NeuroImage 79 (2013) 351–360

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Mean-based neural coding of voices

Attila Andics [a,b,c,*], James M. McQueen [a,d], Karl Magnus Petersson [a,b,e]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, P.O. Box 310, 6500 AH, The Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, P.O. Box 9104, 6500 HE, The Netherlands
[c] Comparative Ethological Research Group, Hungarian Academy of Sciences, Eötvös Loránd University, Budapest, Pázmány Péter sétány 1/c, 1117, Hungary
[d] Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, P.O. Box 9104, 6500 HE, The Netherlands
[e] Cognitive Neuroscience Research Group, Institute for Biotechnology & Bioengineering, CBME, University of Algarve, Faro, Campus de Gambelas, 8500-139, Portugal

ARTICLE INFO

Article history:
Accepted 4 May 2013
Available online 9 May 2013

Keywords:
fMRI
Inferior frontal cortex
Prototype-centered representations
Superior temporal sulcus
Voice identity learning

ABSTRACT

The social significance of recognizing the person who talks to us is obvious, but the neural mechanisms that mediate talker identification are unclear. Regions along the bilateral superior temporal sulcus (STS) and the inferior frontal cortex (IFC) of the human brain are selective for voices, and they are sensitive to rapid voice changes. Although it has been proposed that voice recognition is supported by prototype-centered voice representations, the involvement of these category-selective cortical regions in the neural coding of such "mean voices" has not previously been demonstrated. Using fMRI in combination with a voice identity learning paradigm, we show that voice-selective regions are involved in the mean-based coding of voice identities. Voice typicality is encoded on a supra-individual level in the right STS along a stimulus-dependent, identity-independent (i.e., voice-acoustic) dimension, and on an intra-individual level in the right IFC along a stimulus-independent, identity-dependent (i.e., voice identity) dimension. Voice recognition therefore entails at least two anatomically separable stages, each characterized by neural mechanisms that reference the central tendencies of voice categories.

Introduction

Human listeners can recognize individuals from their voices (i.e., auditory percepts of human vocalizations) alone and can rapidly learn new voice identities (i.e., voice-based percepts of person identity). Cortical regions involved in voice recognition have been mapped out, but it is not yet known how those regions represent voice knowledge. Here we test the hypothesis that in category-selective regions voice identities are represented in a prototype-centered voice processing hierarchy. In particular, we ask whether and how cortical activity reflects typicality in newly-learned voice categories. We will refer to this as mean-based neural coding of voices.

Two cortical regions have been reported to be sensitive to conspecifics' vocalizations. These regions are intriguingly similar in the primate and human brain and include regions along the superior temporal sulcus (STS) (in macaques: Petkov et al., 2008; in humans: Belin et al., 2000, 2011; Ethofer et al., 2009b; Grandjean et al., 2005) and the inferior frontal cortex (IFC) (in macaques: Romanski and Goldman-Rakic, 2002; Romanski et al., 2005; in humans: Fecteau et al., 2005; von Kriegstein and Giraud, 2006). Strong anatomical and functional connections have been found between the STS and the ipsilateral IFC in both primates (Hackett et al., 1998; Romanski et al., 1999) and humans (Ethofer et al., 2012). Furthermore, STS and IFC are not only voice-selective but also sensitive to short-term voice stimulus similarity, as demonstrated in rapid fMRI adaptation and carryover effects (STS: Andics et al., 2010, 2013; Belin and Zatorre, 2003; Latinus et al., 2011; Wong et al., 2004; IFC: Andics et al., 2010, 2013; Latinus et al., 2011). Short-term sensitivity here refers to mechanisms typically active within the range of a few seconds (cf., short-term repetition suppression, Epstein et al., 2008). This short-term sensitivity for voice similarity is an important requirement for the ability to tune in to voice stimuli, but it is not sufficient for the representation of long-term voice knowledge. Long-term here refers to processes relying on representations that need to be stored for longer than a few seconds (cf., long-term repetition suppression, Epstein et al., 2008). We adopt this definition in the present study. Neural storage of voice knowledge in the much longer term (e.g. weeks, months) is a topic for future research. Although it seems plausible that category-selective cortical regions are there to represent category knowledge for more than just in the short term, there is little evidence so far that the voice-selective STS and IFC contribute to representing voice knowledge for more than a few seconds.

This study asks whether the STS and IFC perform this function and elaborates on the recent proposal that long-term voice knowledge is represented in the human brain in a prototype-centered way. Mean-based neural coding appears to be a powerful way to represent individual stimuli in a category space (e.g., Panis et al., 2011). A possible mechanism for mean-based coding is neural sharpening (Hoffman and Logothetis, 2009): the coding of central values in relevant object dimensions becomes sparser with more experience. Neural sharpening reflects long-lasting cortical plasticity and so could be

* Corresponding author at: Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands. Fax: +31 24 3521213.
E-mail address: attila.andics@gmail.com (A. Andics).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.05.002

used for positioning stimuli in long-term object spaces. For faces, mean-based coding was found behaviorally (Leopold et al., 2001; Rhodes and Jeffery, 2006), in primates (Leopold et al., 2006), and also with human fMRI localizing the mechanism in face-selective fusiform regions (Loffler et al., 2005). It has been argued that mean-based coding can also result from long-term adaptation (Kahn and Aguirre, 2012), a mechanism that is sensitive to stimulus distributions. Recent behavioral (Bruckert et al., 2010; Latinus and Belin, 2011; Latinus et al., 2009; Mullennix et al., 2009; Papcun et al., 1989) and neuroimaging studies (Andics et al., 2010) also suggest that there is mean-based coding for voices. In other words, voice representations appear to be centered around prototypes in long-term memory.

Long-term mean-based coding for voices has nevertheless not yet been demonstrated in voice-selective cortical regions. Andics et al. (2010) found mean-based coding for voices in several regions, but some of these regions (the deep posterior STS and the orbital/insular cortex) are not voice-selective. Other regions (the amygdala and the anterior temporal pole) appear to be involved in the multimodal integration of person identity rather than in pure voice identity processing (Andics et al., 2010; Belin et al., 2011; Latinus et al., 2011). Although recent findings suggested IFC involvement in the representation of long-term stored objects (Latinus et al., 2009), to date there is thus no evidence for long-term mean-based voice encoding in the core category-selective cortical regions, namely the STS and the IFC.

It has been proposed that voice recognition involves not only mean-based voice encoding but also separate processing stages for voice-acoustic and voice identity analysis (Belin et al., 2004, 2011; Bestelmeyer et al., 2012; Charest et al., 2013; Scott and Johnsrude, 2003). This proposal, however, has received little direct support so far in the form of functional-anatomical correspondences between voice-processing stages and voice-selective regions. In the framework of mean-based coding, voice-acoustic analysis corresponds to an identity-independent, supra-individual representation of voice typicality, while voice identity analysis corresponds to an identity-dependent, intra-individual representation of voice typicality. These definitions will be adopted in the present study. Note that typicality is thus defined here with respect to the materials in the experiment, and not judgments of typicality collected, for example, in a rating study.

Recently, Latinus et al. (2011) attempted to dissociate acoustic from identity effects in voice processing, but their design focused on short-term effects of acoustic and identity changes. Short-term acoustic processing was found in both the STS and the IFC and short-term identity processing was found in the IFC only. These short-term effects may be indicators of long-term voice processing mechanisms, but those mechanisms have not yet been tested directly. The present study therefore tested the hypothesis that long-term mean-based voice encoding is present both at voice-acoustic (supra-individual) and at voice identity (intra-individual) levels of processing, and aimed to specify the role of the two core voice-selective cortical regions in these two levels.

