PHILOSOPHICAL TRANSACTIONS
OF THE ROYAL SOCIETY B
BIOLOGICAL SCIENCES

Singers show enhanced performance and neural representation of vocal imitation.

Journal:	*Philosophical Transactions B*
Manuscript ID	RSTB-2020-0399.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Waters, Sheena; Barts and The London NHS Trust, Wolfson Centre of Preventative Medicine Kanber, Elise; UCL Lavan , Nadine ; University College London, Belyk, Michel; University College London, Speech, Hearing and Phonetic Sciences; Carey, Daniel; Novartis Cartei, V; University of Chichester, Psychology; University of Lyon, Equipe de Neuro-Ethologie Sensorielle Lally, Clare; UCL Miquel, Marc; Queen Mary University of London, William Harvey Research Institute McGettigan, Carolyn; University College London Division of Psychology and Language Sciences, Speech, Hearing and Phonetic Sciences
Issue Code (this should have already been entered and appear below the blue box, but please contact the Editorial Office if it is not present):	VOICE
Subject:	Neuroscience < BIOLOGY
Keywords:	voice modulation, larynx, expertise, singing, speech, MRI

Author-supplied statements

Relevant information will appear here if provided.

Ethics

Does your article include research that required ethical approval or permits?:
Yes

Statement (if applicable):
The study was approved by the Ethics Committee at the Department of Psychology, Royal Holloway, University of London.

Data

It is a condition of publication that data, code and materials supporting your paper are made publicly available. Does your paper present new data?:
Yes

Statement (if applicable):
Supporting data are available on the Open Science Framework (https://osf.io/6pqkt/).

Conflict of interest

I/We declare we have no competing interests

Statement (if applicable):
CUST_STATE_CONFLICT :No data available.

Authors' contributions

This paper has multiple authors and our individual contributions were as below

Statement (if applicable):
CM designed the study. SW, EK, NL, MB, CL, and CM collected and/or analysed data. DC, MM, and VC developed bespoke tools for data collection and analysis. All authors contributed to the writing and approval of the manuscript.

PHILOSOPHICAL
TRANSACTIONS B

Phil. Trans. R. Soc. B.
doi:10.1098/not yet assigned

Singers show enhanced performance and neural representation of vocal imitation.

Sheena Waters[1,2,*], Elise Kanber[1,3*], Nadine Lavan[3,4], Michel Belyk[3], Daniel Carey[1,5], Valentina Cartei[6,7], Clare Lally[1,3], Marc Miquel[8,9], Carolyn McGettigan[1,3,^, †]

1 Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, UK;
2 Wolfson Institute of Preventive Medicine, Barts and The London School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK;
3 Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, UK;
4 Department of Biological and Experimental Psychology, Queen Mary University of London, Mile End Rd, Bethnal Green, London E1 4NS,
5 Novartis Global Service Center, 203 Merrion Rd, Dublin 4, D04 NN12, Ireland
6 Equipe de Neuro-Ethologie Sensorielle (ENES), Centre de Recherche en Neurosciences de Lyon, Université de Lyon / Saint-Etienne, 21 rue du Dr. Paul Michelon, 42100 Saint-Etienne, FRANCE
7 Department of Psychology, Institute of Education, Health and Social Sciences, University of Chichester, College Lane, Chichester, West Sussex, PO19 6PE, UK
8 Clinical Physics, Barts Health NHS Trust, London EC1A 7BE, UK;
9 William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK.

Keywords: Voice modulation, larynx, expertise, singing, speech, MRI, vocal tract

---

Summary

Humans have a remarkable capacity to finely control the muscles of the larynx, via distinct patterns of cortical topography and innervation that may underpin our sophisticated vocal capabilities compared with non-human primates. Here, we investigated the behavioural and neural correlates of laryngeal control, and their relationship to vocal expertise, using an imitation task that required adjustments of larynx musculature during speech. Highly-trained human singers and non-singer control participants modulated voice pitch and vocal tract length (VTL) to mimic auditory speech targets, while undergoing real-time anatomical scans of the vocal tract and functional scans of brain activity. Multivariate analyses of speech acoustics, larynx movements and brain activation data were used to quantify vocal modulation behaviour and to search for neural representations of the two modulated vocal parameters during the preparation and execution of speech. We found that singers showed more accurate task-relevant modulations of speech pitch and VTL (i.e. larynx height, as measured with vocal tract MRI) during speech imitation; this was accompanied by stronger representation of VTL within a region of right dorsal somatosensory cortex. Our findings suggest a common neural basis for enhanced vocal control in speech and song.

Introduction

Many cognitive, neural and physiological adaptations have been implicated in the evolution of human speech [1-3]. When comparing our species with the other great apes, one major distinction concerns the neural control of the larynx (or voice box). In humans, anatomical studies have revealed that the larynx receives innervation via direct connections from the primary motor cortex to the nucleus ambiguus, while in other apes this pathway is relatively more sparse, and in monkeys it is absent [4-7]. One hypothesis proposes that this direct pathway facilitates the rapidity and precision of laryngeal control in human speech and song, for example in the initiation of vocalisation, the fine tuning of vocal pitch and voice quality, and in switching between voiced and unvoiced segments of spoken words (e.g. consecutive consonants and vowels) [7-15].

---

^Author for correspondence (c.mcgettigan@ucl.ac.uk).
†Present address: Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, UK
*These authors contributed equally to the research

Researchers investigating the evolution of vocal behaviour in humans have been interested in measuring the acoustic correlates of laryngeal control through volitional vocal modulations. Two acoustic parameters have been particularly important in this endeavour: fundamental frequency (F0) and formant spacing (ΔF). F0 relates to the rate of vibration of the vocal folds in the larynx and is perceptually experienced as vocal pitch – in humans, adult males typically have longer and thicker vocal folds than adult females, and thus generate a lower F0 during speech. ΔF is related to the resonant properties of the vocal tract and covaries negatively with vocal tract length (VTL) – thus, adults show lower ΔF than children (whose vocal tracts are typically shorter), and adult male voices typically have lower ΔF than adult female voices due to the secondary descent of the larynx during puberty in human males. Previous research on speech acoustics has shown that humans can readily modulate ΔF in the appropriate direction when attempting to sound larger or smaller [16-17]. Similarly, adults and children will increase F0 and ΔF to sound more feminine, and will decrease these parameters to sound more masculine [18-20]. Such studies have provided crucial insights into the acoustic correlates of laryngeal control, although it should be noted that we are not aware of any study to date that has shown that human ΔF modulations are indeed achieved through changes in larynx height.

The ability to modulate the voice is potentially adaptive for individuals. For example, vocal size exaggeration is effective in changing listeners' evaluations of talker height [17], which may provide advantages in competitive situations. Furthermore, recent evidence on social trait expression has shown that talkers can volitionally modulate their speaking voice to generate exaggerated impressions of specific traits in naïve listeners [21]. Beyond the mere demonstration of vocal modulation in humans, it is of interest to investigate how this skill might vary across individuals. One way to do this is to investigate expert vocal performers, such as singers or voice artists. Formal training in singing involves enhanced training in the fine-tuned sensorimotor control required to support both solo and ensemble vocal performance [22], and a body of work has already demonstrated advantages for singers compared with non-singing controls in a range of language and accent imitation tasks (see [23]). Physiologically, proficient singing requires the efficient control of breathing, the coordination of laryngeal muscles to generate the optimal source signal, and further modulation of that source signal through fine control of vocal tract shape [24]. Thus, it could be predicted that expertise in singing might confer advantages for other vocal tasks requiring specific laryngeal muscle modulations, such as in the exaggeration of body size during speech.

Several studies have reported correlates of laryngeal muscle movements in the human neocortex. These include locations in the dorsal part of human ventral primary motor cortex [25-27], in addition to a more ventral site that may be evolutionarily common to humans and other primates [28-30]. Research in vocalizing humans has associated activation of dorsal larynx motor cortex (LMC) with three primary dimensions of laryngeal muscle activity: 1) adduction versus abduction of the vocal folds to allow phonation and non-voiced exhalation, respectively; 2) adjustments in vocal fold tension leading to changes in the fundamental frequency (F0) of the voice; and 3) vertical shifts in the position of the larynx to change the length of the vocal tract and thus the resonant properties of the voice through concomitant alterations in the formant frequencies (with functional MRI (fMRI): [25-27, 31]; with intracortical recordings/stimulation: [29-30, 32]). However, there are outstanding questions about what the neural activation patterns in speech motor cortex might represent (e.g. acoustic targets of speech or articulator kinematics [33]), and how these are coded during speech planning versus execution.

Primary motor cortex in the precentral gyrus is followed by a parallel somatosensory cortex in the postcentral gyrus which receives proprioceptive feedback from the muscular periphery, among other sensory information. Neuroimaging studies have shown that primary somatosensory cortex is engaged by both overt and covert speech production [15] and thus could be implicated in both the planning and execution of laryngeal muscle activity. Evidence from highly-trained singers has identified regions of somatosensory cortex proximal to the dorsal LMC whose local activity [34], resting-state connectivity [35], and structure [36], are associated with singing experience. One interpretation of this finding is that it reflects the heightened control and somatosensory/kinesthetic awareness of vocal musculature that are associated with extensive musical training in voice [37]. This interpretation is supported by a study which showed that magnetic stimulation of right somatosensory cortex improved pitch-matching in non-singers, but only when acoustic masking forced them to rely on somatosensory feedback [38]. Together, these studies suggest the possible existence of an area of somatosensory cortex that is associated with enhanced laryngeal control for singing. However, this finding has not been underpinned by direct measurements of laryngeal position or kinematics. Furthermore, the work described here was limited to the neural underpinnings of sung behaviours – it is not known if this neural substrate of expert vocal control for singing would extend to speech, for example as it applies in speakers' attempts to manipulate the physical body traits implied by the quality of their voices.

