The response of the anterior striatum during adult human vocal learning

Anna J. Simmonds, Robert Leech, Paul Iverson, and Richard J. S. Wise

Computational, Cognitive and Clinical Neuroimaging Laboratory (C3NL), Division of Brain Sciences, Imperial College London, United Kingdom; and Department of Speech, Hearing and Phonetic Sciences, Division of Psychology and Language Sciences, University College London, United Kingdom

Submitted 20 December 2013; accepted in final form 2 May 2014

Simmonds AJ, Leech R, Iverson P, Wise RJ. The response of the anterior striatum during adult human vocal learning. J Neurophysiol 112: 792–801, 2014. First published May 7, 2014; doi:10.1152/jn.00901.2013.—Research on mammals predicts that the anterior striatum is a central component of human motor learning. However, because vocalizations in most mammals are innate, much of the neurobiology of human vocal learning has been inferred from studies on songbirds. Essential for song learning is a pathway, the homolog of mammalian cortical-basal ganglia “loops,” which includes the avian striatum. The present functional magnetic resonance imaging (fMRI) study investigated adult human vocal learning, a skill that persists throughout life, albeit imperfectly given that late-acquired languages are spoken with an accent. Monolingual adult participants were scanned while repeating novel non-native words. After training on the pronunciation of half the words for 1 wk, participants underwent a second scan. During scanning there was no external feedback on performance. Activity declined sharply in left and right anterior striatum, both within and between scanning sessions, and this change was independent of training and performance. This indicates that adult speakers rapidly adapt to the novel articulatory movements, possibly by using motor sequences from their native speech to approximate those required for the novel speech sounds. Improved accuracy correlated only with activity in motor-sensory perisylvian cortex. We propose that future studies on vocal learning, using different behavioral and pharmacological manipulations, will provide insights into adult striatal plasticity and its potential for modification in both educational and clinical contexts.

fMRI; vocal learning; striatum; non-native

Speech production is a complex motor-sensory process, involving rapid sequential motor movements. It relies on the integration of feedforward motor and feedback sensory signals, with online self-monitoring guiding rapid modification of motor commands to the larynx, pharynx, and articulators, requiring coordination of up to 100 muscles (Ackermann and Riecker 2010; Levelt 1989). This integration allows the maintenance of intelligible speech, even under adverse speaking conditions, and depends on motor (frontal), auditory (temporal), and somatosensory (parietal) cortex as well as the insular cortices, subcortical nuclei, and cerebellum (Argyropoulos et al. 2013; Gollinopoulos et al. 2010; Guenther 2006; Ventura et al. 2009).

Vocal learning covers a range of vocal behaviors including contextual learning (comprehension and usage) as well as production learning (Janik and Slater 2000), and the present study focuses on the latter. Children are remarkable in their ability to learn the complex motor sequences that allow them to speak (Guenther 1994). Acquisition of the native language (L1) occurs by matching speech sounds to articulatory gestures required to produce those sounds (Buchsbaum et al. 2005), starting with the babbling stage. By 1 yr, babbling turns to speech and the accent of the child is clearly that of a native speaker. In contrast, although humans are capable of vocal learning throughout life, when a second language (L2) is learned after L1 has been established, there is no babbling stage, and the learners acquire new words with the explicit knowledge about their meaning. There is also a strong tendency to translate a heard word in L2 into its corresponding word in L1 (Thierry and Wu 2007; Wu and Thierry 2010). Second language learners may feel that they have developed L2 sufficient for their purpose at a relatively early stage, and they may not strive to tune their auditory cortex to generate accurate, long-term representations of L2 syllables, words, and phrases as spoken by a native speaker. Therefore, the online perceptual monitoring of what they are producing is never sufficiently fine-tuned to drive improvements in the long-term motor representations of articulation, and late-acquired languages are spoken with an accent determined by the native speech (Feige 1995; Simmonds et al. 2011).

Repeating non-native bisyllabic words involves generating previously unlearned sequences of movements of the larynx and supralaryngeal articulators during the pronunciation of the first of the two syllables. Novel sensorimotor mappings for the non-native words depend on the accuracy of both the sensory “templates” created for each word and the neural coding for the new sequences of motor commands (Gollinopoulos et al. 2010; Guenther 2006; Guenther and Vladusic 2012). Although the most important sensory component may be auditory, somatosensation also plays a role in the control of voice production (Lametti et al. 2012; Simmonds et al. 2014).

