal imitation skill, charting the behavioral and neural outcomes of imitating trained vowels at word level. We tested whether imitation training with isolated L2 vowels would be associated with more accurate production of those vowels as the session progressed. Later, we tested whether imitation accuracy would be impacted when L2 vowels occurred

in non-words that varied in item complexity (one vs. three syllables), and whether individual differences in learning outcomes would persist at non-word level. Moreover, we aimed to relate individual differences in training success when imitating novel isolated vowels, to neural activation during sensorimotor transformation (ST) (i.e., pre-imitation listening) and during imitation itself.

We found substantial individual differences in imitation success for training on the non-native vowel tokens in isolation: just under half of the cohort showed at least some evidence of learning during imitation training, whereas the remaining participants tended to worsen as training proceeded. Importantly, these individual differences were preserved in a subsequent task, across imitation of the isolated vowels and mono- and trisyllabic words that contained those vowels. Moreover, we found that item complexity influenced imitation success despite individual differences in performance, such that monosyllabic contexts afforded the most acoustically accurate imitations of non-native vowels. fMRI data revealed a widespread network involved in ST, which was modulated by both vowel nativeness and utterance complexity: activation in STG increased monotonically with complexity, and STG activation was greater for non-native than native vowels in isolation and in trisyllabic contexts, but not in monosyllabic contexts. Activation related to imitation was also modulated by complexity, with greater activation found in right anterior cerebellum and bilateral STG for trisyllables than for vowels. Finally, we found that individual differences in the change in acoustic accuracy of non-native vowel imitations during training were related to activation during ST. Specifically, pre-scan acoustic performance (i.e., difference scores during vowel training) predicted activation for non-native versus native items during ST, in insular cortex, pre-SMA, cerebellum and cingulate cortex; thus, activation in these regions varied as a function of participants' performance during pre-scan training, with greater or lesser activation reflecting better versus poorer performance, respectively.

4.1 Behavioral findings

Our findings of individual differences in imitation success for the non-native front rounded vowel (/y/) add to previous behavioral and training studies that have explored non-native vowel imitation. Levy and Law (2010) examined the production of Parisian French vowels (including /y/) by American English speakers. Using ratings of their vowel production accuracy by Parisian French native speakers, Levy and Law (2010) showed that rated production accuracy tended to improve as a function of French language experience. Moreover, the syllabic context in which the vowel /y/ occurred significantly impacted production accuracy: alveolar context (/radVta/) was associated with more accurate production than bilabial context (/rabVpa/), and particularly so for those with moderate French tuition experience.

Our present results were derived from indices of speech signal acoustic distances (see Delvaux et al., 2014). We showed that imitation performance varied considerably across individuals and was impacted by syllabic context. Given Levy and Law's (2010) results, an important consideration is the effect of participants' L2 experience (e.g., French) on performance (see Kartushina et al., 2015, for discussion). We found that very similar numbers of participants across our learners and non-learners had experience with French (with groups not differing significantly in those numbers), and that all of those participants reported elementary French ability (e.g., reading simple words, signs, etc.) More critical was our finding that contextual effects had a significant impact on our participants' imitation accuracy, as when vowels were presented in novel words. Furthering Levy and Law (2010), we found that simple [tVb] contexts were associated with significantly more accurate imitation than trisyllabic [tVtVbV] contexts, or vowels in isolation. Indeed, this was true both for learners and non-learners.

In considering why simple contexts led to more acoustically accurate imitation of non-native items, it is important to note that both of the non-word contexts (mono and trisyllabic) afforded similarly accurate imitations of native vowels; however, only the simpler monosyllabic context facilitated improved imitations of the non-native vowel (relative to imitation of that vowel in isolation or in trisyllables). Based on our predictions, this appears to suggest that the increased articulatory complexity of the trisyllabic context, when combined with the non-native vowels, impacted participants' ability to accurately imitate the mid-stressed /y/ vowel (further to Magen, 1997; see Kühnert & Nolan, 1999).

We can then ask why isolated vowels were associated with less accurate imitations compared to in simple context, particularly in the case of non-native /y/ (cf. context effects in perceiving French vowels; Gottfried, 1984). The initial training data in our experiment offer one possible account: inspecting the formant data across vowel training, F1 was increased for the latter four compared to the initial four blocks in non-learners' imitations of both vowels (although this was not observed for learners). Given that there is an inverse relationship between F1 and tongue height, the increase in F1 for isolated vowels may suggest that non-learners reduced their tongue height over time. Indeed, this may have persisted in the subsequent vowel and non-word imitation task (see increased F1 for vowels and trisyllables, versus monosyllables – Suppl. fig. 2). The precise reasons as to why tongue height might have changed are unclear. However, given that we did not find corresponding changes in F2, it seems unlikely that non-learners systematically assimilated the perceived vowel to an incorrect category (e.g., /u/; Flege et al., 1997; Strange et al., 2009). It is possible however that non-learners may have reverted to a F1 closer to that typical of close front vowels in their own speech (see Kartushina & Frauenfelder, 2014), which might explain the wholesale F1 increase seen for both the native and non-native isolated vowels. Extending this account to address the reduced accuracy for isolated vowels compared with vowels in monosyllables, it may be that some

subjects had increased difficulty in maintaining a fixed tongue height for sustained isolated vowels (as compared to the monosyllables). We may be able to probe this articulatory account of poorer imitation performance in future analyses, using the real-time vocal tract MR images we collected from our participants prior to each fMRI run (further to Pillot-Loiseau et al., 2013).

Finally, the present results call for further consideration of the mechanisms by which training effects emerged. We showed participants an initial articulatory instruction video, however our paradigm did not provide any explicit feedback during training. Studies that have used visual cues to provide trial-by-trial feedback as to the proximity of imitation formants to the formants of the target stimulus (Dowd et al., 1997; Carey, 2004; Kartushina et al., 2015) have shown significantly reduced target distance following training. Nevertheless, recent data also suggest that articulatory training without visually-based formant feedback can yield significant improvements in accuracy ratings of non-native vowels at post-training (Wong, 2013). Here, we suggest that relatively unsupervised imitation training can yield successful reductions in acoustic distance to target for at least some participants. Moreover, an important consideration appears to be the nature of the stimuli used during training; we suggest that monosyllabic contexts could be particularly beneficial for achieving more optimal non-native vowel imitation, over and above vowels in isolation (further to Kartushina et al., 2015; Dowd et al., 1997). A further potential approach in future paradigms could be the provision of online visual articulator feedback, by using real-time MRI of the vocal tract to instruct subjects in the correct positioning of the tongue during imitation (for instance, instructing correct tongue height in the case of the present close vowels; see Pillot-Loiseau et al., 2013).