In order to address research gaps in knowledge about the neural and physiological basis of vocal modulation, the current study set out to measure vocal modulation behaviour in expert and non-expert vocalists, and to investigate the neural representations of the human larynx for speech in both populations. To do this, we conducted a vocal size and pitch imitation task with both highly-trained singers and non-singer control participants. Specifically, we created novel versions of the participants' own speech, in which we manipulated the fundamental frequency (F0) and the formant frequencies to simulate target voices with varying perceived pitch and vocal tract length (VTL; see Figure 1a). In order to mimic these voices, participants were required to adjust two dimensions of larynx motor behaviour - vocal fold tension, and larynx height in the vocal tract. To make the task challenging, we measured voice imitation across two different vowels (the front vowel /i:/ and the low back vowel /a:/), and for different combinations of acoustic F0 and VTL shifts. It was anticipated that the high front tongue position for the vowel /i:/ and the low back tongue position for /a:/would be differentially constraining for larynx raising and lowering, thus adding demand to the vocal control required to articulate the vowel accurately while imitating the voice targets. Further, as F0 and VTL typically covary negatively across human voices (i.e. adult males have longer VTLs and lower pitches than adult females and children), we predicted that including atypical combinations of voice parameters (e.g. short VTL with lowered pitch; long VTL with raised pitch) should also add difficulty to the task. Both design choices were thus made to maximise the discriminability of expert versus non-expert vocal control. In the imitation task, participants produced heard targets after a short delay (1-2 seconds), such that in an fMRI experiment we could model neural activation separately for speech preparation (i.e. hearing targets and planning speech) and speech execution – this allowed us to inspect representations of imitated voice parameters during different stages of speech production. During the fMRI runs, acoustic speech imitations were recorded to allow extraction of F0 as a measure of vocal fold tension, while task-related vertical movements of the larynx were measured during interleaved blocks of real-time anatomical MRI of the vocal tract (see Figure 1b for design). Using multivariate analyses of behaviour and neural activation (Representational Similarity Analysis, RSA; [39]) during speech preparation and execution, we aimed to measure imitation accuracy and locate the representation of pitch/vocal fold tension and VTL/larynx height during the two phases of speech imitation. We predicted that expertise in singing would generalize to greater speech imitation accuracy in the singers, and reveal more robust corresponding neural representations of laryngeal activity in this group.

Methods

Participants

A total of 57 adults (20 male; Mean age = 24.7 years, s.d. = 5.7, range = 19-43 yrs) with healthy hearing and no neurological illness (both self-reported) completed the study. Twenty-seven participants (10 male; Mean age = 27.5 years, s.d. = 6.4, range = 20-43 yrs) were highly trained singers, with the primary recruitment criterion that they should have studied / be studying voice performance as the principle instrument in their first university or music college degree. One singer participant did not meet this criterion, but reported extensive singing experience (32 years) and ongoing engagement with singing practice and performance. The remaining 30 participants (10 male; Mean age = 22.1 years, s.d. = 3.4, range = 19-35 yrs) formed a control group. This included one participant who had reported as a singer with 5 years of experience, but did not meet the degree criterion. All participants completed a questionnaire on their music and language experience, which showed that the singers had on average 16.3 years of experience and/or education in voice (range = 5-35 yrs) and all currently practised singing. Across the sample, participants reported some experience and/or education in voice and musical instruments (Singers: mean = 3.3 instruments, range = 2-6; Controls: mean = 0.8 instruments, range = 0-3) and in languages additional to English (Singers: mean = 1.7 additional languages, range = 0-6; Controls: mean = 1.4 additional languages, range = 0-4). Thus, the main distinction between the participant groups was in their probable level of singing expertise, and we did not control for overall levels of musical or linguistic experience. All participants gave informed consent, and the study was approved by the Ethics Committee at the Department of Psychology, Royal Holloway, University of London.

Seven participants (5 Controls, 2 Singers) were excluded from the fMRI analyses due to an error with slice positioning; the remaining 50 were used in the calculation of the RSA region-of-interest maps. A final sample of 49 participants[1], comprising 24 Singers (9 male; mean age = 28.1 years, s.d. = 6.5, range = 21-43 yrs) and 25 Controls (8 male; mean age = 22.1 years, s.d. = 3.7, range = 19-35 yrs), was used in the statistical analyses of behavioural pitch imitation and in all searchlight analyses of brain activation. Larynx height could not be tracked for two of these participants, due to MR signal dropout (1 Control) and pervasive errors with the automated labelling of larynx height (1 Singer). Thus, the reported group analyses involving VTL imitation behaviour include a group of 47 participants comprising 23 singers (9 male; Mean age = 28.2 years, s.d. = 6.7, range = 21-43 yrs) and 24 controls (8 male; Mean age = 22.1 years, s.d. = 3.7, range = 19-35 yrs). We note that while we achieved good matching of the male-to-female ratio across groups, it was not possible to recruit more males due to a lack of availability of volunteer participants – we therefore do not report analyses on the effects of participant sex.

Stimuli

All audio speech data collected during the behavioural session were recorded with a condenser microphone (Røde NT1-A; RØDE Microphones LLC, Silverwater, Australia) and digitized through a PreSonus AudioBox USB recording system (PreSonus Audio Electronics, Inc., Baton Rouge, LA). The experimental stimuli comprised 18 versions of the monosyllabic words "bead" and "bard", generated from recordings of the participant's voice.

Participants produced 5 instances of "bead" and "bard" in a short carrier phrase (e.g. "Say the word: BEAD"), following instructions to produce the words at a normal pitch and with a slightly longer than natural duration (this was in order to

---

[1] One Singer was excluded at this stage for head movement exceeding our criteria (i.e. 1 or more mid-run jumps of >3mm translation in any of x, y, z, and/or >3 degrees rotation in any of pitch, roll, yaw, occurring in more than 1 block of the fMRI experiment).

obtain a sufficiently long vowel steady state portion for imitation and acoustic/vocal tract analysis in the main experiment). The experimenter selected one representative token of each word, aiming for a duration of 0.6-0.8 seconds and good voice quality (e.g. without vocal fry, which introduces distortions in the synthesis of target stimuli). Tokens were inspected, excised and saved using Praat (www.fon.hum.uva.nl/praat/).

The two selected tokens (one "bead" and one "bard") were then transformed into acoustically manipulated targets using a modified version of a procedure developed by Chris Darwin at the University of Sussex (http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/VTchange) that allows adjustment of the F0 and speech spectrum as ratios of the original stimulus values. To make clear the distinction between this acoustic modulation and actual vocal tract length, we here refer to the manipulated stimulus parameter as "acoustic VTL" or "acVTL".

A central, "normal voice" version of each word was produced, in which the acoustic VTL (acVTL) was unchanged but the F0 was shifted 2 semitones upward from the original (to allow for the generation of lower-pitched targets that would not go beyond the speaker's natural range). In addition, there were 8 modified versions of "bead" and "bard", in which the acVTL and F0 were further adjusted relative to the "normal voice", either by shifting both the F0 and acVTL by 2 or 4 semitones in the same direction (i.e. +2 F0, +2 acVTL; -4 F0, -4 acVTL), or in opposite directions (i.e. +2 F0, -2 acVTL; -4 F0, +4 acVTL). This process yielded final voice targets with F0 ranging from 89% - 140% of the participant's original F0 in Hz. Assuming a linear relationship between formant frequencies and physical vocal tract length, the apparent VTLs of the voice targets ranged from 79% - 126% of the participant's actual vocal tract length in centimetres. Figure 1a depicts the 2 resulting "axes" of voice targets used in the experiment.

The "normal voice" and all 8 modified voices were used in a behavioural practice session (See Supplemental Materials for details), while the "normal voice" and the 4 most extreme modified voices were used in the MRI session. For use in the MRI scanner, stimuli were further filtered with earbud-specific parameters for use with Sensimetrics earbuds (S14; Sensimetrics Corp., Malden, MA), then parametrically equalized (filter CF: 3.5 kHz; 10 dB gain; Q factor = 2), and normalized (root-mean-square) with Adobe Audition (Adobe Inc., San Jose, CA) - these steps ensured that all stimuli were clearly distinguishable against continuous rtMRI acquisition noise.

Behavioural Practice Session

Training video

The participant viewed a short presentation (lasting approx. 4 minutes) in Microsoft PowerPoint (Microsoft Corporation, Albuquerque, NM), in which they were introduced to examples of modified stimuli of the type used in the experiment (presented over headphones) and instructed how to perform the imitation task. The presentation can be found in the supporting data for this paper (https://osf.io/6pqkt/). Additional description of the training can be found in the Supplementary Materials.

Imitation practice

Participants completed a short practice task in which they produced imitations of all 18 voice targets ((1 normal voice + 8 modulated targets) x 2 words). Stimulus presentation and data collection was performed using Matlab (The MathWorks Inc., Natick, MA) with the Psychtoolbox extension [40] – see the Supplementary Materials for further details of the stimulus presentation and recording. Each condition was presented in miniblocks of 5 trials (two miniblocks per condition) and the order of conditions was pseudorandomized. Participants were given the opportunity for a short break every 6 miniblocks. Analyses of these data will not be discussed here.

MRI Session

MRI Procedure

All stimuli were delivered through MR-compatible earbuds; speech was recorded with a fibre-optic microphone (FOMRI-III; OptoAcoustics Ltd., Or Yehuda, Israel). All stimuli were presented, and speech output recorded, digitized and saved, via the Psychophysics toolbox running in Matlab, with back projection for presentation of visual stimuli. For MRI acquisition parameters, please see the Supplementary Materials.