Vocal learning is not unique to humans and is particularly evident in parrots and oscine songbirds (Mooney 2009; Petkov and Jarvis 2012). A limited capacity for vocal learning may also exist in mice (Arriaga and Jarvis 2013). During a critical period, hatchling songbirds create memories of target songs by listening to a tutor bird. By trial-and-error experimentation, they then modify their vocal output until the auditory feedback matches the memorized auditory templates (Olveczky et al. 2005), similar to the mechanism that operates in human infants as they learn to speak (Doupe and Kuhl 1999). A recent study has also demonstrated that both songbirds and human infants take a stepwise approach to the acquisition of vocal transitions, rather than undergoing a single a developmental shift (Lipkind et al. 2013). Many songbirds, including the most commonly studied song learning bird, the zebra finch, learn to sing their single song as a juvenile, and this song remains constant
throughout life (Eda-Fujiwara et al. 2012). Although this is clearly more limited than the unbounded nature of human speech, other species, such as Northern mockingbirds, continue to learn, even in adulthood, and the song repertoire of the brown thrasher can extend into the thousands (Boughey and Thompson 1981; Kroodsma and Parker 1977). Therefore, research on the genetic, molecular, and neural mechanisms underlying song acquisition in birds may inform human speech acquisition (Bohns et al. 2010; Brainard and Doupe 2002; Doupe and Kuhl 1999; Doupe et al. 2005; Enard 2011; Jarvis 2007). This has been made possible by a clearer understanding of the anatomic homologies between avian and mammalian brains (Jarvis et al. 2005). The anterior striatum is the homolog of avian “area X,” the basal ganglia region central to song learning. Neuroscientists studying vocal learning in songbirds have predicted that the human anterior striatum plays a prominent role in infant vocal learning (Jarvis 2004, 2007).

The present study used functional magnetic resonance imaging (fMRI) to investigate the neural networks supporting vocal sensorimotor learning in adults, with an emphasis on anterior striatal function. The study observed the effects of 1 wk’s training on the pronunciation of non-native words, changes in brain activity during and between the scanning sessions performed before and after training, and the relationship between performance and brain activity. It was expected that training on repeating non-native words would result in an increase in accuracy of behavioral performance. However, whereas it may reasonably be assumed that infants maintain anterior striatal activity during the acquisition of their first language, the hypothesis was that adults would demonstrate a rapid decline in the anterior striatal activity as they quickly adapted to the novelty of the new speech motor sequences, mapping them onto their existing prepotent motor sequences for native speech.

METHODS

Participants

Twenty-one right-handed monolingual native speakers of English (10 women), average age 26 yr (range 19–40 yr) participated after giving informed written consent. The study was approved by the local research ethics committee. In the second scanning session there were technical difficulties in one of the three runs for three separate participants, so those runs were excluded from the analyses. In total, data were analyzed from 6 runs for 18 participants and from 5 runs for 3 participants. Participants were paid £10 per hour for the 5 training sessions, each lasting 1 h, giving a total payment of £50 per participant. The subjects were aware that they would be paid irrespective of their response to training; that is, the final reward was not linked to improved non-native word repetition.

Stimuli

For the non-native condition the stimuli came from three different languages, Mandarin, Spanish, and German, based on four phonemes per language (Mandarin: tones 1, 2, 3 and 4; Spanish: /l/, /r/, /j/, /y/; German: /u/, /ü/, /o/, /ö/). The non-native stimuli were specifically chosen to manipulate a different place and manner of articulation in each language (by manipulating either pitch, vowel sounds, or consonants). The words were real bisyllabic words with a consonant-vowel-consonant-vowel (CVCV) structure and were matched for number of phonemes, with the target phoneme and stress on the first syllable and the rest of the word easy to produce for native English speakers (e.g., Mandarin: “bu4 cha1”; Spanish: “jaba”; German “Höhle”). The native stimuli consisted of bisyllabic non-words, also with a CVCV structure and matched for number of phonemes, starting with either /bl/, /pl/, or /l/ (e.g., “beetoo”). Native speakers of each language, one man and one woman, produced the stimuli. Audiovisual recordings were made of the stimuli for inclusion in the training materials, but audio-only versions were used in the scanning sessions. There were four different sounds from each non-native language. Of the four sounds for each language, the pronunciation of two was trained and that of the other two was untrained. For Mandarin, the words involved four different tones (trained: t1 and t4; untrained: t2 and t3); for Spanish, four different consonants were used (trained: /l/ and /y/; untrained: /d/ and /r/); and for German, four different vowel sounds (trained: /u/ and /ü/; untrained: /o/ and /ö/). The native sounds were all untrained (/b/, /p/, /t/) because despite being a novel combination of phonemes (since the non-words would not have been produced by the participants before), the sounds themselves were all highly familiar to the native English participants and no training was required.