4.2 fMRI results

Using an event-related, rapid-sparse fMRI paradigm, we were able to model both the sensorimotor transformation (ST) and imitation of vowels, in isolation as well as in mono- and trisyllabic contexts. We found an extensive network involved in ST, which was modulated by vowel nativeness and utterance complexity. Specifically, we showed that activation in STG increased monotonically with complexity; moreover, greater activation emerged there for non-native versus native vowels in the isolated and trisyllabic contexts, whereas we found no significant difference related to nativeness in the monosyllabic context condition.

Activation related to preparatory speech mechanisms has implicated a range of areas, including dorsolateral frontal cortical regions, anterior insula, SMA and superior cerebellum (e.g., Riecker et al., 2005; Carey et al., 2017). While these regions and others (bilateral somatomotor cortex and IFG) were involved when we contrasted ST in all conditions with rest, we found that activation as a function of nativeness and utterance complexity was largely focused within bilateral STG. Activation in superior temporal gyrus and sulcus has long been associated with perception of intelligible speech (Scott et al., 2000; Davis & Johnsrude, 2003; Wilson & Iacoboni, 2006; but see Leech et al., 2009). With respect to the nativeness of overt speech, increased activation has been found in superior temporal cortex (in addition to insular, pre-motor and inferior frontal cortex) for monolingual English speakers when reading aloud sentences in French (versus rest) (Berken et al., 2015). Further, a fMRI study that used Dynamic Causal Modelling to explore modulation of overt speech activation in native versus non-native English speakers, found that non-native speakers showed lesser suppression of superior temporal cortical regions by pre-central gyrus, together with increased auditory feedback from superior temporal cortex to pre-central regions (Parker-Jones et al., 2013). In tandem, studies of overt speech have found greater recruitment of left superior temporal gyrus and sulcus as a function of increasing semantic and phonological complexity, in monolingual

adults (Krishnan et al., 2013). Further, a meta-analysis of overt speech studies showed that more novel speech items (i.e., pseudowords) were associated with increased likelihood of activation in left superior temporal gyrus, when compared to less novel real words (Davis & Gaskell, 2009). These findings suggest key roles for superior temporal cortex in processing both the content of the speech signal with respect to familiarity or nativeness, in addition to processing the acoustic and phonological complexity of speech utterances, prior to or during articulation (for review, see Myers, 2014; Simmonds et al., 2011b).

Here, we expand on these findings, demonstrating that during ST, superior temporal gyrus activation is modulated by both the complexity and nativeness of the to-be-imitated stimuli. Indeed, an important consideration was that nativeness effects (i.e., non-native > native) on STG activation emerged differentially across complexity conditions, appearing only in the isolated vowel and trisyllable conditions. By contrast, the monosyllabic condition, which was associated with largely more accurate pre-scan acoustic imitations, showed no evidence of significant activation differences within superior temporal regions, when contrasting non-native and native items. We suggest that the activation differences we observed here may have varied as a function of the apparent imitation demands, phonological novelty, and phonological complexity of the to-be-imitated stimuli. In particular, that subjects imitated the non-native monosyllables with relatively reduced distance to target than the non-native isolated vowels or trisyllables, appears to agree with the fact that non-native monosyllables taxed STG processing resources to a similar degree as the native monosyllables. Conversely, the less successful imitations we found for non-native isolated vowels and trisyllables (compared to their native counterparts) fits with the activation differences at STG that we found for the corresponding isolated vowel and trisyllable non-native > native contrasts. Taken together, we suggest that STG activation during ST appears to be mediated by the inherent

demands in sensorimotor transformation, which may hinge upon the phonological novelty of the to-be-imitated item, together with the articulatory complexity of the attendant speech motor sequence.

With respect to imitation, we observed that complexity effects modulated activation in right superior cerebellum (lobules V/VI) and bilateral STG – activation was increased for imitation of the trisyllable compared to the isolated vowel. In light of previous evidence of cerebellar recruitment as a function of articulatory complexity and the frequency of occurrence of consonant clusters (Riecker et al., 2008; Segawa et al., 2015; see also Sörös et al., 2006), our results further support these effects with respect to novel multi-syllabic utterances.

Examining effects of nativeness, we found that non-native imitation (vs. native) was associated with significant activation in left inferior frontal gyrus (IFG). Modulation of left IFG by non-native speech has been well documented (Simmonds et al., 2011a, 2011b), and may reflect the taxing of phonological and articulatory resources by the less familiar vowel. We further observed that native items (across all levels of complexity) activated medial pre-frontal cortex (mPFC) to a greater degree than non-native items. In line with previous findings from our group (Carey et al., 2017) and evidence of default mode network modulation during speech tasks (Geranmayeh et al., 2014), this result may reflect the lesser suppression of mPFC during native compared to non-native speech, owing to the reduced task demands posed by the more familiar native items.

4.3 Individual differences results

A key motivation for our current study was to explore individual differences in speech imitation training outcomes with respect to neural substrates supporting ST and subsequent imitation. Using measures of the change in acoustic performance (i.e., learning or worsening) over the course of pre-scan training, we found that those acoustic

measures were associated with significant modulation of activation for non-native versus native items during ST. We observed significant linear effects, such that with improved acoustic performance (i.e., learning) during training, activation increased within insular cortex, pre-SMA, cingulate gyrus and sulcus, and cerebellum; conversely, worsening acoustic performance was related to reduced activation in these regions. We did not find evidence of significant positive or negative effects when regressing pre-scan acoustic performance onto non-native versus native activation for imitation.

Previous fMRI studies that have investigated short term learning of novel speech items have found evidence of positive linear effects of improved performance on activation during speech production. Exploring the articulation of novel consonant clusters, Segawa et al. (2015) found that successful learning of utterance duration was associated with significantly increased activation in frontal operculum. Further, Moser et al., (2009) found that, for the production of novel pseudowords composed of consonant clusters non-native to English, learning was positively correlated with increases in activation at left anterior insula (see also Schuster, 2009). With respect to ST, recent fMRI evidence has suggested that success in vocal pitch imitation reflects differing extents of cortical recruitment: activation was found to correlate negatively with the degree of successful pitch imitation achieved (in primary auditory regions and supramarginal gyri; Garnier et al., 2013).

That we found greater activation with better pre-scan learning suggests the engagement of a broader range of cortical and cerebellar resources in those who achieved more successful outcomes prior to scanning. A number of these regions have been implicated within speech networks (e.g., insula and cerebellum); in particular, anterior versus middle regions of the insula have been identified as having respective roles in expressive and receptive language (Oh et al., 2014). Agreeing with our findings of its involvement in ST and of activity up-regulation in better learners, the insula may play a key role in articulatory planning (see Brown et al., 2009; see also Price, 2010). A notable point

however was that the modulation of insula activity in better learners emerged most clearly in the most basic context – that of the isolated vowel. This may suggest that insular recruitment during ST occurs differentially where novel phonological items are processed, yet where sequencing of articulatory gestures over time is less critical (but see Riecker et al., 2008).