We performed an fMRI experiment using a within-subject voice-training paradigm. Listeners were trained on two consecutive weeks to categorize voice stimuli on a voice morph continuum as belonging to either of two talkers characterized by the two continuum endpoints (morph0, morph100). During training the entire continuum was sampled and the acoustic center of the trained stimulus space was identical across weeks (morph50). The feedback during training on week1 and week2 specified different voice identity category boundary locations on each week (morph36 or morph64). After each training session, we could separately manipulate two perceptual properties of the voice stimuli: their perceived acoustic centrality (i.e., degree of prototypicality defined by the acoustic space, independent of identity feedback) and their perceived identity centrality (i.e., degree of prototypicality of a new voice identity, as defined by a voice-training procedure, independent of acoustic properties). Our design also allowed us to separately test for short-term effects (e.g., rapid adaptation indicating stimulus similarity sensitivity in the 0–5 second range) and long-term effects (e.g., long-term adaptation or neural sharpening indicating norm-based coding in the >5 second range) within a single experiment.

We hypothesized that cortical representations of the voice-acoustic space are organized along an acoustically central to acoustically peripheral dimension, and thus should not be modulated by voice identity feedback. Acoustically central stimuli should have sharper neural coding than acoustically peripheral stimuli and hence we predicted that there should be less activity for central than for peripheral stimuli in voice-acoustic regions. We also hypothesized that voice identity representations are organized along a feedback-defined typical to atypical dimension, and that this typicality is fully independent of voice-acoustic properties. We predicted that the activity of voice identity representations generated by identity-typical stimuli should therefore be less than the activity generated by atypical stimuli.

Material and methods

Participants

Eighteen Dutch female listeners (19–24 years) with no reported hearing disorders were paid to complete the experiment. Written informed consent was obtained from all participants. One person was excluded because of a failure to perform the task during training. Two further participants were excluded because of poor learning performance during training (i.e., voice identity categorization performance per morph level did not significantly differ from the 50% chance level in the final training block before scanning, one-sampled, two-tailed $t(14) < 1$, $p > .4$). The analyses presented here were based on the remaining 15 subjects.

Stimulus material

We selected two perceptually similar voices from a voice pool that contained recordings from young male nonsmoking adult native speakers of Dutch with no recognizable regional accents and no speech problems pronouncing Dutch monosyllables (Andics et al., 2007). The voices were unfamiliar to the listeners. Recordings were made in a soundproof booth using a Sennheizer Microphone ME62, a MultiMIX mixer panel, and Sony Sound Forge. All stimuli were digitized at a 16 bit/44.1 kHz sampling rate and were volume balanced using Praat software (Boersma and Weenink, 2007). A single token was selected per voice identity, of the word *mes* (knife). The two tokens were acoustically similar: average pitches were 122 Hz and 113 Hz, and stimulus lengths were 482 ms and 492 ms respectively.

We then created a voice morph continuum using the speech manipulating algorithms of STRAIGHT (Kawahara, 2006). The speech signals were decomposed into three parameters: an interference-free spectrogram, an aperiodicity map and a fundamental frequency (F0) trajectory. These parameters were then logarithmically interpolated segment by segment. Finally, a 100-step stimulus continuum with equidistant intermediate levels was resynthesized. The endpoints (levels morph0 and morph100) were also resynthesized. Average syllable duration was 487 ms (audio samples can be found at http://mpi.nl/people/andics-attila/research).

Training design

Listeners received multiple-phase voice identity training on two consecutive weeks. During the entire course of training, listeners were presented with words from the voice morph continuum and were instructed to make forced-choice decisions on talker identity after every word they heard. To allow initial assignment of talker names (Peter and Thomas) on response buttons to voice identities (voice A and voice B), listeners were presented three naturally

produced monosyllables from each talker before the experiment. The whole continuum was sampled each week. The assignment of talker names to voice identities and to dominant or non-dominant index fingers was counterbalanced across participants. The full stimulus range was sampled both during training and at test, but there was no exact stimulus overlap between the two parts (i.e., the morph levels used at training were different from those used at test; see below). Two training conditions were used: listeners were trained on different voice identity boundaries (morph36 or morph64) on the first and second weeks. The category boundary was made explicit by giving feedback according to a predefined boundary at 36% voice B morphs one week and at 64% the other week. Therefore, morphs between the two boundaries were trained to be categorized as voice A one week (when the boundary was at 64%), but as voice B the other week (when the boundary was at 36%). This training manipulation was amplified by presenting more stimuli from the most ambiguous parts of the continuum (Appendix A): The mean of all stimuli from each voice identity category was a 10% distance from the category boundary. The order of training conditions was counterbalanced across participants. Participants were not informed about the category boundary shift. The reason to use a continuum between two highly similar voices was to ensure that the boundary shift, while being large relative to the continuum, is not too large acoustically, and therefore remains unnoticed by the listeners.

Training procedure

Stimuli were presented via headphones binaurally, at a comfortable listening level. In each of two weeks participants received 72 min of training over 2 days, with 3 training sessions of 18 min each on day1 and a single training block of 18 min on day2. Training was followed by an fMRI test session on day2 in each week. Stimuli on consecutive trials were physically different. Stimulus ordering was otherwise random and varied across listeners. Training trials were 3000 ms long and included visual feedback (i.e., whether responses were correct, incorrect or late), presented from 2100 to 2400 ms after trial onset. Training phases contained 360 trials (12 repetitions of 30 morph levels). The manipulation appeared to be successful in that all participants reported, after the experiment, that they thought that they had heard various exemplars of natural voices only and that they were convinced that the trained voices were two actual persons' voices.

Conditions of interest

The critical stimuli in the fMRI test were morphs05, 33, 67 and 95. The categorization training defined identity membership of these stimuli (belonging to voice identity A or B), although these specific morph levels were not presented during training. Morph05 and morph33 always belonged to voice A, while morph67 and morph95 always belonged to voice B. The critical voice morphs also differed in terms of their distributional position on the stimulus continuum: Morph05 and morph95 were close to the endpoints, while morph33 and morph67 were close to the middle of the continuum — these morphs are referred to as peripheral and central stimuli, respectively. The trained voice identity and the centrality of these critical stimuli did not change across training sessions. But, crucially, the perceived typicality of the central voice morph stimuli changed as a function of the training condition. During voice identity boundary 36% training, morph67 was a typical exemplar of voice B (i.e., far from the identity boundary), and morph33 was an atypical exemplar of voice A (i.e., close to the identity boundary); but during voice identity boundary 64% training, morph33 was a typical exemplar of voice A, and morph67 was an atypical exemplar of voice B. These morphs, dependent on whether they were far from (>30 morph steps) or close to (=3 morph steps) the actual voice identity boundary, are referred to as typical and atypical stimuli, respectively. Note that acoustically peripheral stimuli were always far from the trained voice identity boundary, so they were always typical for one of the voices. Therefore, all critical stimuli fall into one of three types: peripheral–typical, central–typical or central–atypical. To control for the distance from the trained voice identity boundary across all typical stimuli when comparing these conditions, only those peripheral–typical stimuli were considered whose distance from the boundary matched central–typical stimuli's distance from the boundary (=31 morph steps). The conditions of main interest are summarized in Table 1.

fMRI test: design and procedure

Every listener was tested twice with fMRI. Stimuli consisted of pairs of tokens, each voice morphs of *mes*. The tokens used in the fMRI tests were morphs05, 33, 50, 67 and 95. There was an onset delay of 800 ms between tokens. Listeners were instructed to ignore the first voice and identify the second one (no feedback was given). FMRI tests were identical across the two weeks, but the pairs could fall into different condition categories on week1 and week2 depending on the identity boundary training. Each test session included 13 token pair types (Appendix B), with 20 repetitions of each type. A silent condition with 40 repetitions was also added. Token pair types were evenly distributed: each chunk of 15 consecutive trials included one of each token pair type and two silent trials. Consecutive trials were always physically different, and also different with respect to the corresponding experimental condition (Appendix B), but stimulus ordering was otherwise random.