Off-Line Pronunciation Training

In between the two scanning sessions, participants undertook a week of computer-based training, 1 h per day, with different exercises each day. Training materials were developed specifically for the purposes of this study. The first part of the training introduced participants to the main articulators involved in speech production, using interactive exercises and multiple-choice tests. The next three sessions looked at each of the speech sound groups (Spanish consonants, German vowels, and Mandarin tones) in turn, with the two speech sounds for each language. The non-native speech sounds were also presented using interactive exercises, including learning of the main articulators used in each sound and phoneme discrimination tasks. In addition, participants watched video clips of a native speaker producing the sound (1 audiovisual presentation of each word). Participants were instructed to record their repetition attempts and were given guidance on creating spectrograms of their speech and comparing those to spectrograms from the native speakers (1 example of each word). The final training session involved many repetitions of all the non-native speech sounds. The training focused on producing the sounds multiple times, with at least 15 attempts on each word. A listen-and-repeat paradigm was used in training, with participants recording their attempts to confirm that they complied with the training.

Off-Line Phoneme Discrimination Testing

With such unfamiliar phonemes as used in this study, it is possible that participants would not be able to discriminate between different sounds and might produce one attempt for a variety of phonemes. To ensure that participants were able to discriminate between the phonemes, their perception of the different sounds used in the study was tested, using pairs of non-native words in a forced-choice same/different task. This test was performed before the first scanning session, and therefore prior to any training.

Functional MRI Experimental Tasks

In both fMRI scanning sessions, participants listened to and repeated native non-words and non-native words. There were two main conditions: “listen-and-repeat,” presented in this report, and “listen only,” not included in this report. A baseline condition of “rest” was also included, during which the subjects fixated on a crosshair without any additional task. During speech trials participants saw two white circles on the screen and heard a word (or non-word). Participants were instructed that if the two circles remained unchanged, it was a listening-only trial and they were not required to speak. If the second
circle turned black, that was the cue to repeat the word. One important point to note from the stimuli presentation design is that when the participants first heard the stimuli, they did not know whether the trial was for a listen-only or a listen-and-repeat condition. In each run (3 in each of the 2 scanning sessions), there were 20 repeat trials for each of the non-native language groups, 10 native repeat trials, and 15 rest trials. For the non-native words, both trained and untrained words were included. In the first session there were words the participants were going to be trained on and words they were not going to be trained on. In the second session, there were words the participants had been trained on (“trained”), along with words they had not been trained on (“untrained”). For the repeat trials, participants heard the word (2,500 ms) and were then cued to repeat it (2,500 ms), followed by a fixation cross (2,200 ms) while the scan was acquired. The fixation cross then remained on the screen until the start of the next trial. Also included in this paradigm were 20 listening trials, which are not included in the present analyses (6 listen trials for each non-native language group and 2 native listen trials per run).

Functional MRI acquisition and analysis. MRI data were obtained on a Philips Intera 3.0-Tesla scanner, using dual gradients, a phased-array head coil, and sensitivity encoding with an undersampling factor of 2. Thirty-two axial slices with a slice thickness of 3.25 mm and an interslice gap of 0.75 mm were acquired in ascending order (resolution, 2.19 × 2.19 × 4.00 mm; field of view, 280 × 224 × 128 mm). Quadratic shim gradients were used to correct for magnetic field inhomogeneities within the anatomy of interest. There were three runs, each of 105 volumes. Functional MR images were obtained using a T$_2$-weighted, gradient-echo, echoplanar imaging (EPI) sequence with whole brain coverage (repetition time = 8 s, acquisition time = 2 s, giving 6 s for the subjects to speak during silence; echo time = 30 ms; flip angle = 90°). A “sparse” FMRI design was used to minimize movement- and respiratory-related artifact associated with speech studies (Hall et al. 1999), as well as to minimize auditory masking. High-resolution, T$_1$-weighted images were also acquired for structural reference. Stimuli were presented visually using E-Prime (Psychology Software Tools) run on an IFIS-SA system (In Vivo).