We observed a variety of regions of association cortex including cingulate gyrus and sulcus (in left and right hemispheres, respectively), in addition to pre-SMA and SMA that were also engaged to a greater extent in better performers, and particularly so in the more complex trisyllabic conditions, where cerebellar activation was also modulated by acoustic learning performance. The recruitment of cingulate and pre-SMA regions may reflect aspects of articulatory planning that appear to have differed between learners and non-learners. Links between anterior cingulate and SMA have been posited to form part of the broader speech production network, particularly given the dense projections of these regions to motor cortex (see Sörös et al., 2006). Recent evidence with respect to SMA has suggested its key role in the formation of auditory images (Lima et al., 2015), whereas pre-SMA appears to show involvement in sound-action related tasks, particularly speech articulation (Adank et al., 2013; see Lima et al., 2016, for review). A possible mechanism at play here – especially in the more complex trisyllable condition – is the up-regulation of regions involved in envisaging the auditory L2 target, and in planning the sequence of articulatory gestures needed to imitate it. These mechanisms may have been differentially engaged by better versus poorer learners, perhaps suggestive of a greater degree of success in imagining the perceived target and planning the subsequent imitation. In addition, the engagement of cerebellar processes agrees with the increased activation typically found as speech rate and/or syllable complexity increase (Riecker et al., 2006, 2008). Considering the potential sources of individual differences between better and poorer learners, we could speculate these may arise from a combination of the articulatory

and motor performance demands posed by the non-native vowels, underpinned by possible variation in functional connectivity (e.g., Parker-Jones et al., 2013) and/or anatomy (e.g., Golestani et al., 2007; Golestani et al., 2011) amongst brain regions involved in transforming from the perceived speech signal to phonemic and subsequent motor targets.

In sum, we propose that the regions which showed increased activation in better learners/decreased activation in poorer learners during ST, reflect a network largely involved in transforming speech sensory and/or phonemic information, to speech articulatory plans. More widespread elements of this network were engaged differentially when those subjects prepared to imitate the non-native items that involved greatest articulatory demands – i.e., as item complexity increased.

Conclusions

Here, we showed that speech L2 learning is associated with individual differences in performance that can account for activation in speech preparatory networks during sensorimotor transformation. Further, we revealed a role for bilateral STG in processing the complexity of novel non-words, together with the familiarity (i.e., nativeness) of the target vowels within those non-words. Our results hold implications for the importance of sensorimotor transformation as a process underlying success in the imitation of non-native speech. Further, our findings inform accounts of the broader network of regions that pertain to articulatory mechanisms necessary for imitating novel speech.

Footnotes

1. The General Certificate of Secondary Education (GCSE) is the secondary education curriculum completed by students in the UK by the end of their fifth year (at ~16 years of age). As foreign language study typically does not form part of the primary education curriculum in the UK, most students begin learning a second language on entry to second level (age 11-12). The curriculum requires compulsory study of English, Mathematics, Science, and typically one foreign language. Languages are taught according to a broad syllabus that includes grammar, written and aural comprehension, written expression, with some practice of basic spoken language (e.g., 'take part in a short conversation, asking and answering questions, and exchanging opinions') (for details of the GCSE curriculum, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/485567/GCSE_subject_content_modern_foreign_langs.pdf). In 2016, 46.2% of overall GCSE French results were B (70-79%) or C (60-69%) grades (see http://www.bstubbs.co.uk/gcse.htm). All subjects had completed the UK A-level curriculum following GCSE, but importantly had not studied a language at A-level. Hence, any school experience our subjects had with languages up to GCSE had ceased at least two years before the study; no subjects reported any continued experience with second language study in the intervening time.

2. Of the 24 participants we included in the analyses, 15 had previous experience with playing a musical instrument or with voice; the remaining 9 had none. We analysed musical experience in a 2 x 2 contingency table, comparing it with learner status (i.e., those with/without musical performance across learner/non-learner outcome from the training stage of the experiment). We found no evidence of any significant differences in cell counts (χ^2 [df=1] = 0.01, $p > 0.9$; those without musical experience: 4 learners, 5 non-learners; those with musical experience: 7 learners, 8 non-learners).

Acknowledgements

Funded by Economic and Social Research Council (grant ES/L01257X/1). We thank Mr. Ari Lingeswaran for assistance with data collection, and Ms. Andreia Freitas for technical assistance with rtMRI data.

References

Adank PM, Rueschemeyer SA, Bekkering H. 2013. The role of accent imitation in sensorimotor integration during processing of intelligible speech. Front Neurosci. doi:10.3389/fnhum.2013.00634.

Berken JA, et al. 2015. Neural activation in speech production and reading aloud in native and non-native languages. NeuroImage. 112: 208-217.

Boersma P, Weenink D. 2016. Praat: doing phonetics by computer. Version 6.0.13.

Bohland JW, Bullock D, Guenther FH. 2010. Neural representations and mechanisms for the performance of simple speech sequences. J Cogn Neurosci. 22(7):1504–1529

Brown S, Laird A, Pfordresher P, Thelen S. 2009. The somatotopy of speech: phonation and articulation in the human motor cortex. Brain Cogn 70:31-41.

Callan DE, Jones JA, Callan AM, 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. 22:1182-1194.

Carey D, Miquel, ME, Evans BG, Adank P, McGettigan C. 2017. Vocal tract images reveal neural representations of sensorimotor transformation during speech imitation. Cereb Cortex. doi: 10.1093/cercor/bhx056.

Carey M. 2004. CALL visual feedback for pronunciation of vowels: Kay Sona-Match. CALICO J. 21:571–601.

Cheung C, Hamilton LS, Johnson K, Chang EF. 2016. The auditory representation of speech sounds in human motor cortex. eLife. 5:e12577. doi: 10.7554/eLife.12577

Chang EF, Niziolek CA, Knight RT, Nagarajan SS, Houde JF. 2013. Human cortical sensorimotor network underlying feedback control of vocal pitch. Proc Natl Acad Sci USA. 110(7):2653-2658.

Cho T. 2004. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. J Phonetics. 32:141-176.

Cogan GB, Thesen T, Carlson C, Doyle W, Devinsky O, Pesaran B. 2014. Sensory-motor transformations for speech occur bilaterally. Nature. 507:94-98.

Davis MH, Johnsrude IS. 2003. Hierarchical processing in spoken language comprehension. J Neurosci. 23(8):3423–3431.