Identical morph pairs were used to test for long-term adaptation (or neural sharpening) effects. We tested acoustically central and peripheral stimuli, and identity-typical and -atypical stimuli, all defined with respect to their positions in the constant acoustic space and the training-varied identity space (Table 1). Short-term adaptation effects were controlled in the tests of long-term effects because the pairs of morphs in each condition were always identical, and consecutive morph pairs were sufficiently distant (>5 s). Short-term effects of voice similarity were tested by comparing responses to identical versus non-identical morph pairs. We assumed that, in voice-selective cortical regions, identical pairs elicit reduced activity compared to non-identical pairs, due to rapid adaptation in response to stimulus repetition. Within non-identical pairs, we further differentiated between coarse and fine within-pair changes, determined by distance in morph steps.

Voice-selective regions were defined in a separate localizer run with blocks corresponding to (1) vocal sounds (verbal and nonverbal), (2) non-vocal sounds (animals, sounds from the environment, music) matched for number of sources, in duration, and overall energy and (3) silence. Participants were instructed to passively listen to the stimuli. Stimuli were controlled using Presentation software (www.neurobs.com). During imaging, stimulus presentation was synchronized by a trigger pulse with the data acquisition. Stimuli were delivered

Table 1
Characterization of conditions.

Condition	Critical morphs		Distance from acoustic center	Distance from identity boundary	Decision difficulty
	Boundary = morph36	Boundary = morph64			
Peripheral–typical	05	95	45 morph steps	31 morph steps	Easiest (96%)
Central–typical	67	33	17 morph steps	31 morph steps	Medium (88%)
Central–atypical	33	67	17 morphs steps	3 morph steps	Hardest (81%)

binaurally through MRI-compatible headphones (Commander XG, Resonance Technology Inc., Northridge, CA).

fMRI data acquisition

Measuring auditorily induced hemodynamic changes with fMRI remains a technical challenge: While continuous sampling methods suffer from scanner noise interference, sparse sampling methods have to cope with a decrease in signal-to-noise ratio caused by the disturbance of steady-state magnetization and subsequent loss of statistical power. We used a 3 T Siemens scanner and an in-house modified scanning protocol with scan-on periods for functional data acquisition and scan-off periods for stimulus presentation. For scan-off periods, gradient switching was removed to reduce scanner noise, but slice selective excitation pulses were played out to keep the magnetization in the steady state (see Schwarzbauer et al., 2006 for a similar protocol). Stimuli were always presented during scan-off periods. To further reduce scanner noise in all periods and to minimize period length at the same time, parallel imaging was used and no fat suppression was applied. A TR of 1200 ms was used. Trial onset-to-onset delay (i.e., the time between trials) was 8400 ms. Five functional volumes were acquired for each trial. For the main tests EPI-BOLD fMRI time series were obtained from 24 transverse slices covering temporal lobes and the inferior part of the frontal lobes with a spatial resolution of 3.5 × 3.5 × 3.5 mm, including a 0.5 mm slice gap (TE = 30 ms, ascending slice order; 300 trials; GRAPPA 2; sequence = SCAN–SCAN–SCAN–SCAN–SCAN–SILENT–SILENT; slice nr = 24; jittering: stimulus1 starts 200–800 ms after silent pulse onset). In total, each test session included 300 trials. The test was conducted as a single run lasting 45 min, including 4 half-minute breaks after each 8.4 min.

For the voice localizer there were 39 transverse slices and a longer silent gap between acquisitions (TR = 2000 ms; sequence = SCAN–SILENT–SILENT–SILENT–SILENT). Stimulus blocks of 8 s, corresponding to vocal sounds, non-vocal sounds and silence were presented after each volume. In total there were 20 blocks of each type (62 volumes including one dummy scan at the beginning and one extra scan at the end). All other parameters were identical to the main test settings. In addition to the functional time series, a standard T1-weighted three-dimensional scan using a turbo-field echo (TFE) sequence with 180 slices covering the whole brain was collected for anatomical reference at the end of the second scanning session, with 1 × 1 × 1 mm spatial resolution.

fMRI data analysis

Image preprocessing and statistical analysis were performed using SPM5 (www.fil.ion.ucl.ac.uk/spm). Phantom image files were added before normal preprocessing to fill missing volume gaps (created by scan-offs). These phantom images were removed again after design specification but before model estimation by editing the design matrices. The functional EPI-BOLD images were realigned, slice-time corrected, spatially normalized, and transformed into a common anatomical space, as defined by the SPM Montreal Neurological Institute (MNI) T1 template. Next, the functional EPI-BOLD images were spatially filtered by convolving the functional images with an isotropic 3D Gaussian kernel (10 mm FWHM). The fMRI data were then statistically analyzed using a general linear model and statistical parametric mapping (Friston et al., 2007). Every token pair was modeled as a separate event, using constant epochs corresponding to the average token length, starting from the onset of the second token. To account for differences in response times (RT), we also performed an a-posteriori confirmatory analysis modeling each event (i.e. token pair) with an epoch length equal to the RT specific to that trial, using the variable epoch approach as described by Grinband et al. (2008). As in the main analysis, the onset of each epoch was positioned at the onset of the second token (also corresponding to response time onset). For the main and confirmatory analyses, condition regressors were constructed per token pair type (Appendix B).

Regressors for silent trials and, to model potential movement artifacts, realignment regressors for each run were also included. A high-pass filter with a cycle-cutoff of 128 s was implemented in the design to remove low-frequency signals. Single-subject fixed effect analyses were followed by random effects analyses on the group level. The whole-volume functional localizer run's statistical test was first thresholded at $p < .001$ uncorrected at the voxel-level (in order to define the supra-threshold clusters) and then family-wise-error (FWE) corrected at the cluster level ($p < .05$). The main run's statistical tests were small-volume corrected using the three significant clusters of the functional localizer as regional masks, and therefore FWE-corrected at the voxel level ($p < .05$).

Results

Flexibility in voice identity learning

During training, overall identification accuracy was 69%. As a large proportion of the training stimuli came from the most ambiguous parts of the morph continuum (Appendix A), this overall hit rate does not directly reflect poor performance, but rather a very demanding training regime. Performance was much better for morph levels corresponding to unambiguous parts of the continuum than for those corresponding to ambiguous parts of the continuum (see Fig. 1). Identification accuracy clearly improved across blocks, especially in week1 for the unambiguous parts of the continuum (from 76% in block 1 to 92% in block 4). The increased accuracy level that was reached in the last block of week1 persisted over a one week delay and was found for all blocks in week2, despite the fact that a new boundary had to be learned (see Fig. 1).

A repeated-measures ANOVA of behavioral responses at the final block of training and during fMRI with the factors boundary (i.e., whether the trained voice identity category boundary was at morph36 or at morph64) and level (training: morph1, 34, 46/54, 66, 99 — matched to those used at test; test: morph5, 33, 50, 67, 95) confirmed that the boundary manipulation led to a training-related shift in voice identity judgments for ambiguous levels of the voice morph continuum (Fig. 2; training: boundary $F(1,14) = 855$, $p < .001$, level $F(4,56) = 730$, $p < .001$, boundary × level $F(4,56) = 146$, $p < .001$, linear component of the interaction $F < 1$, quadratic component of the interaction $F(1,14) = 738$, $p < .001$; test: boundary $F(1,14) = 19.3$, $p = .001$, level $F(4,56) = 330$, $p < .001$, boundary × level $F(4,56) = 2.61$, $p = .089$, linear component of the interaction $F < 1$, quadratic component of the interaction $F(1,14) = 10.4$, $p = .006$). Note that the training-related shift in voice identity judgments for the ambiguous morph levels, as reflected in the quadratic component of the boundary by

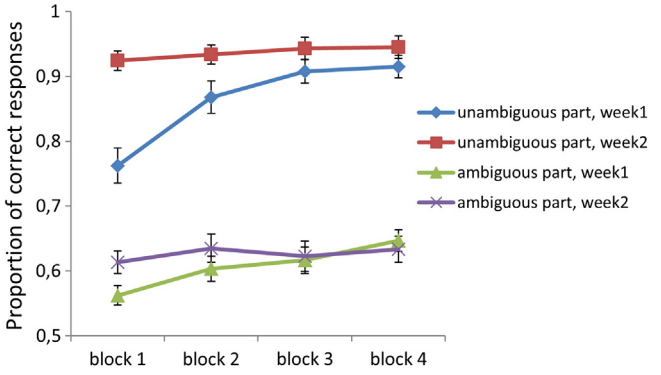


Fig. 1. Voice identification accuracy per week and per training block, for ambiguous and unambiguous parts of the continuum. 'Ambiguous part' and 'unambiguous part' correspond to morph levels that are less than 10 morph steps from the actually trained voice identity boundary or more than 10 steps, respectively. Error bars represent standard error of the mean.