Functional MRI was analyzed with FEAT (FMRI Expert Analysis Tool) version 5.98, using standard within-subject and across-subject analysis pipelines and default settings unless otherwise specified. Preprocessing included motion correction using MCFLIRT (Jenkinson et al. 2002), non-brain removal using BET (Brain Extraction Tool) (Smith 2002), spatial smoothing using a Gaussian kernel of FWHM (full width half-maximum) 5 mm; grand-mean intensity normalization of the entire four-dimensional (4D) data set by a single multiplicative factor; high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 5.00$ s). The blood oxygen level-dependent (BOLD) response was modeled using a double-gamma hemodynamic response function (Glover 1999). In addition, the FSL tool Motion Outliers was used, which identifies time points with a high amount of residual intensity change from the motion corrected data. A confound matrix of outliers is created, and this is included in the FEAT analysis as an additional confound variable. This is more beneficial than simply removing volumes with high levels of motion because it does not require any adjustments to the rest of the model with regard to timing, and it correctly accounts for signal changes on either side of the excluded time point, as well as correctly adjusting the degrees of freedom. Time-series statistical analysis was performed using FILM (FMRIB’s Improved Linear Modeling) with local autocorrelation correction (Woolrich et al. 2001). Z (Gaussianized TIF) statistic images were thresholded using clusters determined by $z > 2.3$ and a corrected cluster significance threshold of $P = 0.05$. Registration to high-resolution structural and standard space images was carried out using FLIRT (FMRIB’s Linear Image Registration Tool) (Jenkinson et al. 2002; Jenkinson and Smith 2001). Higher-level group analysis was carried out using FLAME (FMRIB’s Local Analysis of Mixed Effects) stage 1 (Beckmann et al. 2003; Woolrich et al. 2004).

Data Analysis

Speech ratings. Speech recordings from participants’ scanning sessions, presented randomly across participants and sessions, were rated by two native speakers of each of the three languages (6 individuals). These independent native speakers listened to each speech trial for their respective language for each participant and gave it a score out of seven. A 7-point scale was used to judge the degree of native-like performance (1 = non-native-like, very strong foreign accent, 7 = native-like, no foreign accent). The speech score was calculated by each utterance from each participant being given a score from each of the two native raters, and the mean value for each trial was calculated. These speech trials were then grouped according to the specific speech sound involved (from the 4 in each language), and a group mean for each speech sound was calculated, before combining the two trained sounds and two untrained sounds for each language. The mean score from the two raters for the trained and for the untrained sounds for each participant was used in further analyses. Repetition performance during the scanning sessions was very high, at over 94% for all participants. Rare trials in which five of the participants failed to produce a word during a repeat trial were excluded from further analyses.

Filtering out CSF and WM. As well as removing variance associated with the six motion variables and the motion outliers on an individual basis, variance associated with the time courses of white matter (WM) and cerebrospinal fluid (CSF) from the whole brain functional data was removed using ordinary least-squares linear regression. To calculate the time course for the WM and CSF, a 3-mm-radius sphere was created based on the MNI coordinates $-26, -22, 28$ and MNI 2, 10, 8, respectively, and the mean time course across the sphere was calculated.

Regions of Interest

Following a whole brain 2 × 2 ANOVA looking at language (native and non-native) and session (pre- and posttraining), the strongest activation for the language × session interaction was observed in the basal ganglia. Region of interest (ROI) analyses using anatomic putamen and caudate masks from the Harvard-Oxford probabilistic atlas were subsequently carried out. A single voxel was selected from within these masks, in each hemisphere separately, for the anterior putamen (left: $-22, 8, -2$; right: $18, 8, -2$) and the caudate (left: $-14, 12, 12$; right: $14, 12, 12$). A 5-mm sphere was then created around each of those four voxels. The spheres were then constrained by the interaction functional result to create the four ROI masks.

RESULTS

Twenty-one participants were required to repeat bisyllabic non-native words taken from three languages (German, Spanish, and Mandarin). There was no training on the meaning of the words, to exclude parallel lexical semantic learning. Each word had a target phoneme in the first syllable, with the second syllable represented in the native (English) language. The target phonemes differed in terms of the vowel (German), consonant (Spanish), or pitch (Mandarin). Four phonemes were chosen for each language, and all were included in two scanning sessions. Functional MRI was performed before and after 1 wk of training on the pronunciation of two target phonemes per language, with no training on the other two phonemes. The baseline task was repeating bisyllabic non-words formed from native (English) syllables. No online feedback was provided during the repetition trials.