Davis MH, Gaskell MG. 2009. A complementary systems account of word learning: neural and behavioural evidence. Phil Trans Royal Soc B: Biol Sci. 364(1536):3773-3800.

Delvaux V, Huet K, Piccaluga M, Harmegnies B. 2014. Phonetic compliance: a proof-of-concept study. Front Psychol. doi: 10.3389/fpsyg.2014.01375.

Dowd A, Smith J, Wolfe J. 1998. Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time. Lang Speech. 41:1–20.

Flege JE, Bohn O-S, Jang S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. J Phonetics. 25:437-470.

Flege JE, Hillenbrand J. 1984. Limits on phonetic accuracy in foreign language speech production. J Acoust Soc Am. 76: 708–721.

Garnier M, Lamalle L, Sato M. 2013. Neural correlates of phonetic convergence and speech imitation. Front Psychol. doi: 10.3389/fpsyg.2013.00600.

Geranmayeh F, Wise RJS, Mehta A, Leech R. 2014. Overlapping networks engaged during spoken language production and its cognitive control. J Neurosci. 34(26):8728–8740.

Golestani N, Zatorre RJ. 2004. Learning new sounds of speech: reallocation of neural substrates. NeuroImage. 21:494-506.

Golestani N, Molko N, Dehaene S, LeBihan D, Pallier C. 2007. Brain structure predicts the learning of foreign speech sounds. Cereb Cortex. 17:575-582.

Golestani N, Price CJ, Scott SK. 2011. Born with an ear for dialects? Structural plasticity in the expert phonetician brain. J Neurosci. 31(11):4213-4220.

Gottfried TL. 1984. Effects of consonant context on the perception of French vowels. J Phonetics. 12:91-114.

Guenther FH. 2006. Cortical interactions underlying the production of speech sounds. J Comm Disord. 39(5):350–365.

Hautus MJ. 1995. Corrections for extreme proportions and their biasing effects on estimated values of d'. Behav Res Meth Ins C. 27(1):46–51.

Hickok G. 2012. Computational neuroanatomy of speech production. Nat Rev Neuro. 13:135–145.

Kartushina N, Hervais-Adleman A, Frauenfelder UH, Golestani N. 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. J Acoust Soc Am. 138(2):817–832.

Kartushina N, Frauenfelder UH. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. Front Psychol. doi: 10.3389/fpsyg.2014.01246.

Klein D, Watkins K, Zatorre RJ, Milner B. 2006. Word and nonword repetition in bilingual subjects: a PET study. Hum Brain Mapp. 27:153-161.

Kleiner M, Brainard D, Pelli D. 2007. "What's new in Psychtoolbox-3?" Percep 36: ECVP Abstract Supplement.

Kreitewolf J, Gaudrain E, von Kriegstein K. 2014. A neural mechanism for recognizing speech spoken by different speakers. NeuroImage. 91:375-385.

Krishnan S, Leech R, Mercure E, Lloyd-Fox S, Dick F. 2015. Convergent and divergent fMRI responses in children and adults to increasing language production demands. Cereb Cortex. 25(10):3261-3277.

Kühnert B, Nolan F. 1999. The origin of coarticulation. Hardcastle WJ & Hewlett N (Eds.), Coarticulation: Theory, data and techniques (pp. 1–30). Cambridge: Cambridge University Press.

Leech R, Holt LL, Devlin JT, Dick F. 2009. Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. J Neurosci. 29(16):5234-5239.

Leonard MK, Cai R, Babiak MC, et al. 2016. The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. Brain and Lang. doi: 10.1016/j.bandl.2016.06.001.

Levy ES, Law FF. 2010. Production of French vowels by American-English learners of French: language experience, consonantal context, and the perception-production relationship. J Acoust Soc Am. 128(3). 1290-1305.

Lima CF, Lavan N, et al. (2015). Feel the noise: relating individual differences in auditory imagery to the structure and function of sensorimotor systems. Cereb Cortex. 25(11):4638-4650.

Lima CF, Krishnan S, Scott SK. 2016. Roles of supplementary motor areas in auditory processing and auditory imagery. Trends Neurosci. 39(8):527-542.

Magen H. 1997. The extent of vowel-to-vowel coarticulation in English. J Phonetics. 25:187-205.

McNealy K, Mazziotta JC, Dapretto M. 2006. Cracking the language code: neural mechanisms underlying speech parsing. J Neurosci. 26(29):7629-7639.

Moser D, et al. 2009. Neural recruitment for the production of native and novel speech sounds. NeuroImage. 46:549–557.

Myers EB. 2014. Emergence of category-level sensitivities in non-native speech sound learning. Front Neurosci. doi:10.3389/fnins.2014.00238.

Obleser J, Eisner F. 2009. Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci. 13(1):14-19.

Okada K, et al. 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex. 20(10):2486-2495.

Oh A, Duerden EG, Pang EW. 2014. The role of the insula in speech and language processing. Brain Lang. 135:96–103.

Parker-Jones O, et al. 2013. Auditory-motor interactions for the production of native and non-native speech. J Neurosci. 33(6): 2376-2387.

Parker-Jones O, et al. 2014. Sensory-to-motor integration during auditory repetition: a combined fMRI and lesion study. Frontiers Hum Neurosci 8: doi:10.3389/fnhum.2014.00024.

Perani D, et al. 2003. The role of age of acquisition and language use in early, high-proficient bilinguals: an fMRI study during verbal fluency. Hum Brain Mapp. 19:170–182.

Pillot-Loiseau C, Kocjancic,T, Kamiyama T. 2013. Contribution of ultrasound visualisation to improving the production of the French /y/-/u/ contrast by four Japanese learners. Proc PPLC13: Phonetics, Phonology, Languages in Contact. Contact Varieties, Multilingualism, Second Language Learning. 86–89.

Price CJ. 2009. The anatomy of language: a review of 100 fMRI studies published in 2009. Ann NY Acad Sci. 1191(1):62-88.

Rauschecker AM, Pringle A, Watkins KE. 2008. Changes in neural activity associated with learning to articulate novel auditory pseudowords by covert repetition. Hum Brain Mapp. 29:1231–1242.

Reiterer SM, et al. 2011. Individual differences in audio-vocal speech imitation aptitude in late bilinguals: functional neuroimaging and brain morphology. Front Psychol. doi: 10.3389/fpsyg.2011.00271.

Reiterer SM, Hu X, Sumathi TA, Singh NC. 2013. Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. Front Psychol. doi:10.3389/fpsyg.2013.00782.

Riecker A, Brendel B, Ziegler W, Erb M, Ackermann H. 2008. The influence of syllable onset complexity and syllable frequency on speech motor control. Brain Lang. 107:102–113.