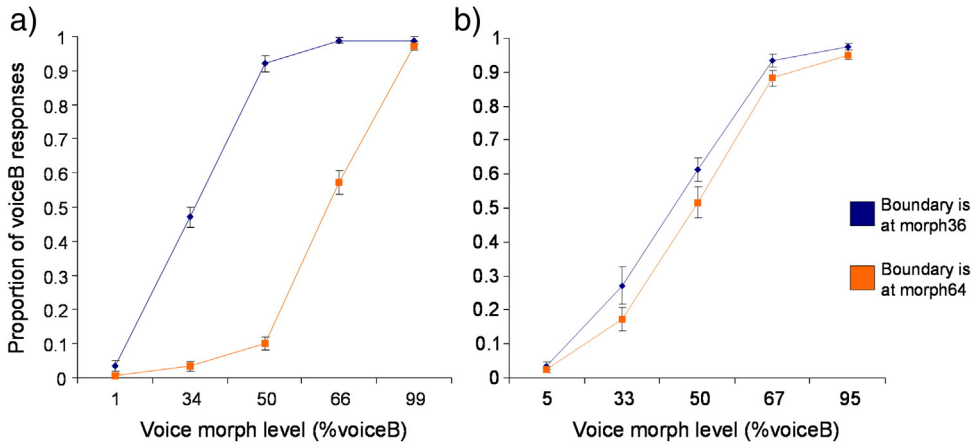


Fig. 2. Voice categorization per voice identity boundary training condition during training and at test. (a) Training: categorization performance in the final training block of each training session, data for morph levels matched to those used at test (e.g., morph50 refers to the average of two trained morph levels neighboring morph50). (b) Test: categorization during scanning sessions, data for morph pairs with no change. Error bars represent standard error of the mean.

level interaction, was significant not only at the end of training but also during the fMRI test. This is evidence that the within-participant, across-week boundary manipulation indeed worked. The reason why the boundary shift is smaller in the scanner than at the end of training is likely to be because of the more demanding experimental settings during fMRI: longer inter-trial intervals; voice pairs were presented but responses had to be made on the second stimuli only; scanner noise; and no feedback.

The proportion of correct decisions was used to judge decision difficulty per condition. We found that at test peripheral–typical trials were easier than central–typical trials (mean difference (%correct) = 7.66, $t(14) = 4.78$, $p < .001$); and central–typical trials, in turn, were easier than central–atypical trials (mean difference (%correct) = 7.54, $t(14) = 3.07$, $p = .008$; Table 1). These differences in decision difficulty were also reflected in RTs during fMRI. Responses for peripheral–typical trials were faster than those for central–typical trials (mean difference (RT) = 107 ms, $t(14) = 4.10$, $p < .001$); and responses for central–typical trials, in turn, were faster than those for central–atypical trials (mean difference (RT) = 38 ms, $t(14) = 2.55$, $p = .023$).

Voice-selective regions

Voice-selective regions were defined in a separate localizer run (Belin et al., 2000), contrasting vocal and non-vocal sounds (see Methods). Four regions survived an uncorrected $p < .001$ threshold ($t(14) > 3.79$): the bilateral STS and the bilateral IFC, but the left IFC region did not reach a cluster-level family-wise error (FWE) corrected level of significance (Table 2). These findings confirmed that the voice-selective regions include both superior temporal and inferior frontal regions.

Table 2
Voice sensitive regions as determined by the functional localizer.

Voice > non-voice	Size (voxels)	p (cluster-corr)	t(14)	x	y	z
Right STS	2647	<0.001	11.87	48	−32	4
			10.81	60	0	−8
			9.16	56	−20	−2
Left STS	2350	<0.001	8.96	−60	−16	4
			8.57	−44	10	−24
			8.32	−58	−44	16
Right IFC	467	0.002	6.24	56	18	24
			5.14	42	14	32
			4.94	48	6	34
Left IFC	30	0.785	4.98	−52	32	6

Height threshold was $p < 0.001$ ($t(14) = 3.79$). For each cluster, the table displays at most 3 local maxima more than 8.0 mm apart.

Mean-based coding of acoustic properties

The effect of "distance from acoustic center" (i.e., distance from morph50) was investigated by contrasting acoustically peripheral and acoustically central stimuli. We predicted that, in regions that code acoustic centrality, peripheral stimuli would elicit greater activity than central stimuli, independently of how typical those stimuli are in the feedback-driven identity space (i.e., peripheral–typical > central–typical = central–atypical; Table 1). We found that only a single voice-selective cluster in the right STS was sensitive to stimulus position in the acoustic space set by the experiment (Table 3). In this region response reduction was found for acoustically central compared to peripheral voice stimuli. As this contrast controlled for short-term adaptation effects (by presenting no-change morph pairs in each of the contrasted conditions), we propose that the response reduction found in the STS was caused by long-term adaptation or a neural sharpening mechanism acting on a long-term stored representation in the voice-acoustic space organized around the acoustic center. This finding of mean voice representations in the right STS is similar to proposed mean face representations in the fusiform face region (Loffler et al., 2005). We suggest, however, that the mean voice we describe here is derived from the experiment, and not from a life-time of experience

Table 3
Significant BOLD effects in the main analysis.

Contrast	ROI	p	t(14)	x	y	z
Long-term acoustic centrality						
Peripheral–typical > central–atypical	Right STS	0.003	6.63	64	−26	0
Peripheral–typical > central–typical	Right STS	0.008	5.79	66	−34	4
Long-term identity centrality						
Central–atypical > central–typical	Right IFC	0.021	4.16	44	16	30
Central–atypical > peripheral–typical	Right IFC	0.022	4.07	48	8	36
Short-term similarity						
Coarse change > no change	–					
Coarse change to central > no change, central	Right STS	0.050	4.47	66	−36	2
	Left STS	0.022	4.98	−64	−20	0
Coarse change to peripheral > no change, peripheral	–					
Fine change between identities > no change (matched)	–					
Fine-change within identity > no change (matched)	–					

ROIs were defined using the voice localizer run's voice vs nonvoice contrast, thresholded at $p < .001$ (uncorrected). Contrasts were thresholded at $p < .001$ ($t(14) = 3.79$). The table displays FWE-corrected p values where significant. No significant effects were found with these contrasts for other ROIs, nor with any further contrasts (e.g., with the reversed tests) for any of these ROIs.

(as in Loffler et al.), thus demonstrating the adaptive nature of the mean-based code.

The long-term stored representation of the voice-acoustic space was further investigated to see whether activity in the space was modulated by voice identity training. We found no evidence suggesting that this was the case, that is, there was no stronger response in the right STS or anywhere else to morph33 for the test sessions where listeners were trained on morph64 as the identity category boundary (i.e., to central–typical stimuli) compared to the test sessions where listeners were trained on morph36 (i.e., to central–atypical stimuli). This suggests that the acoustic space representation was independent of voice identity feedback.