J Neurophysiol • doi:10.1152/jn.00901.2013 • www.jn.org
Behavioral Results

Off-line phoneme discrimination. Before the first scan, a within-language pairwise comparison task (chance score 50%) determined whether the participants could discriminate between the phonemes employed in the study. The mean score across all languages was 91.4% correct (range 72.0–98.0%). The scores for German and Mandarin were identical (93.3%; range 58.8–100%), although individual participants were better at either German or Mandarin. The scores were only a little lower for Spanish (mean 87.5%, range 68.7–100%). Because performance was above chance, and the most common type of error was misidentifying a matching pair as a mismatch, this behavioral measure demonstrated that participants were able to perceive the differences between unmatched pairs. Although this is not a measure of the accuracy of the auditory memories (“templates”) the participants created, these scores indicated that the subjects were able to detect acoustic differences between the phonemes.

Rating scores of online speech production. Repetition performance during the scanning sessions was very high, at over 94% for all participants. Sixteen participants repeated back each non-native word and native non-word that they heard. Five participants occasionally failed to produce a word during a repeat trial, and the data for these trials were excluded from the fMRI analyses. Speech recordings from each participant at both scanning sessions were presented randomly to two native speakers for each of the three languages (see Methods). The raters scored accuracy of pronunciation on a 7-point scale, from 1 (very strong foreign accent) to 7 (native pronunciation). The inter-rater reliability was significant for each of the three languages (Mandarin: \( r = 0.594, P < 0.001 \); German: \( r = 0.543, P < 0.01 \); Spanish: \( r = 0.483, P < 0.01 \)). Therefore, the mean of the two raters’ individual scores for each language and each participant was used in the analyses (Fig. 1).

Performance in the two sessions varied for the three different non-native languages. Highest scores were for the German vowels, closely followed by the Mandarin tones, with Spanish consonants scoring the least (Fig. 1). This pattern was the same in both sessions. For averaging across all sounds in each language group, there were significant improvements for all non-native languages. Highest scores were for the German language, there was a small but significant improvement in pronunciation between scanning sessions. On the untrained words, there was no significant differences across session for the untrained sounds. The table (bottom) presents post hoc 2-tailed paired t-tests following a 3 (Language) x 2 (session) x 2 (training) ANOVA for the behavioral scores from the functional magnetic resonance imaging (fMRI) experimental task.

![Image](http://jn.physiology.org/)

Fig. 1. Speech scores from all languages in both sessions. Mean scores from session 1 are shown in black and from session 2 in gray. Error bars display 95% confidence intervals. A: average scores for the 3 languages in both sessions. One-tailed paired t-tests revealed significant improvements across session for all languages (\( P < 0.05 \)). B: scores for the trained sounds only for each of the 3 languages. One-tailed paired t-tests revealed significant improvements for the trained sounds in Spanish (\( P > 0.005 \)) and Mandarin (\( P < 0.005 \)). C: scores for the untrained sounds only. There were no significant differences across session for the untrained sounds only. The table presents post hoc 2-tailed paired t-tests following a 3 (Language) x 2 (session) x 2 (training) ANOVA for the behavioral scores from the functional magnetic resonance imaging (fMRI) experimental task.
improvement for any language, although there was a trend ($P = 0.08$) for improvement on the German words.

Within the first session, comparing the first and third runs, there was a significant improvement for the non-native languages averaged together ($P = 0.05$), which was driven by improvements for Mandarin ($P = 0.001$). Spanish and German did not show any within-session significant changes.

Overall, the behavioral results demonstrated that participants were able to detect acoustic differences between the non-native phonemes included in the study, and there was a small but significant improvement in pronunciation between scanning sessions.

**Imaging Results**

There were two types of trial, listen only and listen-and-repeat (see Methods). Only the data from the repeat trials are presented in this report. A visual cue signaled to the participant to repeat the non-native word or native non-word that had just been heard, after which the scan data were acquired over 2 s. A low-level baseline condition, when the participants fixated on a crosshair, was also included.