Riecker A, et al., 2005. fMRI reveals two distinct cerebral networks subserving speech motor control. Neurology. 64: 700-706.

Shuster LI. 2009. The effect of sublexical and lexical frequency on speech production: an fMRI investigation. Brain Lang. 111(1):66-72.

Scott SK, Blank CC, Rosen S, Wise RJ. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 123:2400-2406.

Segawa JA, Tourville JA, Beal DS, Guenther FH. 2015. The neural correlates of speech motor sequence learning. J Cogn Neurosci. 27(4):819–831.

O'Shaughnessy D. 1987. Speech communication: human and machine. New York: IEEE Press.

Sörös P, Sokoloff LG, Bose A, McIntosh AR, Graham SJ, Stuss DT. 2006. Clustered functional MRI of overt speech production. NeuroImage. 32:376-387.

Strange W, Levy ES, Law FF. 2009. Cross-language categorization of French and German vowels by naïve American listeners. J Acoust Soc Am. 126(3): 1461-1476.

Tyler MD, Cutler A. 2009. Cross language differences in cue use for speech segmentation. J Acoust Soc Am. 126(1): 367-376.

Simmonds AJ. 2015. A hypothesis on improving foreign accents by optimizing variability in vocal learning brain circuits. Frontiers Hum Neurosci. 9: ISSN:1662–5161

Simmonds AJ, Leech R, Collins C, Redjep O, Wise RJS. 2014a. Sensory-motor integration during speech production localizes to both left and right Plana Temporale. J Neurosci. 34(39):12963–12972.

Simmonds AJ, Leech R, Iverson P, Wise RJS. 2014b. The response of the anterior striatum during adult human vocal learning. J Neurophysiol. 112:792–801.

Simmonds AJ, Wise RJS, Dhanjal NS, Leech R. 2011a. A comparison of sensory-motor activity during speech in first and second languages. J Neurophysiol. 106:470–478.

Simmonds AJ, Wise RJS, Dhanjal NS, Leech R. 2011b. Two tongues, one brain: imaging bilingual speech production. Front Psychol. doi: 10.3389/fpsyg.2011.00166.

Wells JC. 1982. Accents of English (vol. 1). Cambridge, UK: Cambridge University Press.

Wilson SM, Iacoboni M. 2006. Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. NeuroImage. 33:316–325.

Wong JWS. 2013. The effects of perceptual and or productive training on the perception and production of English vowels /I/ and /i:/ by Cantonese ESL learners. Proc. Interspeech. 14:2113–2117.

Wong PCM, Perrachione TK, Parrish TB. 2007. Neural characteristics of successful and less successful word learning in adults. Hum Brain Mapp. 28:995-1006.

Supplemental table 1: Participants' L2 experience and proficiency estimates

Participant	Language learned	Duration learned (years)	Proficiency	Learned during pre-scan training (yes/no)
1	French	5	2	Yes
2	German	7	1	No
3	German	5	2	Yes
4	French	5	1	No
5	French	5	1	Yes
6	French / German	4 / 6	1 / 2	No
7	Gujarah	N/A	2	Yes
8	French	5	1	No
9	German / French	2 / 5	1 / 2	Yes
10	Welsh	15	3	No
11	N/A			Yes
12	French / German	N/A / N/A	1 / 1	Yes
13	N/A			No
14	French / German	0.5 / N/A	1 / 1	Yes
15	German / Russian	5 / 4	1 / 1	Yes
16	Spanish / French	2 / 2	2 / 1	Yes
17	French	4	1	No
18	German	4	1	Yes
19	French	5	1	No
20	French	1	1	No
21	Kurdish	N/A	3	No
22	French / Spanish	5 / 5	1 / 2	No
23	French	4	2	No
24	N/A			No
25 *	Urdu	N/A	3	N/A
26 *	Spanish	2	1	N/A
27 *	Gujarah	N/A	2	N/A
28 *	French	5	1	N/A

* Participants excluded from all analyses due to exceeding fMRI head motion criteria.
N/A - not applicable; N/A for duration learned indicates no formal experience with language study (e.g., infrequently spoken at family gatherings, on holidays, etc.) Proficiency estimates key: 1) 'I can understand simple signs and words' 2) 'I can understand simple conversations' 3) 'I can read magazines and/or have conversations with friends'

Supplemental table 2: mean (±SD) F1 and F2 for stimulus categories (Hz & Mels)

Category	Item complexity	F1 (Hz)	F2 (Hz)	F1 (Mels)	F2 (Mels)
/i/	Vowel	308.60 (10.53)	2694.09 (32.68)	411.58 (11.79)	1779.15 (10.82)
	Monosyllable	290.90 (10.97)	2692.07 (20.99)	391.61 (12.42)	1778.51 (6.97)
	Trisyllable	320.30 (4.11)	2584.74 (33.33)	424.62 (4.54)	1742.24 (11.40)
/y/	Vowel	275.89 (7.02)	2198.98 (23.10)	374.44 (8.11)	1601.47 (8.96)
	Monosyllable	284.77 (8.75)	2163.84 (21.26)	384.64 (10.01)	1587.73 (8.37)
	Trisyllable	287.81 (6.95)	2151.92 (30.80)	388.12 (7.94)	1583.00 (12.24)

Supplemental table 3: cluster significance and peak co-ordinates for all ST and all imitation (vs. rest)

Contrast	Cluster FDR	Cluster size (voxels)	t-stat (peak)	x	y	z	Location
All Listen pre-imitate > rest			14.15	48	-10	36	RH post-central
			14	-50	-12	36	LH post-central
	< 0.00001	44645	12.01	10	-54	-20	RH cerebellum (lobule V)
			8.82	8	2	48	RH SMA
			8.33	-4	4	48	LH SMA
	0.002	745	7.2	-6	10	38	LH mid. cingulate
All Imitation > rest							
	< 0.00001	2215	10.48	-60	-8	8	LH Rolandic operculum
			8.68	-54	-14	10	LH Rolandic operculum/STG
			4.53	-36	4	6	LH Insula
	< 0.00001	1426	8.33	58	-4	12	RH Rolandic operculum
			7.81	64	-4	4	RH Rolandic operculum/STG
			6.51	66	-6	22	RH ventro-lateral post-central
	0.009	361	8.06	-14	-62	-20	LH cerebellum (lobules V/VI)
	0.014	292	7.07	16	-60	-22	RH cerebellum (lobules V/VI)

Figure captions

Fig. 1. Experimental training protocol, in-scanner procedure, and analyses. Top left: participants trained on imitating a native unrounded front vowel (/i/) and its non-native front rounded partner (/y/). 16 blocks of 10 trials were presented, with tokens of a single vowel category presented per block (each of 5 tokens presented twice, non-consecutively). Top right: in-scanner procedure. Participants completed three block 'trios' of real-time vocal tract MRI (~3 mins each; not analysed here), with each followed by a block of fMRI (~13 mins each). During both types of scans, participants imitated the isolated vowels learned during training, along with the one- or three-syllable non-words comprised of those vowels (see Stimuli - bottom left). Non-words were practised prior to scanning in a task that mirrored the fMRI procedure (bottom row, middle) (see Materials and Methods). Bottom right: separate flexible factorial ANOVA models were specified in SPM for the pre-imitation listening and imitation portions of the task. Passive listening trials were included interleaved with 'listen then imitate' trials; these passive trials are not analysed here.