A confirmatory analysis that modeled trial-specific RTs using a variable epoch approach (Grinband et al., 2008; see Methods) yielded very similar results for the same contrasts (Table 4), but note that in one of these tests coding of acoustic centrality in the voice-selective STS was found bilaterally. The similarity of results across analyses nevertheless suggests that the STS findings cannot be explained by across-condition differences in voice identity decision difficulty, as reflected in the RTs.

Mean-based coding of voice identity

The effect of "distance from identity boundary" (i.e., distance from morph36 or morph64) was tested by contrasting identity-atypical and typical stimuli. We predicted that in regions that code identity centrality, identity-atypical would elicit greater activity than identity-typical stimuli, independently of how central or peripheral those stimuli are in the acoustic space (i.e., central–atypical > central–typical = peripheral–typical; Table 1). We found that only a single voice-sensitive cluster in the right IFC was modulated by voice identity training (Table 3). In this IFC region response reduction was found for the same voice stimuli when trained as more prototypical versus less prototypical encounters of a talker. This contrast controlled for both short-term adaptation effects (by presenting no-change morph pairs in each of the contrasted conditions) and for acoustic variation (by contrasting conditions with exactly the same stimuli, but after differing voice identity training). We therefore propose that the response reduction found in the IFC was caused by a neural sharpening mechanism acting on long-term stored, prototype-centered representations in a voice identity space.

Table 4
Significant BOLD effects in the confirmatory analysis accounting for RTs.

Contrast	ROI	p	t(14)	x	y	z
Long-term acoustic centrality						
Peripheral–typical > central–atypical	Right STS	0.035	4.65	50	−28	6
Peripheral–typical > central–typical	Right STS	0.029	4.73	54	−26	4
	Left STS	0.015	5.11	−58	−10	8
Long-term identity centrality						
Central–atypical > central–typical	Right IFC	0.004	5.15	50	8	38
Central–atypical > peripheral–typical	Right IFC	0.032	3.67[+]	46	4	34
Short-term similarity						
Coarse change > no change	Right STS	0.059	4.38	66	−22	8
	Left STS	0.016	5.24	−62	−24	16
Coarse change to central > no change, central	–					
Coarse change to peripheral > no change, peripheral	–					
Fine change between identities > no change (matched)	–					
Fine-change within identity > no change (matched)	–					

ROIs were defined using the voice localizer run's voice vs nonvoice contrast, thresholded at $p < .001$ (uncorrected). Contrasts were thresholded at $p < .001$ ($t(14) = 3.79$). The table displays FWE-corrected p values where significant. No significant effects were found with these contrasts for other ROIs, nor with any further contrasts (e.g., with the reversed tests) for any of these ROIs.

[+] Thresholded at $p < .002$ ($t(14) = 3.44$).

Importantly, this response reduction was found for acoustically distant identity-typical voice stimuli that were associated with different person identities. A repeated-measures ANOVA on percent signal change values in the peak coordinate of the central–atypical vs central–typical test in the right IFC [44, 16, 30] was also performed with the factors voice identity (A, B) and identity centrality (identity-typical, identity-atypical). Beyond an obvious main effect of identity centrality ($F(1,14) = 16.95$, $p = .001$), we found no main effect of voice identity ($F < 1$) and no interaction of the two factors ($F < 1$). These data confirm that the identity centrality effect in IFC is equally present for each of the two voice identities we tested. This suggests that IFC maintains multiple prototype-centered voice identity spaces — perhaps one for each voice identity.

Further analyses confirmed that the IFC findings are not caused by across-condition differences in decision difficulty. First, no IFC modulation was found for an analog contrast with a similar difference in decision difficulty (Table 1) but without a difference in the distance from the trained category boundary (namely, for the central–typical > peripheral–typical contrast). Second, a confirmatory analysis that accounted for RT differences on a trial-by-trial basis yielded the same pattern of results (Table 4).

Rapid adaptation for voice changes in the STS

Further tests included non-identical morph pairs with coarse or fine voice changes that, through comparison to identical morph pairs, were used for investigating short-term adaptation effects. We demonstrated short-term adaptation for voice stimuli in voice-sensitive regions of the STS. Response reduction was found bilaterally in the STS for identical voice stimulus pairs compared to voice pairs with a coarse voice change, but no adaptation effect was found with a finer voice change. The loss of adaptation effect with finer voice changes was not modulated by voice identity properties (i.e., we found no adaptation in voice-selective regions for either fine between-identity changes or for fine within-identity changes). This pattern of activity indicates short-term coarse acoustic processing in the voice-selective STS. Interestingly, however, the adaptation effect with coarse voice changes was only present when no-change stimuli were acoustically central, and disappeared when no-change stimuli were acoustically peripheral. That is, short-term adaptation was modulated by long-term acoustic centrality in the voice-sensitive STS (Table 3). Note that the RT-modulated follow-up analysis confirmed the presence of the adaptation effect with coarse voice changes, but note that it was modulated by acoustic centrality (Table 4).

Discussion

We aimed at specifying the role of voice-selective cortical regions in maintaining long-term voice knowledge. Earlier studies have indicated that voices may be represented in prototype-centered voice spaces (Andics et al., 2010; Bruckert et al., 2010; Latinus and Belin, 2011; Latinus et al., 2009; Mullennix et al., 2009; Papcun et al., 1989) and that the STS (Andics et al., 2010; Belin and Zatorre, 2003; Latinus et al., 2011; Wong et al., 2004) and IFC (Andics et al., 2010; Latinus et al., 2011) are core voice processing regions, showing voice selectivity and short-term sensitivity to voice similarity. But these voice-selective regions of the STS and the IFC have not previously been shown to be involved in long-term mean-based voice coding, and indeed there has to date been no other evidence of long-term neural coding of voice prototypes. Here we performed an auditory fMRI study combined with a training manipulation. Listeners were trained on the same voice morph continuum but with different voice identity category feedback on two consecutive weeks, each time followed by scanning. After each training session, we could separately manipulate two perceptual properties of the voice stimuli: their perceived acoustic centrality (independent of identity feedback) and their perceived identity centrality (independent of acoustic

properties). The main results are: (1) there is long-term encoding of acoustic centrality of voices in the right STS, and (2) there is long-term encoding of identity centrality in the right IFC (Figs. 3a, b). We also confirmed that the bilateral STS is sensitive to short-term acoustic similarity of voices.

The present study therefore not only supports a hierarchical model of voice recognition, that is, that there exist distinct voice processing functions with distinct anatomical locations (Belin et al., 2004), but, critically, it also characterizes the neural mechanisms of these processing stages: our results provide evidence that both long-term acoustic and identity processing mechanisms are based on mean-based neural coding, and that these long-term codes are maintained in voice-selective regions of the STS and the IFC.

With respect to the role of the STS, previous work has established that regions of the bilateral (but right-lateralized) STS are voice-selective and play a key role in voice recognition (Andics et al., 2010; Belin et al., 2000; Formisano et al., 2008; Gervais et al., 2004; Latinus et al., 2011; von Kriegstein and Giraud, 2004; Warren et al., 2006) and talker normalization in word recognition (Wong et al., 2004). Even though there is agreement that the STS is a functionally highly heterogeneous region (Beauchamp et al., 2004), with distinct subregions having different properties and functions, even within the domain of voice processing (von Kriegstein and Giraud, 2004), its exact role in the hierarchical model of voice recognition is still debated. Crucially, there are differing views on whether the voice-selective right STS is also involved in identity processing of voices (Warren et al., 2006), or whether it is involved in acoustic processing exclusively (Andics et al., 2010; Latinus et al., 2011). In other words, does STS keep track of who is speaking or does it only encode how the voice sounds in relation to other voices? Andics et al. (2010) found that listeners' individual sensitivity to voice similarities in a right mid STS region correlated with pre-scan voice recognition performance, but they suggested that this measure reflected sensitivity to short-term acoustic similarity rather than long-term identity similarity. The present results show that the STS is involved in both short-term acoustic processing and in long-term acoustic processing (with a clear right-hemisphere dominance), but not in long-term identity processing.