In the scanner, participants listened to and imitated the central ("normal") voice condition and the 4 most extreme voice transformations (i.e. the endpoints of the axes tested in the behavioural practice; Figure 1a) only. A pair of rtMRI runs (63s each) was presented before each of the 3 fMRI runs (~12mins each), and the session ended with a T1-weighted whole-brain structural scan. The total duration of the scans was around 1 hour (Figure 1b).

fMRI data were acquired using a rapid-sparse, event-related protocol, with auditory stimuli and speech production events timed to occur during short silent periods between acquisition of whole-brain volumes. Each listen-then-imitate trial occurred over 2 dynamic acquisitions (i.e. 2 periods of acquisition + delay). Participants listened to a particular voice target condition, and imitated it when cued after the next acquisition. This enabled us to separately capture BOLD activation reflecting speech preparation and the subsequent execution of the speech. Listen-only and rest trials occurred in a single dynamic acquisition (see Figure 1b). Similarly to our previous work [41-42], we distinguish speech preparation from passive listening using the event labels "listen pre-imitate" and "listen only", respectively. Three trial types were thus presented during fMRI: listen-then-imitate (comprising listen pre-imitate and imitation events), listen only, and rest. Results of listen-only trials are not discussed here.

The structure of fMRI trials is illustrated in Figure 1b. Four miniblocks of 35 trials (20 listen-then-imitate (2 per speech target), 10 listen only (1 per speech target), and 5 rest) were presented per fMRI run, for a total of 140 trials. Trial order was randomized separately within each miniblock. Each fMRI run lasted approximately 13 minutes. Before entering the scanner, participants completed a practice fMRI miniblock of 35 trials (no speech data were recorded during this practice).

rtMRI blocks comprised pairs of 63-second runs. Across a pair of runs, participants imitated all 10 voice targets. Each target condition (e.g. “normal bead”) was delivered in a miniblock of 4 consecutive trials, for a total of 20 trials per run. The order of miniblocks was randomized across the two runs. Each trial began with delivery of an audio stimulus and a visual prompt (“Listen”), followed after 1.2 sec by a prompt to imitate (“Repeat”) and a 1.5 sec gap in which the participant produced their imitation.

Data Processing

Acoustic data

All participant imitations from the fMRI runs were subjected to an acoustic analysis in Praat to extract trialwise mean fundamental frequency (F0) in Hz from the vocalic portion of each utterance. Stimuli were analyzed in batch per condition, with trial-by-trial visual inspection of the F0 and adjustment of the measurement parameters if necessary (see Supporting Information for exclusion criteria). We calculated the mean condition-wise F0 shifts separately for “bead” and “bard” by subtraction from the mean F0 for the “normal” voice condition, such that performance was expressed in terms of the shift of F0 in semitones relative to the central voice target in Figure 1a.

Vocal tract MRI data

Vocal tract MRI images were compiled into one AVI file per run pair. From each video, images were cropped to 68 x 68 pixels covering the whole vocal tract area. Larynx coordinates were identified and extracted frame-by-frame using a custom Matlab toolbox [43]; larynx y-coordinates (in pixels) were averaged across the steady-state portion of the vowel in each imitated word, then across all trials for that condition (see Supporting Information for exclusion criteria). Separately for “bead” and “bard”, the mean coordinate for each modulated condition was normalized relative to the mean of the “normal” voice tokens for that run, then averaged across the three runs. These values were used in the construction of vocal tract-derived dissimilarity matrices of larynx height for RSA analyses (see below). Figure 1c illustrates example frames from the output of the larynx-tracking analysis from one Singer.

Functional MRI analysis

Functional MRI images were preprocessed within Matlab using the SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/). Per subject, raw EPI images were realigned, coregistered to the anatomical image, normalized to MNI space (and resampled to 2mm isotropic), and smoothed with a Gaussian kernel of 8mm FWHM. Data were then analyzed in a first-level general linear model, in which listen-only, listen pre-imitate and imitate events were modelled as regressors — separately for each “bead” and “bard” target —and convolved with the canonical haemodynamic response function in SPM. Listen-only and listen pre-imitate events were modelled at the onset of the auditory stimulus. Imitate events were modelled as coincident with the appearance of cue to speak (a green cross; Figure 1b). Six motion parameters (describing translations and rotations about the x, y and z axes) were included as regressors of no interest. For each subject, T-contrasts were calculated for 1) All listen pre-imitate events > Rest (conditions collapsed), 2) All imitate events > Rest (conditions collapsed), 3-12) Each listen pre-imitate (speech preparation) condition > Rest (i.e. separate contrasts for each “bead” and “bard” target), and 13-22) Each imitate condition > Rest.

Statistical Analysis

Behavioural Data

Analysis of larynx displacement and Fo shifts

Behavioural data were analyzed using linear mixed effects models within the *lme4* [44] package in the *R* environment. Outcome variables were 1) mean vertical larynx displacement (pixels) and 2) mean F0 shift. Fixed factors were Group (Singers, Controls), VTL (long, short), Pitch (high, low) and Word (bead, bard). Participants were modelled as random intercepts. Significance of interactions and main effects was established via likelihood ratio tests, in which a model containing the effect of interest was contrasted with a reduced model lacking the effect. For both outcome measures, the full linear model including the effect of Word produced a singular fit, therefore this factor was removed. For F0 shifts, removing the main effect of Pitch generated a singular fit, so for this main effect we instead report the coefficient statistic and its associated significance, obtained using the *sjPlot* [45] package in the *R* environment.

Representational Similarity Analysis

In order to model performance on the behavioural task, we constructed two 10 x 10 representational dissimilarity matrices (RDMs) for each participant. Cells within these matrices described the absolute pairwise distances between the different “bead” and “bard” targets in 1) F0 (semitones) and 2) Larynx height (pixels). For each participant, these matrices were then compared with two ideal 10 x 10 model RDMs describing the underlying relationships between target stimuli in pitch (semitones) and VTL (semitones), using Spearman correlation tests within the CoSMoMVPA toolbox [46] implemented in

Matlab. Figures 2b and 3b depict the model matrices alongside the group averaged performance matrices for Singers and Controls.

Group analyses of these Spearman correlation scores were conducted in the *R* environment: Mann-Whitney tests within the *coin* package [47] were used to compare performance between the two groups, and one-sample Wilcoxon tests to compare performance against zero. Finally, Spearman correlation was used to test the significance of the relationship between performance and years of experience in voice, separately for pitch and VTL, in Singers only.

Functional MRI data

Representational Similarity Analysis

Representational similarity analysis (RSA) on functional neuroimaging data was carried out using the searchlight function within the CoSMoMVPA toolbox. Two candidate representational dissimilarity matrices (RDMs) – the ideal pitch model and the ideal VTL model - were used to searchlight neural activation separately for 1) the listen pre-imitate (speech preparation) phase and b) the imitate (speech execution) phase of speech imitation trials. The neural data were RDMs generated from smoothed T-maps of the single-subject contrasts of each condition > rest. To constrain the searchlight analyses to regions showing significant activation associated with speech preparation and speech execution, respectively, we used group masks of 1) All listen pre-imitate > Rest for the listen pre-imitate data and 2) All imitate > Rest for the imitate data. The group ROIs were generated using second-level one-sample T-tests on all participants, calculated in SPM. In order to ensure that each mask was of comparable volume, the listen pre-imitate (i.e. speech preparation) mask was created at a voxel height threshold of $p < 1 \times 10^{-7}$ FWE and a corrected cluster threshold of p < .05 FWE (yielding 18128 voxels), while the imitation (i.e. speech execution) mask had a more liberal voxel height threshold of p< .05 FWE and a corrected cluster threshold of p < .05 FWE (yielding 10897 voxels; see Figure 1d). The searchlight process involved extracting 10 x 10 RDMs describing the distances (as Spearman correlation coefficients) between activation (listen pre-imitate or imitate) in spherical searchlight volumes (radius: 4mm) centred around each voxel in the ROI. Spearman correlation tests were applied iteratively to compare these neural RDMs with the relevant candidate RDM (i.e. ideal pitch model or ideal VTL model) across the brain – the resulting correlation coefficients were Fisher z-transformed before being converted back to Pearson correlations for use in the group analyses. Each searchlight analysis thus generated a map of correlation coefficients per subject.

Group analyses of the searchlight maps were carried out using nonparametric permutation-based tests implemented in the SnPM toolbox (version 13.1.06; http://warwick.ac.uk/snpm). For within-group comparisons of coefficients with zero we used the "One Sample T test" module: this test was applied separately for each searchlight analysis on 1) Singers only, 2) Controls only, and 3) all participants. For comparisons of the searchlight maps between Singers and Controls, we used the "Two Sample T test". For an exploratory analysis of the effects of experience on representations of VTL in Singers only, we used the "Simple Regression" module. For all analyses, we applied 10,000 permutations and no variance smoothing.

Results

Imitation of vocal tract length

During imitation, Singers displaced their larynx on average by 1.6 pixels/4mm upward (SD=1.3 pixels /3.3 mm; Range = 0.9 pixels / 2.3 mm downward – 4.9 pixels / 12.3 mm upward) and 2.4 pixels/6mm downward (SD=2.1 pixels / 5.3 mm; Range = 7.7 pixels / 19.3 mm downward – 0.7 pixels / 1.8 mm upward) relative to the normal voice to imitate modulated targets with short and long VTLs, respectively. This compared to an average of 0.7 pixels/1.8mm upward (SD=1.2 pixels; Range = 0.8 pixels / 2 mm downward – 5.2 pixels / 13 mm upward) and 1.0 pixels /2.5mm downward (SD=1.3 pixels; Range = 3.4 pixels / 8.5 mm downward – 1.3 pixels / 3.3 mm upward) for Controls.

Analysis using linear mixed models identified a significant two-way interaction of Group x Length (χ^2[1]=42.36, p<.001), and main effects of Group (χ^2[1]=7.06, p = .008), Length (χ^2[1]=187.54, p < .001), and Pitch (χ^2[1]=18.63, p < .001). Figure 2a illustrates these results: Singers made more pronounced vertical displacements for both the long VTL and short VTL targets compared with controls, while both groups showed a lower vertical larynx position when imitating longer vocal tracts and lower-pitched targets.

Representational similarity analysis was used to compare vertical larynx movements in each participant with a model describing ideal performance on VTL imitation. This identified significant correlations with the model in Singers (median Spearman's rho: 0.569, z = 4.09, p < .001), and in Controls (although this relationship was weaker; median Spearman's rho: 0.149, z = 2.24, p = .012). A direct comparison of the two groups confirmed a significantly better fit to the model for Singers than Controls (z = 2.99, p = .003; see Figure 2b). However, a further Spearman correlation analysis revealed no significant relationship between Singers' RSA scores and the number of years of experience in voice.