Fig. 2. Imitation training behavioral results. (a) Individual differences in training outcomes. Training difference scores for 2D Euclidean distance to target for individual participants (in Mels), for non-native (left) and native (right) trained vowels. Difference score per participant calculated as: mean 2D Euclidean distance to target for blocks 1-4 minus mean 2D Euclidean distance to target for blocks 5-8 (i.e., positive difference scores indicate learning). Participants grouped according to 'learner' and 'non-learner' status for /y/ vowel (learner defined as difference score for /y/ > 0). Group means, standard errors of mean (SEM) and standard deviations (SD) shown as red, blue and cyan lines, respectively. Symbols denote individual participants (square: learner; x: non-learner). (b) Group mean 2D Euclidean distance to target (±SEM) for initial and latter four blocks, replotted by learner (left) and non-learner (right) status, across vowels, and block grouping. (c) (left) Mean 2D Euclidean distance to target for non-word imitation task (i.e., fMRI practice), plotted across levels of complexity (isolated vowels, monosyllables, trisyllables) and nativeness (bracketed lines denote pairwise comparisons). (Right) Mean 2D Euclidean distance to target for non-word imitation task, plotted by learner/non-learner status from (a), and across native/non-native item.

Fig. 3. fMRI results. (a) t-contrasts for main effects of listen pre-imitate (all listen pre-imitate > rest; left) and imitation (all imitation > rest; right). (b) t-contrasts for main effects over levels of imitation. (left) main effects of nativeness on activation during imitation: non-native imitation > native imitation (top); native imitation > non-native imitation (bottom). (right) main effects of complexity on activation during imitation: trisyllable > vowel (t-contrasts specified outside 2 x 3 flexible factorial ANOVA). (c) Significant nativeness x complexity interaction for listen pre-imitate (from 2 x 3 flexible factorial ANOVA). Interaction manifests at bilateral superior temporal plane, revealing modulatory effects of nativeness within the vowel and trisyllable conditions (mean±SEM parameter estimates plotted as inset; see Materials and Methods for details of parameter estimate extraction). Note: F statistics displayed on surface. (d) t-contrasts for listen pre-imitate non-native > native, across each level of complexity (t-contrasts specified outside 2 x 3 flexible factorial ANOVA). Note absence of superior temporal activation for monosyllables. All effects presented at cluster level false-discovery rate corrected $q < 0.05$ (achieved with voxel height and extent threshold, $p < 0.0015$, k=50).

Fig. 4. Individual differences analyses for sensorimotor transformation (ST). Pre-scan acoustic performance (x-axes - 2D Euclidean distance-to-target difference scores for /y/

from training; Fig. 2a, left) used as regressor for 'listen pre-imitate' parameter estimates for ST non-native > native contrast. Participant age was entered as a covariate of no interest in the model. Cyan clusters indicate regions where significant (FDR $q < 0.05$) positive linear fits emerged (with exception of * - marginal FDR $q < 0.065$). Learner/non-learner status is marked with symbols as in Fig. 2a. Cluster peak co-ordinates (MNI space) reported in parentheses.

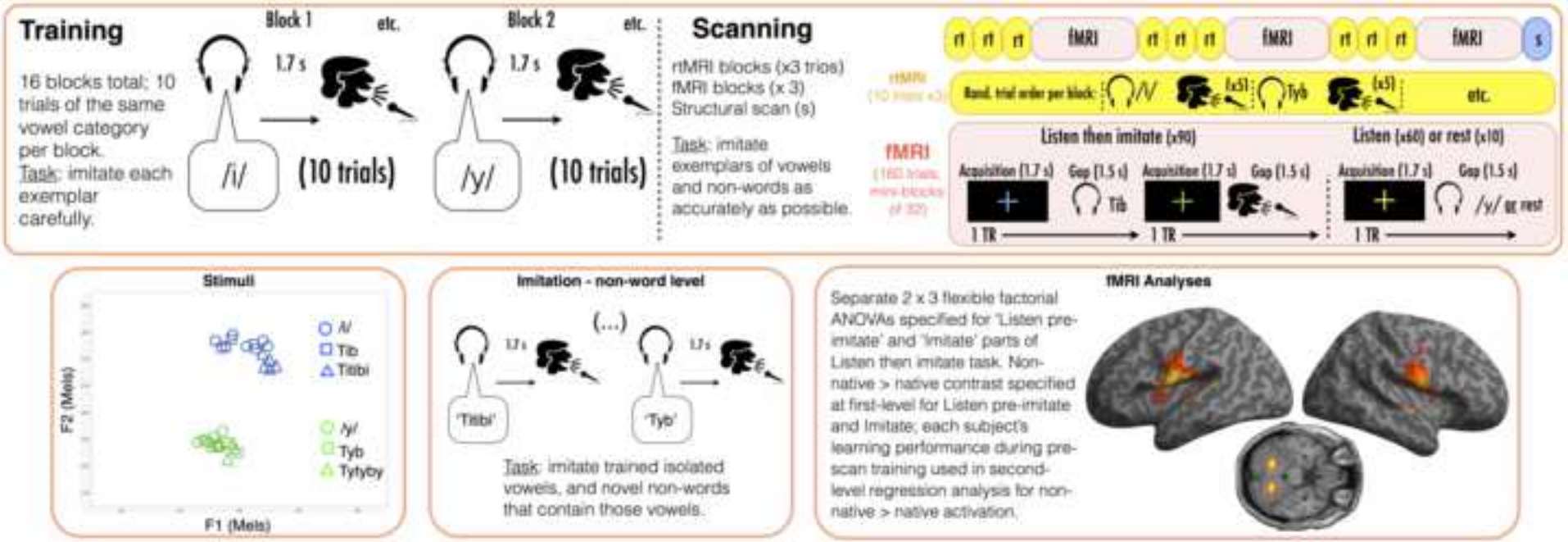
Supplemental figure 1. F1-F2 plot (average in Mels) for acoustic training data, for learners (squares) and non-learners (x symbols). Mean F1-F2 values for /i/ and /y/ training stimuli are shown as filled blue and green diamonds, respectively. Unfilled coloured symbols/coloured x symbols show the final F1-F2 values achieved by learners/non-learners over the latter four training blocks (see legend, right).