We also tested for short-term identity sensitivity, but found no significant regions. Previous studies claiming to have found short-term identity processing in the STS have possible acoustic confounds. Warren et al. (2006) found that regions along the bilateral STS responded more strongly to change than to no change of speaker. They argued that the STS is therefore crucial for voice identity processing. However, this contrast had possible acoustic biases, since the changing speaker condition necessarily contained greater acoustic variation than the fixed speaker condition. So these findings may be evidence of short-term acoustic processing. The mid STS certainly appears to be a crucial stage of the voice recognition pathway, but we suggest that it does not encode person identity (i.e., intra-individual voice typicality) information. Based on the present findings we can make the case that the voice-selective right mid STS encodes acoustic centrality by maintaining a supra-individual, feedback-independent, norm-based acoustical voice space.

It should be noted that neural sharpening and long-term adaptation (long-term in the sense that its time-scale is longer than a few seconds, but still within the time-scale of the experiment and not over months or

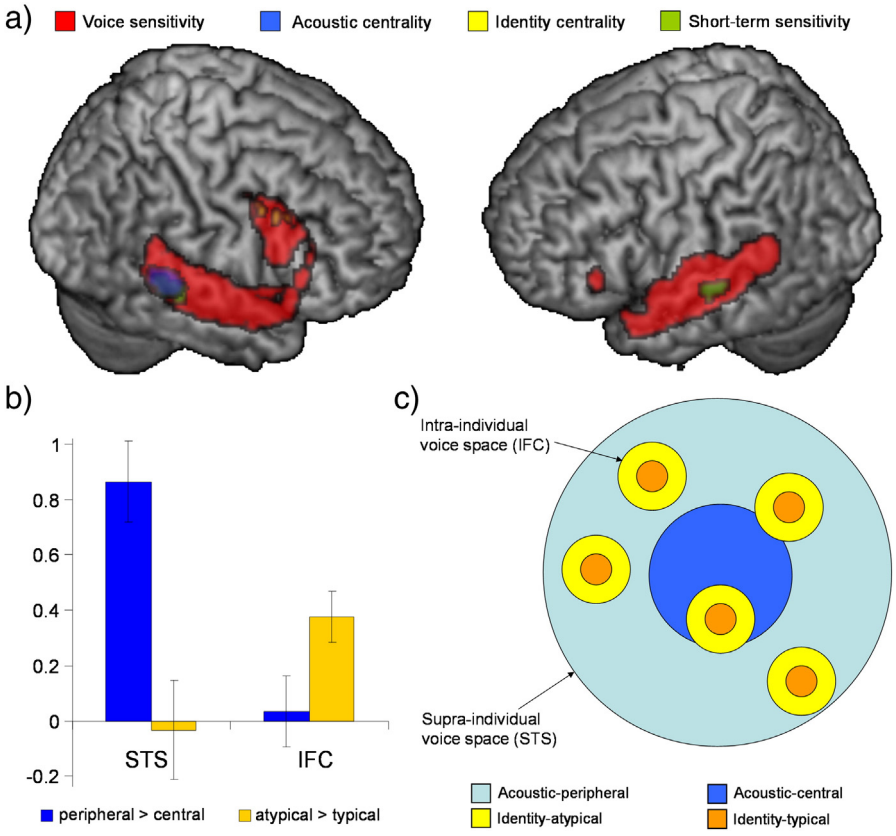


Fig. 3. Acoustic centrality and identity centrality representations of voices. (a) Contrast maps overlaid on a rendered brain, displaying voice sensitivity: voice vs nonvoice localizer (red), acoustic centrality: peripheral–typical vs central–typical (blue), identity centrality: central–atypical vs central–typical (yellow) and short-term sensitivity: coarse change (to central) vs no change (central) (green) contrasts. (All tests are thresholded at $p < .001$, $t(14) = 3.79$; and masked by the voice localizer, thresholded at $p < .001$, $t(14) = 3.79$.) (b) Bar graph displaying percent signal change in the peak coordinate of the acoustic centrality test (peripheral–typical vs central–typical) in the right mid STS [66, −34, 4] and in the peak coordinate of the identity centrality test (central–atypical vs central–typical) in the right IFC [44, 16, 30]. Error bars represent standard error of the mean. (c) A schematic illustration of mean-based representations of acoustic and identity properties in intra-individual and supra-individual voice spaces.

years) are both possible (and plausible) neural mechanisms behind these long-term acoustic centrality effects in the right STS. The present experiment was not designed to distinguish between sharpening and adaptation accounts. Instead, our study distinguishes between short-term and long-term centrality effects, and between acoustic and identity effects. Nevertheless, we argue that our findings demonstrate that STS responses reference the central tendencies of the voice stimulus distribution, as derived from the experiment, and not from a life-time of experience. This is the phenomenon we refer to as 'mean-based coding', and it is independent of the underlying mechanism. Note that the concept of the adaptive mean-based codes for voices described here is in contrast with the proposed life-long mean-based codes for faces in Loffler et al. (2005).

With respect to the role of the rIFC, the importance of prefrontal regions in the processing of voices has been demonstrated only recently, in extracellular recording experiments with primates (Romanski and Goldman-Rakic, 2002; Romanski et al., 2005). These studies showed that neurons in the macaque ventrolateral prefrontal cortex respond stronger to conspecifics' vocalizations than to nonvocal auditory stimuli. An analog region with a similar response pattern was identified in the human brain (Fecteau et al., 2005), responding more strongly to speech and to nonlinguistic vocalizations than to non-voice stimuli, and to emotional than to neutral vocalizations. Other studies have also suggested that the IFC is involved in voice processing (Andics et al., 2010; Bestelmeyer et al., 2012; Charest et al., 2013; Ethofer et al., 2009a; Latinus et al., 2011; Stevens, 2004; von Kriegstein and Giraud, 2004, 2006), that IFC responses to voices are enhanced after learning more about the voices (von Kriegstein and Giraud, 2006), and that the IFC is sensitive to short-term voice-acoustic (Andics et al., 2010; Latinus et al., 2011) and voice identity changes (Latinus et al., 2011). The present study provides the first demonstration that individual voice identities are represented in a prototype-referenced manner in the human prefrontal cortex. A single region in the right IFC responded more strongly to identity-atypical than to identity-typical stimuli when all acoustic properties of the stimuli were controlled. Our results thus suggest that the right IFC contributes to long-term voice knowledge. More specifically it appears to encode voice identity centrality (i.e., how far a given voice stimulus is from an average of the listener's memory of that specific person's voice). Recent findings in voice gender and voice attractiveness processing come to similar conclusions. Charest et al. (2013) proposed that the IFC reflects stimulus ambiguity and long-term voice gender representations. Bestelmeyer et al. (2012) demonstrated that less attractive voices elicit greater IFC activity, independently of acoustic properties. These studies and the present findings converge on the claim that the voice-sensitive IFC is involved in linking voice representations to basic, long-term social concepts such as person identity, person gender and person attractiveness.