RSA searchlight analyses of neural activation supported these findings (Figure 2c; Supplementary Table S1), with a stronger representation of the ideal VTL model for Singers (vs Controls) during speech preparation in right dorsal pre-/post-central gyrus (with the peak in somatosensory area S1). Taken alone, the Singers showed significant representation of VTL during speech preparation in an overlapping region of right central sulcus/post-central gyrus, and in additional volumes within the hippocampus and thalamus. However, there was no significant correlation between the strength of neural representations and the number of years of experience in voice. An analysis of all participants revealed significant representation of VTL in left ventral post-central gyrus during speech preparation. There was no evidence of significant representations in the Control group alone at the chosen threshold.

Imitation of pitch

During imitation, Singers shifted voice F0 on average by 3.7 semitones up (SD = 0.8; Range: 0 up – 5.2 up) and 2.5 semitones down (SD = 1.0; Range: 5.1 down – 0.1 down) relative to the normal voice for high and low-pitched targets. This compared to an average of 2.8 semitones up (SD = 1.1; Range: 0.1 up – 4.2 up) and 1.6 semitones down (SD = 1.3; Range: 5.3 down – 0.8 up) for controls (see Figure 3a).

Analysis using linear mixed models identified significant two-way interactions of Group x Pitch ($\chi^2[1]=69.13$, $p < .001$) and Pitch x Length ($\chi^2[1]=7.75$, $p = .005$), and a significant main effect of Pitch ($t = -43.24$, $p < .001$). The effects can be observed in Figure 3a: All participants distinguished between high and low pitched targets through shifts in the F0 of their imitations. Within this, Singers tended to make more pronounced upward and downward shifts in F0 than Controls, while both groups showed relatively smaller excursions in F0 for short VTL targets compared with long VTL targets.

Representational similarity analysis of the F0 of the spoken imitations showed that both the Singers and Controls performed well (Figure 3b), with median Spearman's correlation coefficients between each participant's performance and the ideal pitch model well above chance for both Singers (median Spearman's rho: 0.931, z=4.40, p < .001) and Controls (median Spearman's rho: 0.834, z=4.38, p < .001). When directly compared, there was a significant difference between the groups (z = 2.18, p = 0.03), indicating that trained Singers performed better than non-singing Controls at adjusting F0 upward and downward to match the voice targets. However, a Spearman correlation analysis revealed no significant relationship between singers' RSA scores and the number of years of experience in voice.

Despite the behavioural advantage for Singers, our searchlight analyses of neural activation data found no difference between groups in the neural representation of the ideal pitch model during speech preparation or speech execution. Further, we found no significant evidence for representation of the ideal pitch model during speech preparation or speech execution in either group separately, or in the combined participant group.

Discussion

We measured imitation of voice pitch and vocal tract length (VTL) in adult singers and non-singers, and probed the neural representations of laryngeal muscles during preparation and execution of imitations. Each participant imitated speech targets that were selectively manipulated relative to their normal voice. By using acoustic measures of F0 alongside larynx position metrics from vocal tract MRI images, we could directly and precisely measure the contributions of intrinsic (vocal fold) versus extrinsic larynx musculature to speech imitations. Furthermore, by comparing performance in trained singers and a group of non-singer control participants, we harnessed differences in vocal expertise to reveal the underlying neural representations of VTL for speech imitation.

We showed that both singers and controls can volitionally modulate vocal parameters in a goal-directed fashion to imitate voices of different sizes and pitches, in line with previous work investigating volitional vocal size exaggeration [16-17]. Specifically, we showed that both groups adjusted F0 downward and upward to imitate lower- and higher-pitched voice targets, respectively and, for the first time, we also showed that modulations to imitate longer and shorter VTLs were achieved via appropriate upward and downward movements of the larynx in the vocal tract. As predicted, singers showed larger modulations of both parameters, which in both cases were more closely correlated with an ideal model of imitation behaviour. Thus, we replicate previous findings that expertise in singing generalises to enhanced performance on speech tasks [23], here for two parameters of laryngeal sensorimotor control.

Using multivariate searchlight analysis of neural activation data, we identified representation of VTL in both cortical and subcortical sites during preparation to speak. A region of left somatomotor cortex identified in the whole participant group did not correspond topographically to previous reports of the larynx motor cortex (LMC). However, a further direct comparison of singers and controls revealed an expertise-related enhancement of VTL representation in right somatosensory cortex, just posterior to the reported location of the dorsal LMC in humans [16-18]. We speculate that this dorsal site could represent a larynx sensory cortex that is closely coupled to its corresponding LMC during speech motor control [9]: in line with this, probabilistic diffusion tractography analyses of LMC connectivity have revealed dramatically stronger connectivity with somatosensory and inferior parietal cortices in humans than in macaques [48]. However, we also note that although the precentral gyrus is predominately associated with motor-related activity and the postcentral gyrus with somatosensation, recent neuroimaging and neurostimulation data suggest that these functional divisions do not always align with gross anatomical landmarks [32,49]. Hence, we refrain from claiming the precise nature of the representations here as somatosensory.

To date, only one study has explicitly investigated the neural correlates of extrinsic laryngeal muscle activity, using univariate analysis of BOLD fMRI data. Belyk & Brown [31] scanned (non-expert) participants while they displaced the larynx in a downward direction, or in both downward and upward directions, and compared the spatial distribution of activation with that measured during phonation (i.e. vibration of the vocal folds). When participants were asked to move the larynx vertically, without speaking, the investigators observed extensive activation covering ventrolateral sensorimotor cortex in both hemispheres, which included the dorsal LMC. Our results extend this finding, as we show that the postcentral gyrus cortex houses representations of vocal tract length during speech that are associated with expertise-related group differences in voice modulation through larynx movement.

Previous work on imagined speech and song suggests that imagery can engage similar neural responses to overt execution of spoken and sung vocal behaviour [15,37]. Our findings of representation of VTL during preparation to speak echo those of our previous study on vowel imitation, in which we reported robust evidence for neural representation of articulatory information (using both vocal tract MR images and acoustic models of formant characteristics) prior to speech execution [41]. In that study, we showed that the raw acoustic properties of the target vowel stimuli were insufficient to account for our findings, suggesting that the identified regions thus contained information related to articulation rather than acoustics per se. In the current study, we demonstrate that representation of VTL during speech preparation was stronger in trained singers, who could more effectively imitate VTL through vertical displacement of the larynx. We argue that the regions implicated here may be critically involved in the conversion of auditory input to motor output [50], although we cannot rule out the contribution of actual larynx movement during this phase. Also in line with our previous study, we found no evidence for representation of VTL during activation related to speech execution. The current paradigm was sufficient to obtain robust univariate activation during imitation (see Figure 1d) - nevertheless, as we previously described [41], there may be specific considerations for probing the properties of overt speech behaviour that are not well suited to the current method of investigation. For example, due to the somatotopic arrangement of motor cortex, it may be that the overall activation of laryngeal motor regions during phonation is sufficiently high to obscure relational differences associated with F0 or larynx height. These may therefore may be better captured before speech onset.

Despite robust representation of the pitch model in the imitative behaviour of both singers and controls, we found no evidence of pitch representation in neural activation patterns. We deliberately constrained pitch targets to be within a comfortable range of ±4 semitones. In contrast, the 8-semitone range in VTL in the current study was quite extreme: changes in VTL sufficient to yield a percept of a change in talker identity are around half as large as for F0 (pitch), suggesting that talkers typically vary VTL much less than F0 in everyday speech [51]. Indeed, even when participants are asked to exaggerate body size volitionally during speech, they tend to make more substantial changes in F0 than VTL [16]. The extent of the F0 shifts chosen for our task, in terms of their perceptual salience and/or the physiological demand of imitating them, may therefore have been insufficient to detect pitch representations in the neural data, in comparison with the more exaggerated VTL targets. However, a recent study with choral singers explored responses to four levels of sung pitch spanning a much wider range (21 semitones), and found no evidence for representation (using a searchlight with a 4-way multivariate classifier; [52]). An alternative possibility is that the larynx's intrinsic musculature may be represented neurally in a more fine-grained way linked to ongoing prosodic modulation rather than mean pitch. This argument is supported by recent work using electrocorticography in pre-surgical patients, in which the intonation contour of spoken sentences and sung phrases was tracked by high gamma activity of electrodes located in dorsal LMC [21].

Several previous studies have explored the neural correlates of vocal expertise, revealing effects on regional activation and structure, as well as connectivity [34-36, 52-54]. In the current study, we found a significant difference between singers and non-singer controls in the spoken imitation of VTL, and in the neural representation of this vocal parameter. The neural locus of stronger VTL representations in singers has been previously linked to singing experience [34-36] and proposed as a correlate of enhanced larynx control and kinesthetic awareness in singing [37] - our MRI data on larynx position and neural representations corroborate this claim, and extend it to the imitation of speech. There is substantial overlap between the neural systems engaged during speech and song production [55], and the components of vocal imitation tested here – perceiving an auditory target, converting it to a motor plan, activating that plan, and monitoring and compensating for sensory feedback errors – are likely to share commonalities across these domains. But it remains unclear whether the expertise-related activations reported here indeed reflect singers' enhanced sensorimotor processing within a common vocal control system for speech and song, or if they arise because singers were using a singing strategy to perform our speech imitation task. Using a wider range of spoken and sung tasks in future work will help to delineate this further.

Our analyses suggested that performance on our vocal imitation task was not related to the number of years of singing experience. However, our sampling strategy was not appropriate to investigate effects of the frequency and recency of singing practice, which might have impacted this result [34]. We also did not control for broader musical experience across our sample of singers and controls. Thus, the observed group differences in our study could be the result of specific training in voice, general musical training (56; though see [23]), the level of ongoing singing practice [34], aspects of innate pre-disposition toward vocal/musical activities [57], or some combination of these. Investigation of a variety of expert groups (e.g. instrumentalists, voice artists) can resolve these factors to better understand the specific contributions of singing expertise to vocal imitation. Further, future studies with non-singing controls should explore the extent to which task-specific training on speech imitation (e.g. with real-time vocal tract feedback of larynx position) can enhance the performance of vocal imitation and its neural representation.