This study was designed to investigate training effects of producing non-native speech sounds. A 2 (language: native and non-native, each against the silent rest baseline, with trained and untrained word sets combined) × 2 (session: pre- and posttraining) factorial design was used, and a whole brain ANOVA was performed. The main effect of language was observed in extensive bilateral cortical and subcortical regions (Fig. 2Ai). In addition to activity in bilateral auditory, premotor, and sensorimotor cortices and in the basal ganglia, thalamus, and cerebellum, much of which can be attributed to sensory-motor processes involved in hearing and repeating a word, there was activity in two separate task-dependent, domain-general networks (Dosenbach et al. 2007, 2008; Seeley et al. 2007). The first, the cingulo-opercular or salience network (Fig. 2Bi), encompasses, in both cerebral hemispheres, the midline dorsal anterior cingulate cortex and adjacent superior frontal gyrus, anterior lateral prefrontal cortex, frontal operculum and adjacent anterior insular cortex, and thalamus. The second, the frontoparietal or central executive network (Fig. 2Bii), encompasses, in both cerebral hemispheres, the dorsolateral prefrontal cortex, dorsal inferior parietal cortex and adjacent intraparietal sulcus, middle cingulate cortex, and precuneus. The cerebellar hemispheres are functionally connected to these two networks. In contrast, the structures showing both a main effect of session (Fig. 2Aii) and a language × session interaction (Fig. 2Aiii) were the left and right basal ganglia. The main peaks for the interaction were located in the left and right caudate nuclei and anterior putamina (Fig. 2C).

When 2 × 2 ANOVAs were performed separately for the trained and untrained words in all three non-native languages as one level of the language factor and the native non-words as...
the other, the language × session interaction was observed for both word sets. Separate ANOVAs for each non-native language (combined trained and untrained word sets) with the native non-words confirmed a language × session interaction in the anterior striatum for all three languages, although for Spanish it was only visualized when the statistical threshold was reduced to a $z$ value of 2 (Fig. 2D).

Spherical ROIs, 5 mm in diameter, were used to demonstrate the profiles of activity across conditions in the anterior striatum. The positioning of the ROIs, described in methods and illustrated in Fig. 3A, was based on the whole brain ANOVA. The only statistics performed on these profiles were to investigate within-session and between-language effects, although 95% confidence intervals for the effect sizes are included for illustrative purposes. Activity in the anterior striatum was present during the repetition of native non-words, relative to the baseline condition of rest, but was considerably greater during the repetition of non-native words. These profiles also illustrated the sharp decline in activity between sessions 1 and 2 during the repetition of non-native words (Fig. 3B).

To further investigate how rapidly changes in anterior striatal activity occurred, activity across the three runs in each of the two scanning sessions were investigated on the ROI data (Fig. 3C). To limit the number of statistical comparisons, only the data from the left striatal ROIs (caudate nucleus and anterior putamen) were analyzed. Language (native non-words and non-native words) × run (first and second) ANOVAs revealed that there was a language × run interaction in both sessions, although less marked in session 2. Thus, in session 1, there was a significant interaction in anterior putamen [$F_{(1,20)} = 23.6, P < 0.001$] and caudate nucleus [$F_{(1,20)} = 9.4, P < 0.01$], with a corrected level of significance set at $P = 0.01$. In session 2, the interactions were significant at an uncorrected level in anterior putamen [$F_{(1,17)} = 6.2, P < 0.05$] and caudate nucleus [$F_{(1,17)} = 4.85, P < 0.05$]. In both sessions, paired $t$-tests confirmed that activity when native non-words were repeated, relative to that at rest, did not change across runs ($P > 0.5$).

### DISCUSSION

This study required participants to learn to pronounce single non-native words, with repeated practice on the required articulatory movements and associated self-corrections based on anticipated and actual sensory feedback. This resulted in overall improvement on the trained words, but with variable success. The prescanning phoneme discrimination task demonstrated that participants could detect auditory differences be-

---

**Correlations of imaging results with speech score.** To investigate how changes in brain activity were linked to improvements in behavioral score, the contrast of non-native repeating $\gg$ rest difference across sessions was run with each participant’s individual speech score difference as an additional variable. It was assumed that activity during the rest condition did not change between sessions. Changes in activity in response to repeating non-native words between the two sessions were correlated with changes in the raters’ scores of proficiency at pronunciation for all words, trained and untrained. When all non-native languages were considered together, no significant linear increases in activity were observed, but activity declined in a number of cerebral and cerebellar cortical regions, the main clusters being located in left central operculum (ventral motor-sensory cortex), anterior and posterior left superior temporal gyrus, and right lateral cerebellum (Fig. 4). Within this network of regions with changes during repeating in response to proficiency, each language contributed differently. There was some overlap between Spanish and German within left primary sensorimotor cortex and left frontal operculum, whereas improvement in Mandarin sounds was responsible for reduction in activity in left anterior superior temporal sulcus. Improvement in the Spanish sounds correlated with a reduction in activity in left posterior superior temporal sulcus. Figure 4B presents the whole brain correlations for individual languages, with correlation plots for the peak voxels. The correlation plots in Fig. 4B confirm that as the speech score improved, activity decreased.