Recently, Latinus et al. (2011) made an attempt to dissociate acoustic from identity effects in voice processing, using a training paradigm with voice morph continua, but despite these similarities there are major design differences between it and the present study. First, the study by Latinus and colleagues focused on short-term sensitivity effects but was not designed to capture long-term effects. Stimulus relations were systematically manipulated within morph pairs, but there were no long-interval comparisons across the different types of pairs. Their contrasts, however, were not free of long-term acoustic effects. In the present study, however, the multi-level manipulation of conditions (i.e., both within and across morph pairs) allowed us to identify effects of short-term and long-term similarity sensitivity simultaneously. Second, the acoustic and identity contrasts in the Latinus et al. study were not fully independent. In the present study, in contrast, the within-subject, multi-session training paradigm allowed us to test for identity effects with acoustic variation fully controlled. In spite of these design differences, our results can easily be reconciled with those of Latinus et al. (2011). In our view, the results of both studies converge in suggesting that the STS is involved in short-term acoustic similarity processing. Latinus et al.'s findings also indicate that the IFC is involved in short-term processing of either acoustic or identity similarities of voices and in Andics et al. (2010) it was found to be involved in short-term acoustic processing. In the present study, however, the IFC was not found to be involved in short-term identity processing. We therefore suggest that to date there is no convincing evidence for the involvement of the IFC, and, in fact, of any other cortical regions, in short-term identity processing. Instead, IFC appears to support short-term acoustic processing and, critically, long-term voice identity processing.

Andics et al. (2010) found that several other cortical regions contribute to long-term identity-based voice knowledge, including a deep posterior STS region, the anterior temporal poles and the amygdala — but, unlike in the present study, not the voice-selective IFC. No long-term effects in STS and IFC were found in Andics et al. (2010), probably because these are not large effects and those findings were based on whole volume tests only, while here we restricted our search space to the voice-selective regions.

One important difference between the two studies is that here we trained two voice identities, one at each endpoint of the morph continuum, while in the Andics et al. (2010) study a single voice identity was trained, in the middle of the morph continuum. Consequently, here we could ask whether there are multiple intra-individual voice spaces maintained in a certain brain region — this could not test be tested in the Andics et al. study. We indeed found that IFC's long-term voice identity codes reference the central tendencies of each of the two trained voice identities.

The present results show that short-term adaptation is not independent of long-term acoustic centrality in the voice-sensitive STS. One possibility is that this interaction could reflect differences in the magnitude of the response in this STS region. But if that was the only reason for the dependence, then we would expect greater short-term adaptation effects for peripheral than for central stimuli, proportional to the magnitude of the response. On the contrary, we found that short-term adaptation is stronger for acoustically central (i.e., more expected) than for acoustically peripheral (i.e., less expected) stimuli: this is in accordance with recent findings demonstrating greater short-term repetition suppression for expected than for non-expected stimuli in category-selective regions (Andics et al., 2013; Summerfield et al., 2008). That is, although acoustically peripheral (less expected) stimuli elicit greater brain responses than central (more expected) stimuli, it seems that the repetition of peripheral (less expected) stimuli leads to weaker (and not to stronger) adaptation effects than the repetition of central (more expected) stimuli. This shows that the size of a short-term adaptation effect may not necessarily be proportional to the corresponding response magnitude. Short-term predictability (whether in the short-term a stimulus is more or less expected) thus seems more crucial in the case of a central (long-term more expected) stimulus than in the case of a peripheral (long-term less expected) stimulus.

In previous studies, short-term acoustically driven adaptation effects were found in both the STS and the IFC (Andics et al., 2010, 2013). The discrepancy in short-term adaptation results between the present study and that of Andics et al. (2010) is probably due to differences in design. First, short- and long-term effects were temporally more distinct in this study. We presented voice pairs with a within-pair stimulus-onset asynchrony (SOA) of 800 ms and an across-trial SOA of 8400 ms. In the Andics et al. (2010) study single stimuli were presented with an SOA of 2500 ms. So 'short-term' here means 800 ms, while in the earlier study it meant 2500 ms; 'long-term' here means at least 7000 ms (8400 minus 800 ms minus maximal jitter of 600 ms), while in the previous study it meant at least 5000 ms. So the time gap between adaptor and target was shorter here, and the chance for carry-over from earlier trials was lower. Second, listeners here had a voice A or voice B task, while in the previous study there was a voice A or not A task. Third, here we associated each voice with a name, while in the Andics et al. (2010)

study the trained voice was associated with a name and a face. Fourth, the present study had a classical adaptation design (with repetition and alternation trials), similar to that in Andics et al. (2013), but with a lower overall proportion, and thus a lower expectation of repetition trials than in the (2010) study. Indeed, short-term adaptation effects are known to be extremely sensitive to design details such as time gap between adaptor and target stimulus (Grill-Spector et al., 2006), possible carry-over from earlier trials (Aguirre, 2007), task (Cohen Kadosh et al., 2010; Wagner et al., 2000), cross-modal associations during a pre-test training (Latinus et al., 2011), and attention or expectation effects (Andics et al., 2013; Larsson and Smith, 2012; Summerfield et al., 2008).

It is worth noting that we found mean-based voice coding almost exclusively in the right hemisphere. This converges with clinical (Van Lancker and Canter, 1982) and neuroimaging studies (Belin and Zatorre, 2003; von Kriegstein and Giraud, 2004) reporting greater sensitivity for talker-related features of voice stimuli on the right side of the brain.

Furthermore, the voice learning task listeners were exposed to in this study is very far from the complexity of voice learning under natural circumstances. Many aspects of voice variability are not accounted for, and it is possible that the mechanisms described here are not specific for voice processing. Nevertheless, we argue that we have described neural mechanisms that are crucial for voice identity processing. First, our participants reported that they thought they had heard natural voices only. Second, we first specified cortical regions that are known to be selective for voices, using a well-established functional localizer (as in Belin et al., 2000), and then we tested how brain responses in these voice-selective regions are modulated by voice stimuli's distributional properties and by voice identity feedback. So, rather than optimizing for stimulus relevance and specificity, we optimized for cortical relevance and specificity, and tried to characterize neural coding mechanisms in brain regions that are clearly relevant for natural voice processing. Third, in a behavioral study (reported in Chapter 3 of Andics, 2013), we tested if voice recognition based on training with the stimuli used here generalizes to other, untrained words. We found that it does, even after a shorter training, and even to words with no segmental overlap.

Finally, it is important to note that our findings should be generalized to other voice learning contexts only with caution. We did not test whether prototype-centered category formation takes place in other kinds of voice learning situations. Indeed, it has been argued that category representations are in part determined by the learning regime (Goldstone, 1994). Along these lines, Gentner and Margoliash (2003) demonstrated that neural responses of starlings to conspecific vocalizations are not independent of the training method. The findings presented here are consistent with the mean-based coding account, but further research is clearly required to fully understand how human listeners cope with the difficulties of voice learning in less constrained contexts.

In conclusion, we propose that the right middle STS processes incoming voice stimuli with respect to their distance from the representation of a supra-individual "mean voice" category (i.e., the average across talkers of the listener's recent voice-acoustic history). This representation does not seem to be biased by voice identity information, rather it collapses across individual voices. The right IFC, in contrast, processes voice stimuli with respect to their distance from representations of "individual mean voices" that are the average of the listener's recent memories of the voices of specific individuals. According to this view, the IFC maintains multiple "individual mean voice" representations, one for each voice remembered (see Fig. 3c for a schematic illustration of the proposed representations). In this study, we presented the first evidence for this multilevel long-term mean-based coding in voice-selective cortical regions.

Acknowledgments

This study was conducted as part of A.A.'s PhD project, funded by the Max Planck Society. This work was supported by the Hungarian Academy of Sciences (MTA 01 031) and the Fundação para a Ciência e a Tecnologia (IBB/CBME, LA, FEDER/POCI 2010 to K. M. P.). We thank Benedikt A. Poser for developing the sparse scanning sequence.