Conclusions

We have provided a novel representational account of laryngeal control in the human cerebral cortex by combining speech acoustics with MR imaging of the brain and vocal tract. We have demonstrated generalization of singing expertise to enhanced performance in a vocal size and pitch imitation task, and identified a possible common neural substrate in somatosensory cortex.

Acknowledgments

The study was funded by a Research Grant (ES/L01257X/1) from the Economic and Social Research Council. Carolyn McGettigan is currently funded by a Research Leadership Award (RL-2016-013) from The Leverhulme Trust.

Data Accessibility

Supporting data are available on the Open Science Framework (https://osf.io/6pqkt/).

Authors' Contributions

CM designed the study. SW, EK, NL, MB, CL, and CM collected and/or analysed data. DC, MM, and VC developed bespoke tools for data collection and analysis. All authors contributed to the writing and approval of the manuscript.

Competing Interests

We have no competing interests.

References

1. Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.
2. Fitch, W. T. (2018). The biology and evolution of speech: A comparative analysis. *Annual Review of Linguistics*, *4*, 255-279.
3. Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, *366*, 50–54.
4. Iwatsubo, T., Kuzuhara, S., & Kanemitsu, A. (1990). Corticofugal projections to the motor nuclei of the brainstem and spinal cord in humans. *Neurology*, *40*, 309–312.
5. Kuypers, H. G. J. M. (1958a). Corticobulbar connexions to the pons and lower brain-stem in man. *Brain*, *81*(3), 364–388. https://doi.org/10.1093/brain/81.3.364
6. Kuypers, H. G. J. M. (1958b). Some projections from the peri-central cortex to the pons and lower brain stem in monkey and chimpanzee. *The Journal of Comparative Neurology*, *110*, 221–255. https://doi.org/10.1002/cne.901100205
7. Simonyan, K., & Jürgens, U. (2003). Efferent subcortical projections of the laryngeal motorcortex in the rhesus monkey. *Brain Research*, *974*(1–2), 43–59.
8. Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, *37*, 529–604.
9. Belyk, M, & Brown, S. (2017). The origins of the vocal brain in humans. *Neuroscience & Biobehavioral Reviews*, *77*, 177–193.
10. Fischer, J., & Hammerschmidt, K. (2011). Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: Insights into the evolution of vocal communication. *Genes, Brain and Behavior*, *10*(1), 17–27.
11. Fitch, W. T. (2011). The evolution of syntax: An exaptationist perspective. *Frontiers in Evolutionary Neuroscience*, *3*, 1–12.
12. Fitch, W. T., Huber, L., & Bugnyar, T. (2010). Social Cognition and the Evolution of Language: Constructing Cognitive Phylogenies. *Neuron*, *65*(6), 795–814.
13. Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences*, *1016*, 749–777.
14. Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: A window into the origins of human vocal control? *Trends in Cognitive Sciences*, *20*, 304–318.
15. Simonyan, K., & Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist*, *17*(2), 197–208.
16. Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., & Feinberg, D. R. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Scientific reports*, *6*(1), 1-8.
17. Pisanski, K., & Reby, D. (2021). Efficacy in deceptive vocal exaggeration of human body size. *Nature communications*, *12*(1), 1-9.
18. Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PloS ONE*, *7*(2), e31353.
19. Cartei, V., Cowles, W., Banerjee, R., & Reby, D. (2014). Control of voice gender in pre-pubertal children. *British Journal of Developmental Psychology*, *32*(1), 100-106.
20. Cartei, V., Garnham, A., Oakhill, J., Banerjee, R., Roberts, L., & Reby, D. (2019). Children can control the expression of masculinity and femininity through the voice. *Royal Society open science*, *6*(7), 190656.
21. Guldner, S., Nees, F., & McGettigan, C. (2020). Vocomotor and social brain networks work together to express social traits in voices. *Cerebral Cortex*, *30*(11), 6004-6020.
22. Zarate, J. M. (2013). The neural control of singing. *Frontiers in human neuroscience*, *7*, 237.
23. Christiner, M., & Reiterer, S. M. (2015). A Mozart is not a Pavarotti: Singers outperform instrumentalists on foreign accent

imitation. *Frontiers in human neuroscience*, *9*, 482.

24. Lã, F. M., & Gill, B. P. (2015). Physiology and its impact on the performance of singing. In The Oxford Handbook of Singing (pp. 1-19). Oxford University Press, Oxford.
25. Brown, S., Ngan, E., & Liotti, M. (2008). A larynx area in the human motor cortex. *Cerebral Cortex*, *18*(4), 837–845.
26. Loucks, T. M. J., Poletto, C. J., Simonyan, K., Reynolds, C. L., & Ludlow, C. L. (2007). Human brain activation during phonation and exhalation: Common volitional control for two upper airway functions. *NeuroImage*, *36*(1), 131–143.
27. Simonyan, K., Saad, Z. S., Loucks, T. M. J., Poletto, C. J., & Ludlow, C. L. (2007). Functional neuroanatomy of human voluntary cough and sniff production. *NeuroImage*, *37*(2), 401–409.
28. Belyk, M, Pfordresher, P. Q., Liotti, M., & Brown, S. (2016). The neural basis of vocal pitch imitation in humans. *Journal of Cognitive Neuroscience*, *28*, 621–635.
29. Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, *495*(21), 327–332.
30. Dichter, B. K., Breshears, J. D., Leonard, M. K., & Chang, E. F. (2018). The control of vocal pitch in human laryngeal motor cortex. *Cell*, *174*(1), 21-31.e9.
31. Belyk, M., & Brown, S. (2014). Somatotopy of the extrinsic laryngeal muscles in the human sensorimotor cortex. *Behavioral Brain Research*, *270*, 364-371.
32. Breshears, J. D., Molinaro, A. M., & Chang, E. F. (2015). A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. *Journal of Neurosurgery*, 123(2), 340-349.
33. Simonyan, K., Ackermann, H., Chang, E. F., & Greenlee, J. D. (2016). New developments in understanding the complexity of human speech production. *Journal of Neuroscience*, *36*(45), 11440-11448.
34. Kleber, B., Veit, R., Birbaumer, N., Gruzelier, J., & Lotze, M. (2009). The brain of opera singers: experience-dependent changes in functional activation. *Cerebral Cortex*, *20*(5), 1144-1152.
35. Zamorano, A. M., Zatorre, R. J., Vuust, P., Friberg, A., Birbaumer, N., & Kleber, B. (2020). Singing training predicts increased insula connectivity with speech and respiratory sensorimotor areas at rest. bioRxiv, 793083. https://doi.org/10.1101/793083
36. Kleber, B., Veit, R., Moll, C. V., Gaser, C., Birbaumer, N., & Lotze, M. (2016). Voxel-based morphometry in opera singers: Increased gray-matter volume in right somatosensory and auditory cortices. *Neuroimage*, *133*, 477-483.
37. Kleber, B., & Zarate, J. M. (2014). The neuroscience of singing. In *The Oxford Handbook of Singing*. Oxford University Press Oxford, UK.
38. Finkel, S., Veit, R., Lotze, M., Friberg, A., Vuust, P., Soekadar, S., ... & Kleber, B. (2019). Intermittent theta burst stimulation over right somatosensory larynx cortex enhances vocal pitch-regulation in nonsingers. *Human brain mapping*, *40*(7), 2174-2187.
39. Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.
40. Brainard, D. H. (1997) The Psychophysics Toolbox. *Spatial Vision 10, 443-446.*
41. Carey, D., Miquel, M. E., Evans, B. G., Adank, P., & McGettigan, C. (2017). Vocal tract images reveal neural representations of sensorimotor transformation during speech imitation. *Cerebral Cortex*, *27*(5), 3064-3079.
42. Carey, D., Miquel, M. E., Evans, B. G., Adank, P., & McGettigan, C. (2017). Functional brain outcomes of L2 speech learning emerge during sensorimotor transformation. *NeuroImage*, *159*, 18-31.
43. Kim, J., Kumar, N., Lee, S., & Narayanan, S. (2014). Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *Proceedings of the International Seminar on Speech Production ISSP* (pp. 222-225).
44. Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).
45. Lüdecke, D. (2017). sjPlot: data visualization for statistics in social science. Retrieved from: https://CRAN.R-project.org/package=sjPlot.
46. Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in neuroinformatics*, *10*, 27.
47. Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a class of permutation tests: The coin package." *Journal of Statistical Software*, 28(8), 1–23.
48. Kumar, V., Croxson, P. L., & Simonyan, K. (2016). Structural organization of the laryngeal motor cortical network and its implication for evolution of speech production. *Journal of Neuroscience*, *36*(15), 4170-4181.
49. Carey, D., Krishnan, S., Callaghan, M. F., Sereno, M. I., & Dick, F. (2017). Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract. *Cerebral Cortex*, *27*(1), 265-278.
50. Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory–motor transformations for speech occur bilaterally. *Nature*, *507*(7490), 94.
51. Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. D. (2009). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. In *Tenth Annual Conference of the International Speech Communication Association*.
52. Belyk, M, Lee, Y. S., & Brown, S. (2018). How does human motor cortex regulate vocal pitch in singers? *Royal Society Open Science*, *5*, 172208.
53. Zarate, J. M., Wood, S., & Zatorre, R. J. (2010). Neural networks involved in voluntary and

involuntary vocal pitch regulation in experienced singers. *Neuropsychologia*, *48*(2), 607-618.