Supplemental figure 2. F1-F2 plot (average in Mels) for vowel and non-word imitation task, for learners (coloured symbols) and non-learners (black symbols). Mean F1-F2 values for /i/ and /y/ vowel, monosyllable and trisyllable stimuli are shown as filled blue and green symbols, respectively (see legend, top left). Unfilled colour symbols show the mean F1-F2 values achieved by learners over the full task; unfilled black symbols show the mean F1-F2 values achieved by non-learners over the full task (see legend, right).

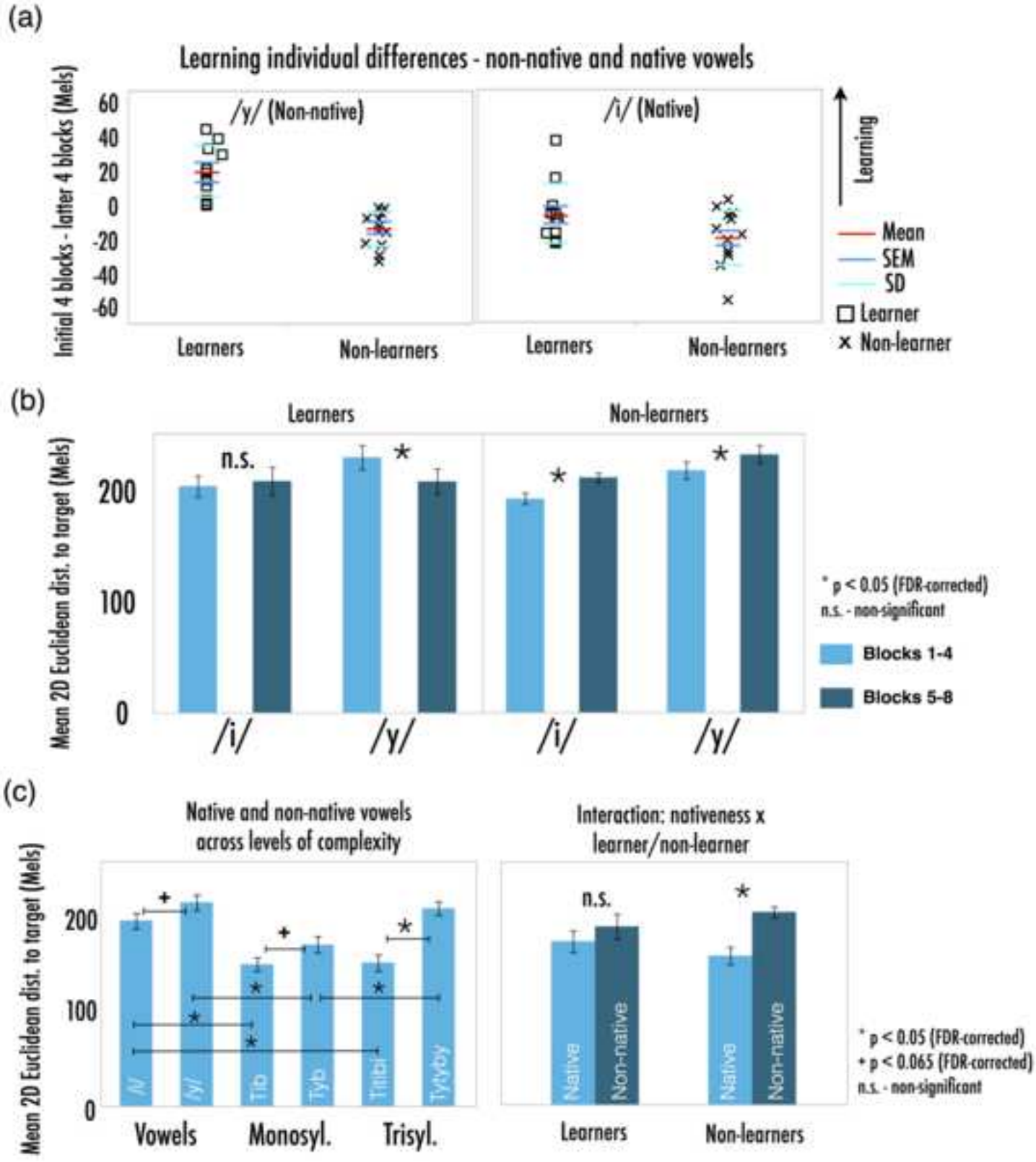
Supplemental figure 3. Individual differences analyses for sensorimotor transformation (ST), split by levels of item complexity. Pre-scan acoustic performance (x-axes - 2D Euclidean distance-to-target difference scores for /y/ from training; Fig. 2a, left) used as regressor for 'listen pre-imitate' parameter estimates for ST non-native > native contrast, specified at each level of complexity – vowel, monosyllable, trisyllable (y-axes). Participant age was entered as a covariate of no interest in all models. All other parameters as per Figure 4.