---

**Fig. 3. Regions of interest (ROI) for native and non-native repeating.** A: ROI masks are shown in both hemispheres for the anterior putamen (red) and caudate (blue) on a coronal slice ($y = 10$). ROI masks are displayed on a standard brain template (MNI152). L, left; R, right. B: across-session effects are shown with the blood oxygen level-dependent (BOLD) response for non-native (black and gray) and native (open and hatched) repeating in both sessions (session 1, black and open; session 2, gray and hatched). Error bars represent 95% confidence intervals. Caudate ROIs are shown in blue, anterior putamen in red. C: within-session effects are shown with the mean time series for native and non-native repeating in each of the 3 runs for both scanning sessions for the left (i) and right caudate (ii) and the left (iii) and right anterior putamen (iv).
bween the non-native speech sounds included in this study, with performance near ceiling for Mandarin and German and well above chance for Spanish. Therefore, we have concluded that the participants were able to form accurate auditory templates of the non-native speech sounds they were required to repeat. Based on this assumption, the vocal training the participants received was directed at the formation of these auditory memories and then modification of their speech output pathways to match more closely accuracy of articulation to these novel templates.

The most striking finding was the sharp focal decline in activity in the anterior striatum over time. However, the results indicate that this decline was dissociated from individual variability in performance and training. The change in anterior striatal function was not a nonspecific effect related to rapid habituation to task novelty, because repeating native non-words was also novel. Although repeating bisyllabic native non-words required only habitual sequential movements to produce each syllable, the combinations of syllables had not previously been encountered by the participants. Although the response of the anterior striatum during native non-word repetition was less than that to non-native word repetition, activity was significantly greater than during the rest condition. Importantly, it did not change across runs in the first session.

Fig. 4. Whole brain correlation with speech score. A: results for all non-native languages averaged together. Top: 2 sagittal slices of a standard brain template (MNI152) are shown, $z = -52$ and 52, revealing activity in bilateral frontal opercula (1), left primary sensorimotor cortex (2), and anterior and posterior left superior temporal gyrus (3). Z-statistic images were thresholded using clusters determined by $z > 2.3$ and a corrected cluster significance threshold of $P < 0.05$. Bottom: axial slices ($z = -34, -10, 6, 26, 50$) with an orthogonal sagittal slice. B: whole brain correlations with mean speech score improvements across sessions for individual languages. Activity for Spanish consonants is shown in cyan, Mandarin tones in violet, and German vowels in green. Activity is shown on axial slices ($z = -34, -10, 6, 26, 40$) with an orthogonal sagittal slice. Correlation plots at right show the BOLD difference across sessions and the speech score improvement for each language separately ($n = 21$). The peak voxel correlations between the BOLD difference across sessions and the speech score improvement were $r = -0.77$ for Spanish consonants (voxel coordinates $-58, -14, 14$), $r = -0.81$ for Mandarin tones (voxel coordinates $-64, -10, -6$), and $r = -0.59$ for German vowels (voxel coordinates $-40, -26, 28$).
No previous functional neuroimaging study in humans has demonstrated novel language learning in relation to striatal activity. Evidence from bilingual studies has demonstrated that the striatum is involved in non-native speech production, but previous studies have investigated languages that have already been acquired. The caudate has been shown to be involved in the cognitive control of multiple languages and in language switching (Abutalebi and Green 2007; Costa and Santesteban 2004; Crinion et al. 2006; Hernandez et al. 2000; Price et al. 1999; Rodriguez-Fornells et al. 2002) and the putamen in articulation (Abutalebi et al. 2013; Frencq-Mestre et al. 2005; Golestani et al. 2006; Klein et al. 1994, 1995, 1999). Activity in the left putamen may be related to language proficiency, with increased activity observed when controlling articulation in a less proficient language (Abutalebi et al. 2013), and may also be attributed to the precise motor timing of speech output, which is less automatic and more “effortful” in a language acquired later in life (Klein et al. 1994).