54. Krishnan, S., Lima, C. F., Evans, S., Chen, S., Guldner, S., Yeff, H., ... & Scott, S. K. (2018). Beatboxers and guitarists engage sensorimotor regions selectively when listening to the instruments they can play. *Cerebral Cortex*, *28*(11), 4063-4079.
55. Cohen, A. J., Levitin, D. J., & Kleber, B. A. (2020). Brain mechanisms underlying singing. In The Routledge Companion to Interdisciplinary Studies in Singing (pp. 79-96). Routledge.
56. Coumel et al. (2019) Second Language Accent Faking Ability Depends on Musical Abilities, Not on Working Memory. *Frontiers in Psychology,* 10
57. Golestani, N., Price, C. J., & Scott, S. K. (2011). Born with an ear for dialects? Structural plasticity in the expert phonetician brain. *Journal of Neuroscience*, *31*(11), 4213-4220.
58. Phillips, N. D. (2017). Yarrr! The pirate's guide to R. *APS Observer*, *30*(3).

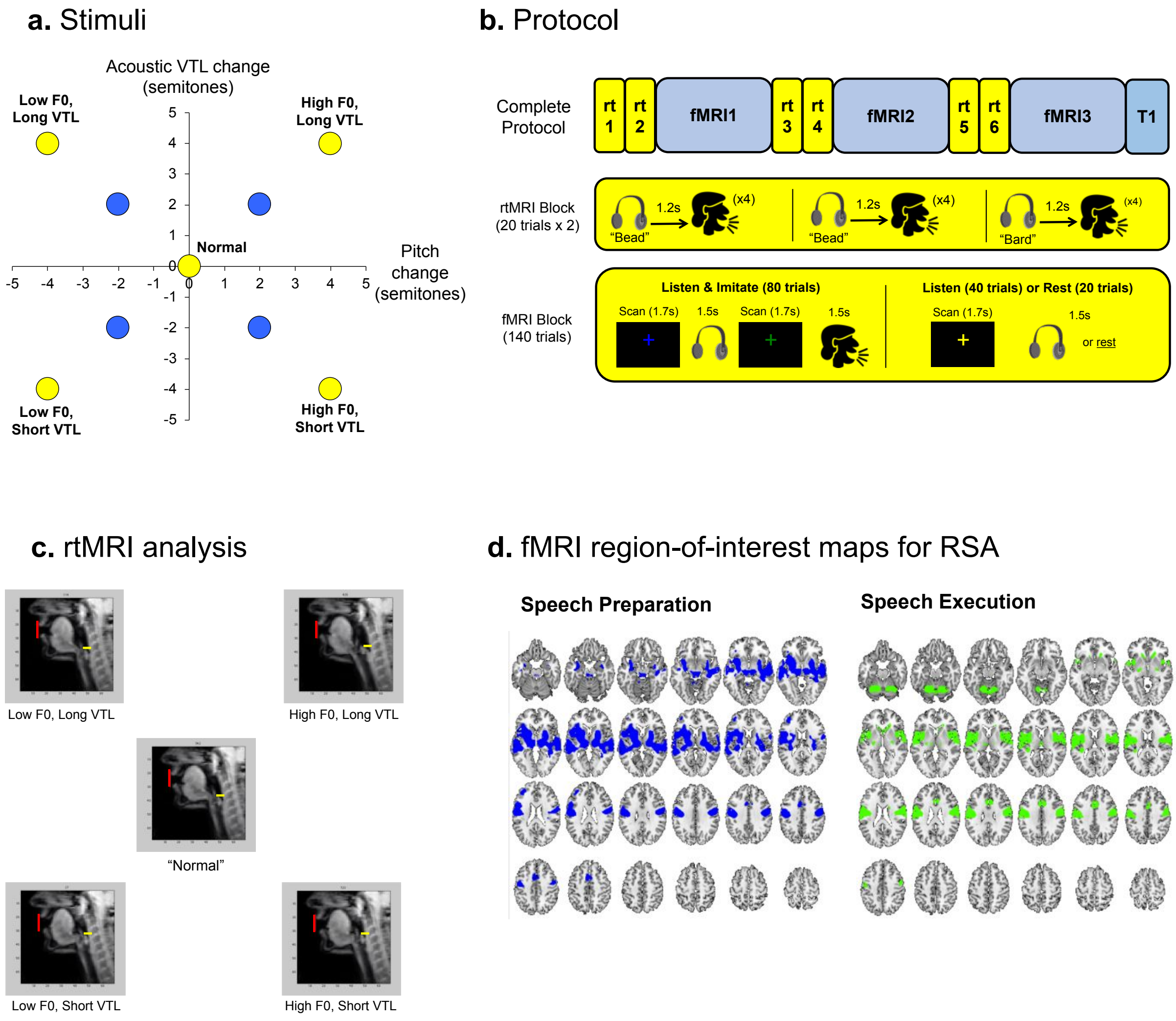


Figure 1: a) Schematic of the voice conditions used in the current study. Yellow dots indicate the acoustic conditions used in the MRI experiment; blue dots indicate additional voice targets imitated in the behavioural practice session. **b)** Experimental protocol. Top row: Overall ordering of scans, where "rt" stands for real-time anatomical MRI scans of the vocal tract, "fMRI" stands for functional MRI scans of brain activation, and T1 represents the whole-brain anatomical scan. Middle row: Details of the real-time MRI blocks. Participants heard a word over headphones and after 1.2 seconds were cued to provide a spoken imitation. Stimuli were presented in miniblocks of 4 trials per condition; condition order was randomized across the block pair. Bottom row: Details of the functional MRI trial types. A rapid-sparse routine was employed, in which listen and imitation events occurred during 1.5s pauses between EPI volume acquisitions. There were three trial types, cued through the colour of an onscreen fixation cross: 1) Listen & Imitate (blue → green for speech onset), 2) Listen only (yellow) and 3) Rest (white). The Listen & Imitate trials were used to calculate activation related to speech preparation and execution. **c)** Example real-time MR images of a singer performing imitation of the five voice target conditions for "bead". Each image shows a frame extracted from the steady state of the vocalic portion of the word, labelled according to the target's displacement from the "normal" voice in pitch and VTL. The yellow and red lines show the vertical position of the larynx and the horizontal position of the lips as obtained from a semi-automated image segmentation routine implemented in Matlab [43]. Only the larynx height data were analyzed for the current study. **d)** Axial whole-brain slices showing the group region-of-interest maps for Speech Preparation (calculated using a contrast of All listen pre-imitate > Rest including all participants; voxel height threshold $p < 1 \times 10^{-7}$ FWE, cluster threshold $p < .05$ FWE) and Speech Execution (calculated with a contrast of All imitate > Rest including all participants; voxel height threshold $p < .05$ FWE, cluster threshold $p < .05$ FWE). See Methods and Supplemental Materials for further details.