5. Figure

Click here to download high resolution image

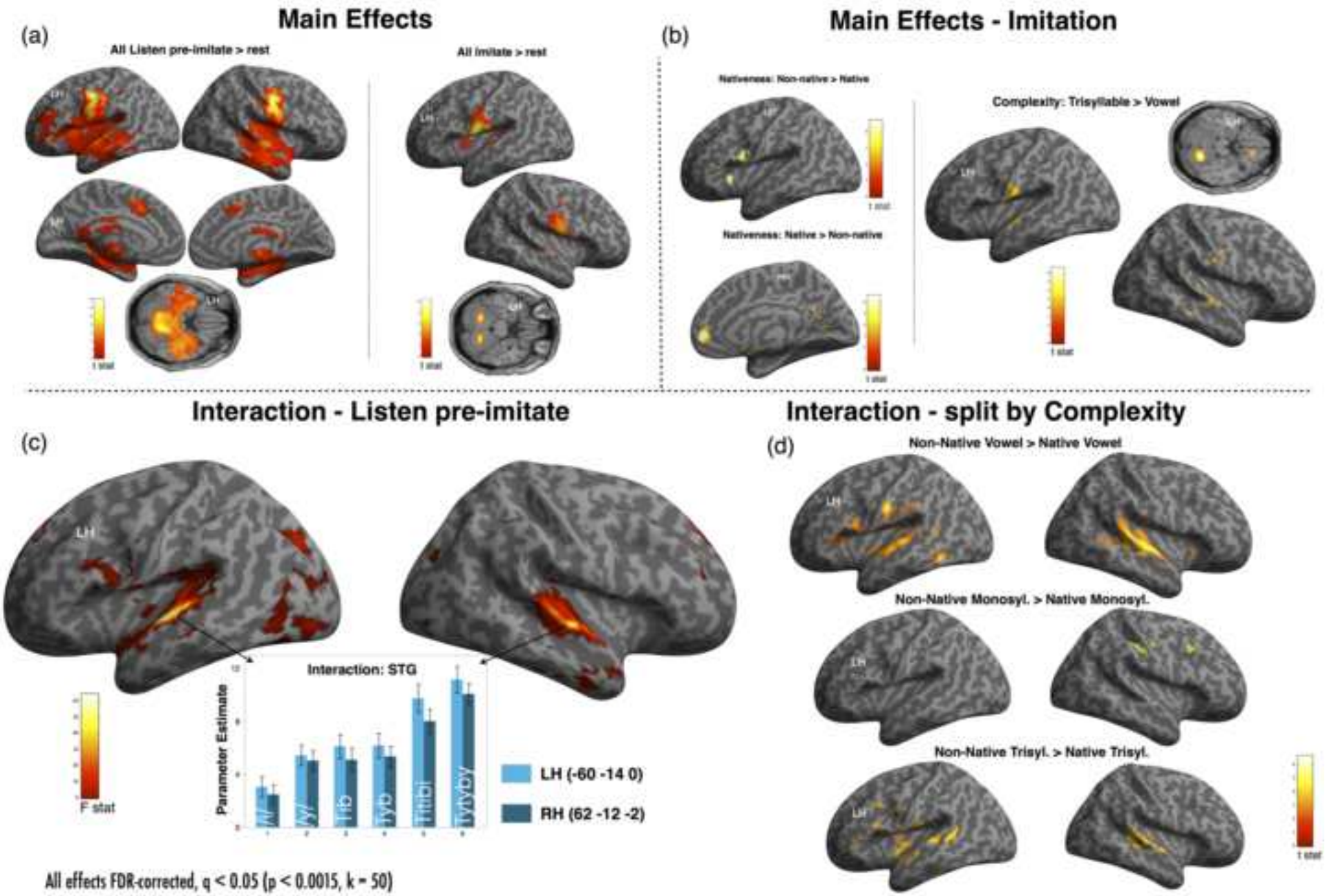


6. Figure
Click here to download high resolution image



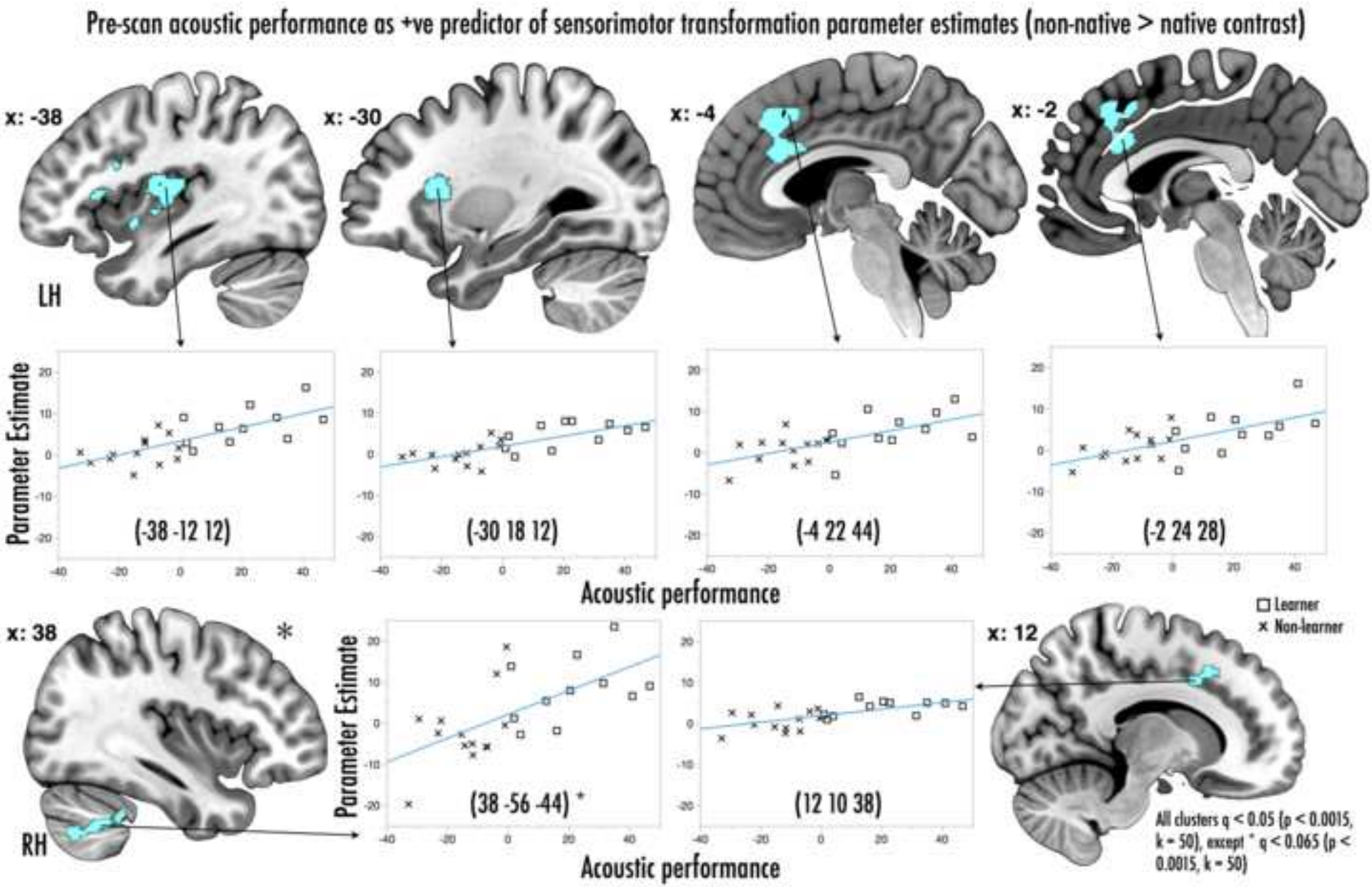
7. Figure

Click here to download high resolution image



8. Figure

Click here to download high resolution image



9. Supplementary Figure 1

Click here to download 10. Supplementary Material: Supplemental_Fig1.tiff

10. Supplementary Figure 2

Click here to download 10. Supplementary Material: Supplemental_Fig2.tiff

11. Supplementary Figure 3

Click here to download 10. Supplementary Material: Supplemental_Fig3.tiff