The analogy from the birdsong literature about human language learning relates specifically to infants acquiring their first language (Jarvis 2004). One obvious difference between the infant and adult human brain is the absence and presence of prepotent articulatory motor sequences, ones that are habitual to the adult. Therefore, if it were possible to perform fMRI scans on awake babbling infants, it is expected that anterior striatal activity would be maintained across scanning runs and sessions. The results presented here on adult brains are compatible with initial activation of the anterior striatum in response to the task demand to attempt completely novel articulatory motor sequences, followed by a rapid decline as the new sequences are incorporated, as an approximation, to the habitual motor sequences of the first language. This is in accordance with the observation that once birdsong has been acquired, ablation of area X has only a small impact on continued “crystallized” song production.

However, even though the role of area X is reduced in crystallized song production, the use of gene expression studies and electrophysiology has shown that area X, a core component of the anterior forebrain pathway (AFP), is still active during singing in adulthood (Hilliard et al. 2012; Jarvis 2004). The output from the AFP to the dorsal song production pathway introduces trial-to-trial variability and is required to make small changes in adult song (Goldberg et al. 2012), and it has been shown that variability in the neural activity in the AFP correlates with performance variability (Kao et al. 2005). Although it is difficult to directly compare the fMRI activation with gene expression and electrophysiological measures, this is one possible difference between humans and song learning birds.

The rapid decline in anterior striatal activity within and between scans by adult vocal learners may have demonstrated a rapid “habituation” to the pronunciation of non-native words even when there had been no objective improvement of articulatory performance. This was probably the product of imperfect self-monitoring, resulting in a form of motor learning in which the learners settled for a less than proficient (native-like) performance. This finding may account for the persistence of a skill self-monitoring, resulting in a form of motor learning in regulatory performance. This was probably the product of imperceptible articulatory improvements even when there had been no objective improvement of articulation. This finding may account for the persistence of a skill self-monitoring, resulting in a form of motor learning in regulatory performance. This was probably the product of imperceptible articulatory improvements even when there had been no objective improvement of articulation.

The whole brain ANOVA demonstrated a main effect of repeating non-native words, one that involved extensive bilateral cortical and subcortical, cerebral, and cerebellar regions, comprising motor, sensory, and higher order systems. This replicates the finding from our previous work on bilingual speech production (Simmonds et al. 2011). As well as premotor and sensorimotor regions, there was activity in cingulo-opercular (salience) and frontoparietal (central executive) networks (Dosenbach et al. 2007, 2008; Seeley et al. 2007). These domain-general regions were strongly active during the repetition of non-native words relative to activity during the repetition of native non-words. It has been proposed that the frontoparietal network initiates and adjusts control during each trial of a task, whereas the cingulo-opercular network maintains task set over many trials, so-called adaptive and stable task control, respectively (Dosenbach et al. 2008). Evidence from other studies that have investigated temporal ordering and “chunking” of novel motor sequences (whether the movements involve the fingers or the articulators) also observed activity in bilateral inferior frontal and parietal networks (Boland and Guenther 2006; Jubault et al. 2007; Koechlin and Jubault 2006; Majerus et al. 2006). The greater activity in these frontoparietal systems during non-native word repetition did not change significantly after 1 wk of training, for either trained or untrained words. Therefore, overall cortical “effort” was not modulated by either familiarity with the repetition task or modest improvement in performance on the trained words.

Correlations between behavioral improvement and changes in activity across sessions for all languages together were observed only in cortical regions, the left and right frontal opercula, left ventral sensorimotor cortex, left superior temporal sulcus (STS), and right lateral cerebellum. The specific requirements of the different types of non-native sounds resulted in different patterns of activation for the correlations between the BOLD response and improvement in behavioral performance. Improvement in producing the novel pitch shifts associated with the tonal words of Mandarin correlated with a reduction in activity in the anterior STS, known to be sensitive to the intelligibility of an auditory stimulus. In contrast, improvement in behavioral performance on the novel vowel sounds in German and novel consonant sounds in Spanish, probably requiring greater dependence on the subtile placement of the articulators, correlated with a decline in activity in ventral sensorimotor cortex. However, precisely which aspects
of the speech production tasks result in specific reductions in activity in these regions must be speculative without further studies.

One prediction from this study is that first language learning in early childhood is associated with much more sustained anterior striatal activity, but this hypothesis will be difficult to determine given the limited cooperation of toddlers and the intimidating environment of MRI scanners. However, the results from this study suggest behavioral changes to training (e.g., external feedback with reward on each correct trial), and observations on subjects who demonstrate a wide range of proficiency at learning accurate pronunciation of a novel language may add more information about the modulation of anterior striatal function.

REFERENCES