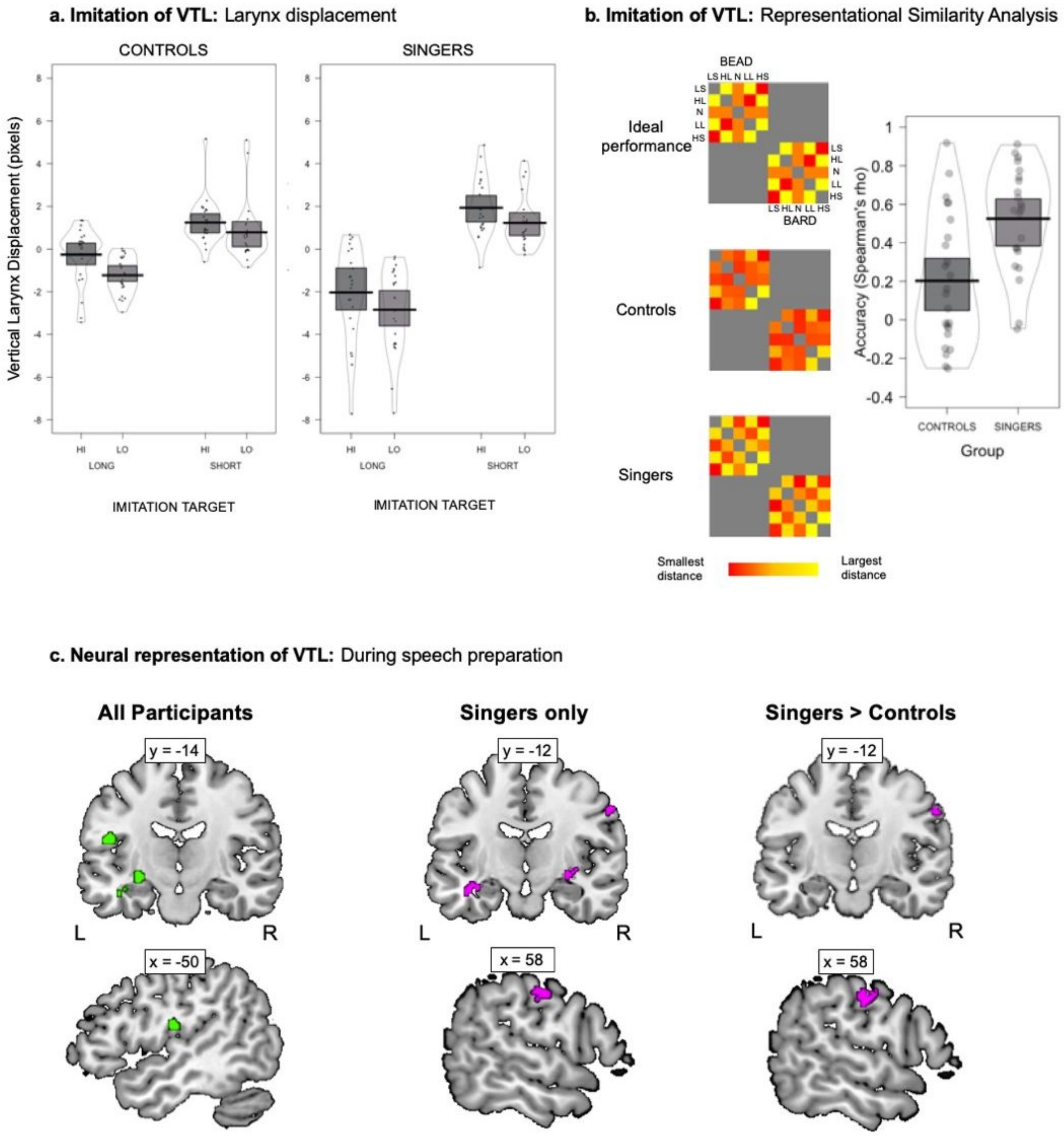


Figure 2: Imitation and neural representation of vocal tract length (VTL) a) Vertical larynx movement during the imitation of modulated speech targets. Data are plotted as the mean upward/downward larynx height excursion from the normal voice, where downward movements are negative. Results are shown collapsed across both word contexts (Bead, Bard). HI = Higher pitched targets, LO = Lower pitched targets; Long = longer VTL targets, Short = shorter VTL targets. Plots were created using the pirateplot function within the yarrr package in R [58]. Solid horizontal lines show the means per group and condition, boxes indicate 95% confidence intervals, dots indicate data from individual participants. **b)** Behavioural Representational Similarity Analysis (RSA) for the imitation of VTL. Measures of vertical larynx displacement in pixels (relative to the "normal voice") were used to generate representational dissimilarity matrices (RDMs) for each participant, which were compared with an ideal model (based on the inter-stimulus VTL distances in semitones) using Spearman's correlation tests. The figure shows the ideal model (LS = low F0, short VTL; HL = high F0, long VTL; N = Normal Voice; LL = low F0, long VTL; HS = high F0, short VTL) as well as the corresponding mean group RDMs for the singers and controls and a plot of accuracy by group (created using the pirateplot function within the yarrr package in R [58]). Solid horizontal lines show the means per group, boxes indicate 95% confidence intervals, dots indicate data from individual participants. **c)** Results of neural RSA searchlight analyses conducted within the CoSMoMVPA toolbox [46] implemented in Matlab. Areas of activation indicate regions showing a significant correlation between neural activation patterns during speech preparation and the ideal performance model for VTL imitation. Group images are shown at a voxel height threshold of $p < .001$ and a corrected cluster threshold of $p < .05$ FWE. Coordinates are in Montreal Neurological Institute stereotactic space.

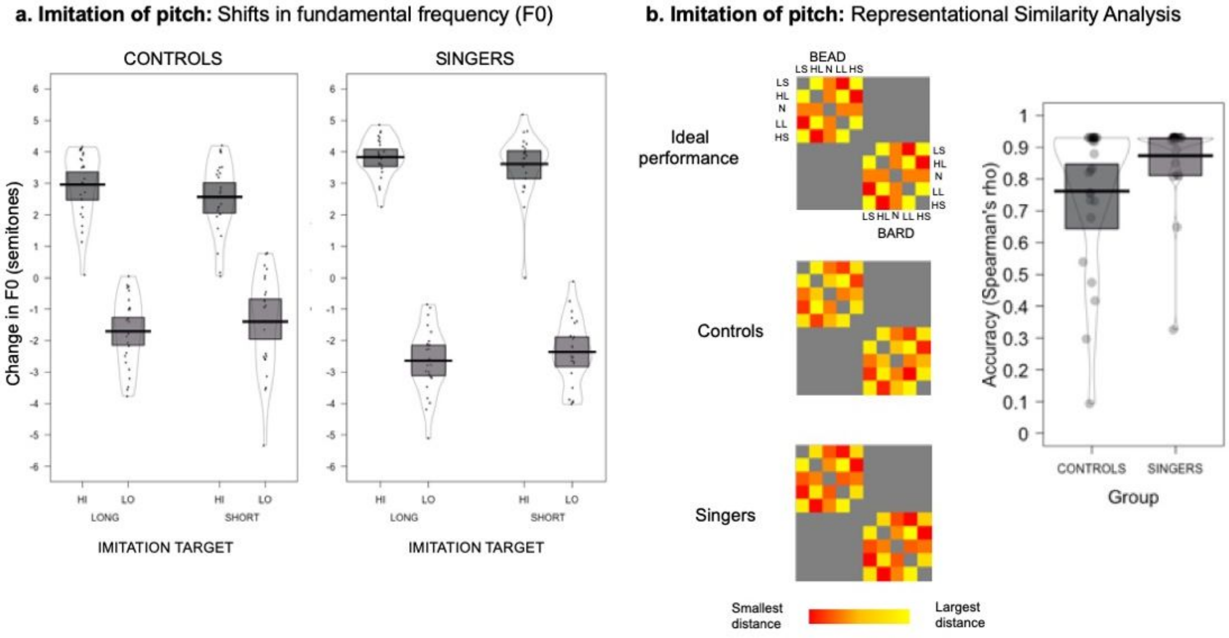
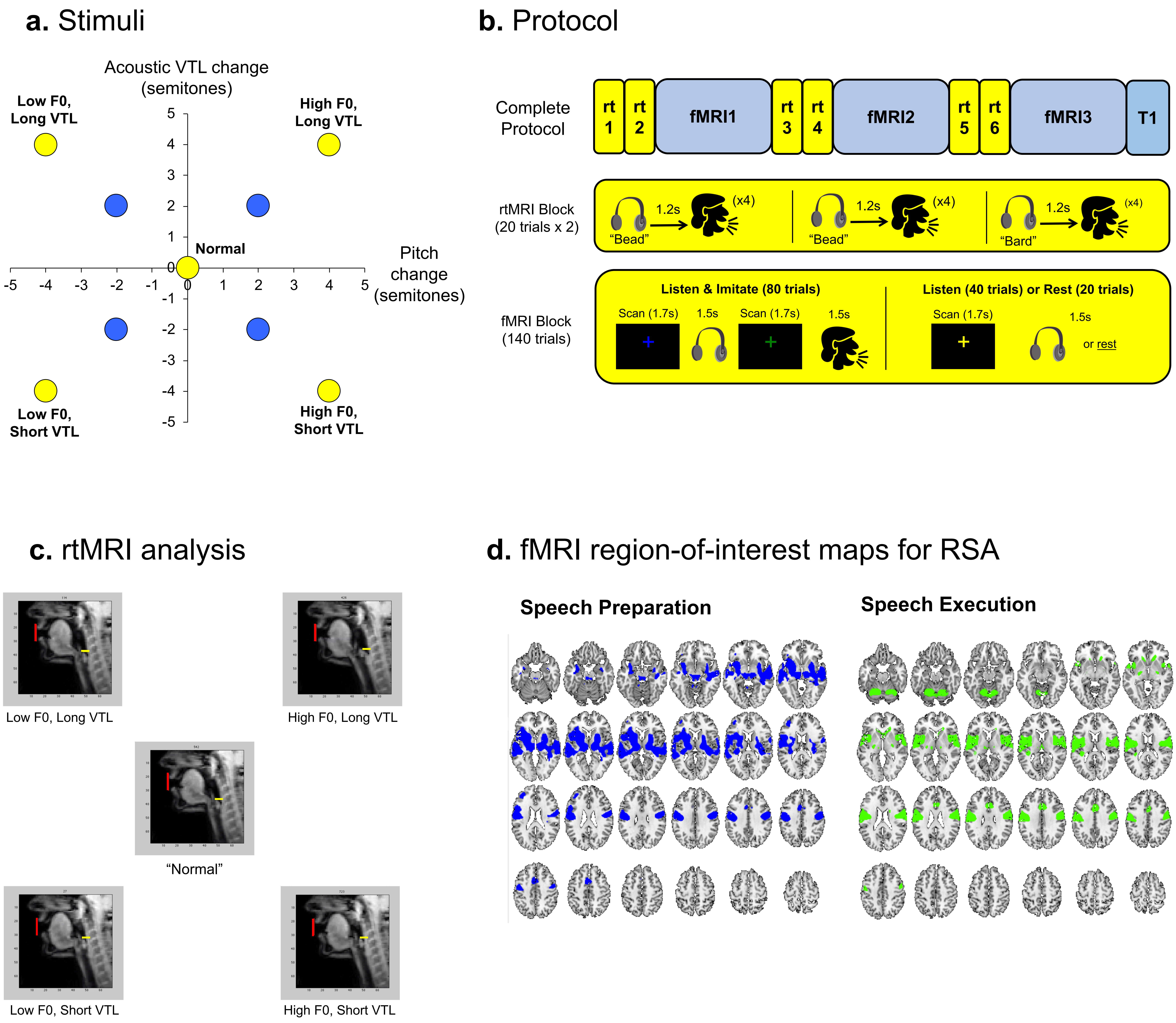
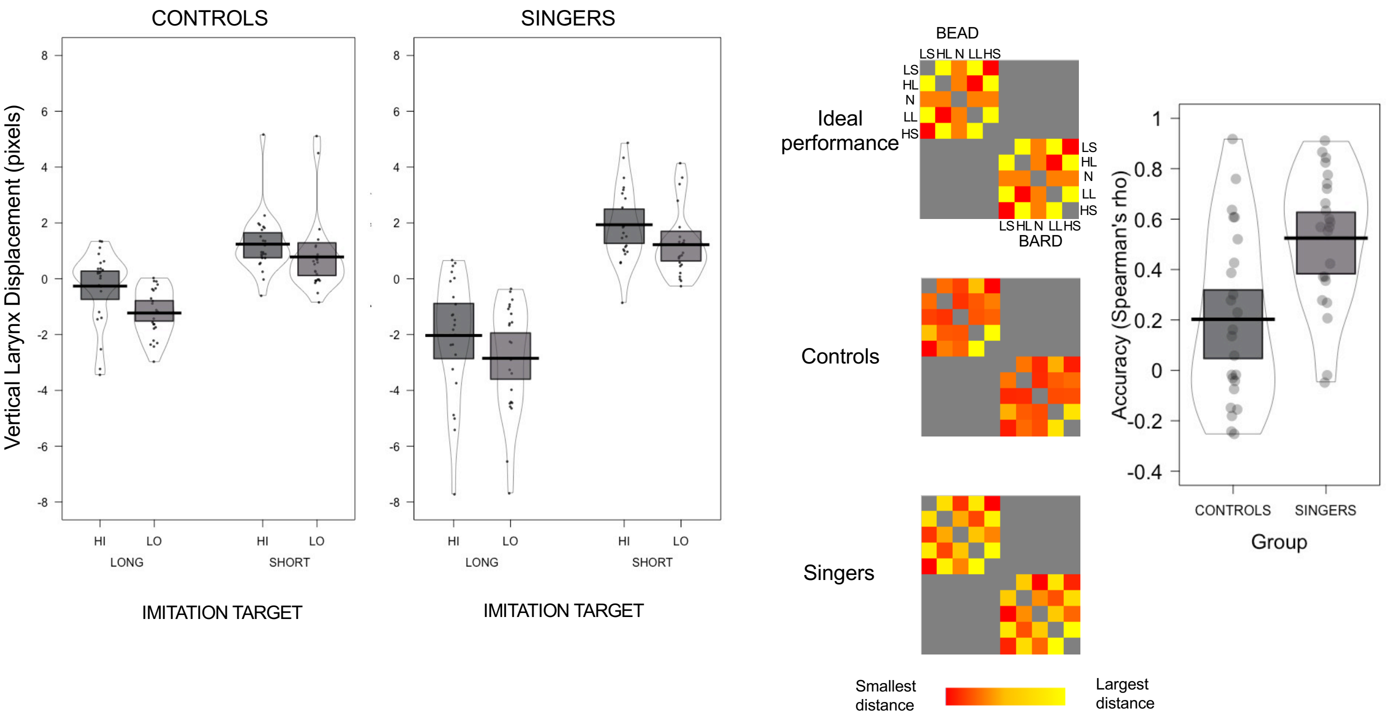


Figure 3: Imitation of pitch a) Changes in F0 during the imitation of modulated speech targets. Data are plotted as the mean upward/downward pitch excursion from the normal voice, where downward changes are negative. Results are shown collapsed across both word contexts (Bead, Bard). HI = Higher pitched targets, LO = Lower pitched targets; Long = longer VTL targets, Short = shorter VTL targets. RDI plots were created using the pirateplot function within the yarrr package in R [58]. Solid horizontal lines show the means per group and condition, boxes indicate 95% confidence intervals, dots indicate data from individual participants. **b)** Behavioural Representational Similarity Analysis (RSA) for the imitation of pitch. Measures of F0 change in semitones (relative to the “normal voice”) were used to generate representational dissimilarity matrices (RDMs) for each participant, which were compared with an ideal pitch model (based on the inter-stimulus F0 distances in semitones) using Spearman’s correlation tests. The figure shows the ideal model (LS = low F0, short VTL; HL = high F0, long VTL; N = Normal Voice; LL = low F0, long VTL; HS = high F0, short VTL) as well as the corresponding mean group RDMs for the singers and controls and a plot of accuracy by group (created using the pirateplot function within the yarrr package in R [58]. Solid horizontal lines show the means per group, boxes indicate 95% confidence intervals, dots indicate data from individual participants.

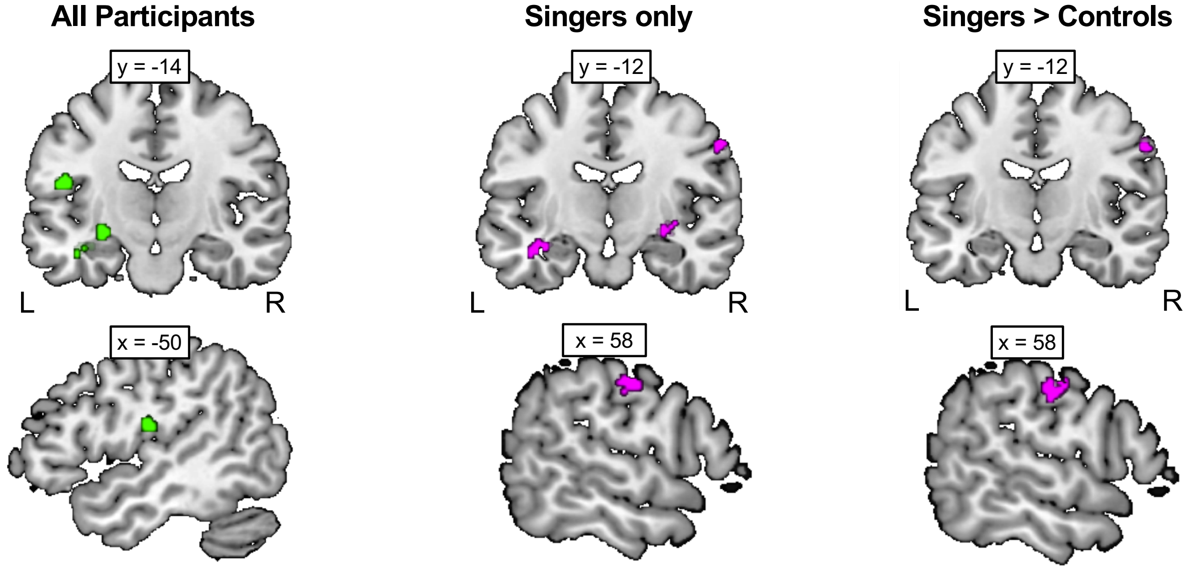

a. Stimuli
Acoustic VTL change (semitones)
Low F0, Long VTL
High F0, Long VTL
Normal
Low F0, Short VTL
High F0, Short VTL
Pitch change (semitones)
-5 -4 -3 -2 -1 0 1 2 3 4 5
b. Protocol
Complete Protocol
rt 1
rt 2
fMRI1
rt 3
rt 4
fMRI2
rt 5
rt 6
fMRI3
T1
rtMRI Block (20 trials x 2)
"Bead"
1.2s
(x4)
"Bead"
1.2s
(x4)
"Bard"
1.2s
(x4)
fMRI Block (140 trials)
Listen & Imitate (80 trials)
Scan (1.7s)
1.5s
Scan (1.7s)
1.5s
Listen (40 trials) or Rest (20 trials)
Scan (1.7s)
1.5s
or rest
c. rtMRI analysis
Low F0, Long VTL
High F0, Long VTL
"Normal"
Low F0, Short VTL
High F0, Short VTL
d. fMRI region-of-interest maps for RSA
Speech Preparation
Speech Execution

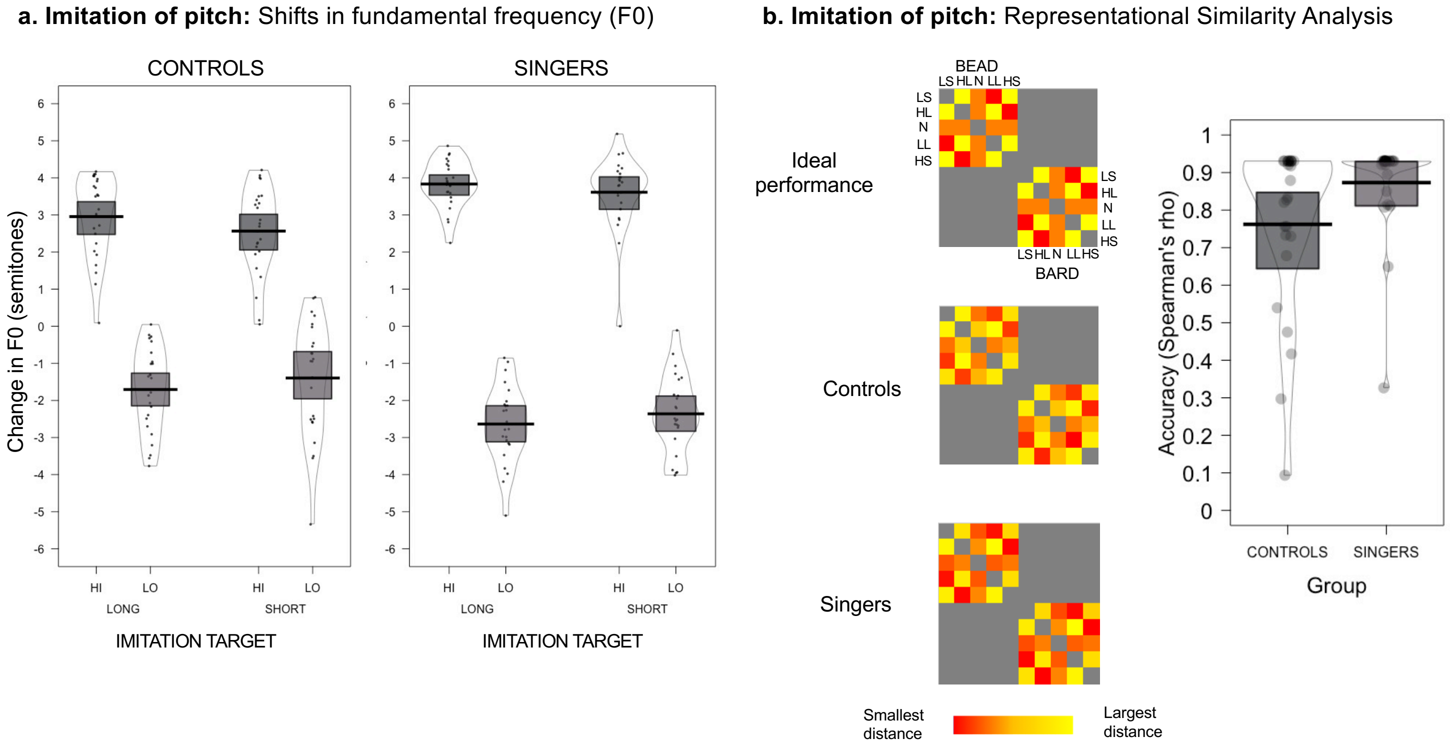


a. **Imitation of VTL:** Larynx displacement

b. **Imitation of VTL:** Representational Similarity Analysis

c. **Neural representation of VTL:** During speech preparation





a. Imitation of pitch: Shifts in fundamental frequency (F0)

b. Imitation of pitch: Representational Similarity Analysis