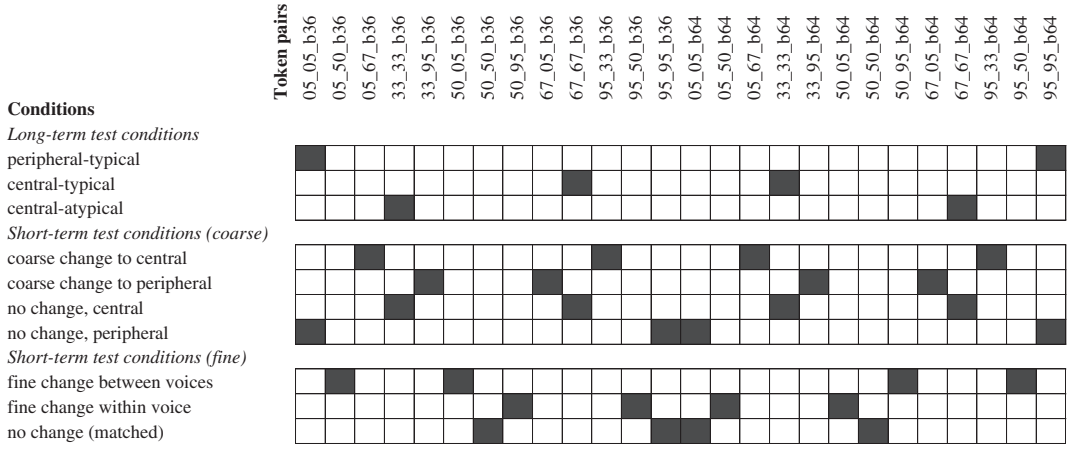
Conflict of interest

The authors declare that they have no conflict of interest.

Appendix A. Training stimuli

Trained voice identity	Trained identity boundary	Mean of all trained morphs	Stimulus morph levels used during training														
A	36	26	1	10	17	22	25	27	28	29	30	31	32	34	34	35	35
B	36	46	37	37	37	38	38	38	39	39	39	40	42	46	55	66	99
A	64	54	1	34	45	54	58	60	61	61	61	62	62	62	63	63	63
B	64	74	65	65	66	66	68	69	70	71	72	73	75	78	83	90	99

Appendix B. Experimental conditions as defined by token pair types of the fMRI tests. For example, '05_50_b36' refers to the token pair type in which the first stimulus was morph 05, the second stimulus was morph50, and the trained identity boundary was at morph36



References

Aguirre, G.K., 2007. Continuous carry-over designs for fMRI. Neuroimage 35, 1480–1494.

Andics, A., 2013. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. PhD dissertation Radboud University Nijmegen (Max Planck Institute Series in Psycholinguistics, 73). Ipskamp Drukkers, Enschede.

Andics, A., McQueen, J.M., Van Turennout, M., 2007. Phonetic content influences voice discriminability. In: Trouvain, J., Barry, W.J. (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007). Dudweiler, Pirrot, pp. 1829–1832.

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage 52, 1528–1540.

Andics, A., Gál, V., Vicsi, K., Rudas, G., Vidnyánszky, Z., 2013. FMRI repetition suppression for voices is modulated by stimulus expectations. Neuroimage 69, 277–283.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A., 2004. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat. Neurosci. 7, 1190–1192.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105–2109.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. 8, 129–135.

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R., 2011. Understanding voice perception. Br. J. Psychol. 102, 711–725.

Bestelmeyer, P.E.G., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., Belin, P., 2012. Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. Cereb. Cortex 22, 1263–1270.

Boersma, P., Weenink, D., 2007. Praat: Doing Phonetics by Computer (Version 4.2.07) ([Computer program]).

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Curr. Biol. 20, 116–120.

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2013. Cerebral processing of voice gender studied using a continuous carryover fMRI design. Cereb. Cortex 23, 958–966.

Cohen Kadosh, K., Henson, R., Cohen Kadosh, R., Johnson, M., Dick, F., 2010. Task-dependent activation of face-sensitive cortex: an fMRI adaptation study. J. Cogn. Neurosci. 22, 903–917.

Epstein, R.A., Parker, W.E., Feiler, A.M., 2008. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. J. Neurophysiol. 99, 2877–2886.

Ethofer, T., Kreifelts, B., Wiethoff, S., Wolf, J., Grodd, W., Vuilleumier, P., Wildgruber, D., 2009. Differential influences of emotion, task, and novelty on brain regions underlying the processing of speech melody. J. Cogn. Neurosci. 21, 1255–1268.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P., 2009. Decoding of emotional information in voice-sensitive cortices. Curr. Biol. 19, 1028–1033.

Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., Vuilleumier, P., 2012. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cereb. Cortex 22, 191–200.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Sensitivity to voice in human prefrontal cortex. J. Neurophysiol. 94, 2251–2254.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970–973.

Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T.E., Penny, W.D. (Eds.), 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, London.

Gentner, T.Q., Margoliash, D., 2003. Neuronal populations and single cells representing learned auditory objects. Nature 424, 669–674.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nat. Neurosci. 7, 801–802.

Goldstone, R.L., 1994. Influences of categorization on perceptual discrimination. J. Exp. Psychol. Gen. 123, 178–200.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., Vuilleumier, P., 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. Nat. Neurosci. 8, 145–146.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10, 14–23.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J., 2008. Detection of time-varying signals in event-related fMRI designs. Neuroimage 43, 509–520.

Hackett, T.A., Stepniewska, I., Kaas, J.H., 1998. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. J. Comp. Neurol. 394, 475–495.

Hoffman, K.L., Logothetis, N.K., 2009. Cortical mechanisms of sensory learning and object recognition. Philos. Trans. R. Soc. B 364, 321–329.

Kahn, D.A., Aguirre, G.K., 2012. Confounding of norm-based and adaptation effects in brain responses. Neuroimage 60, 2294–2299.

Kawahara, H., 2006. STRAIGHT, exploration of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. Acoust. Sci. Technol. 27, 349–353.

Larsson, J., Smith, A.T., 2012. FMRI repetition suppression: neuronal adaptation or stimulus expectation? Cereb. Cortex 22, 567–576.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Front. Psychol. 2, 1–12. http://dx.doi.org/10.3389/fpsyg.2011.00175.

Latinus, M., Crabbe, F., Belin, P., 2009. FMRI investigations of voice identity perception. Organization for Human Brain Mapping 2009 Annual Meeting, July 2009. Neuroimage 47 (Supplement 1), S156.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb. Cortex 21, 2820–2828.

Leopold, D.A., O'Toole, A.J., Vetter, T., Blanz, V., 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. Nat. Neurosci. 4, 89–94.

Leopold, D.A., Bondar, I., Giese, M.A., 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature 442, 572–575.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. FMRI evidence for the neural representation of faces. Nat. Neurosci. 8, 1386–1390.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2009. Typicality effects on memory for voice: implications for earwitness testimony. Appl. Cogn. Psychol. http://dx.doi.org/10.1002/acp.1635.

Panis, S., Wagemans, J., Op de Beeck, H.P., 2011. Dynamic norm-based encoding for unfamiliar shapes in human visual cortex. J. Cogn. Neurosci. 23, 1829–1843.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. J. Acoust. Soc. Am. 85, 913–925.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374.

Rhodes, G., Jeffery, L., 2006. Adaptive norm-based coding of facial identity. Vision Res. 46, 2977–2987.

Romanski, L.M., Goldman-Rakic, P.S., 2002. An auditory domain in primate prefrontal cortex. Nat. Neurosci. 5, 15–16.

Romanski, L.M., Bates, J.F., Goldman-Rakic, P.S., 1999. Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. J. Comp. Neurol. 403, 141–157.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J. Neurophysiol. 93, 734–747.

Schwarzbauer, C., Davis, M.H., Rodd, J.M., Johnsrude, I.S., 2006. Interleaved silent steady state (ISSS) imaging: a new sparse imaging method applied to auditory fMRI. Neuroimage 29, 774–782.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26, 100–107.

Stevens, A.A., 2004. Dissociating the cortical basis of memory for voices, words and tones. Cogn. Brain Res. 18, 162–171.

Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.M., Egner, T., 2008. Neural repetition suppression reflects fulfilled perceptual expectations. Nat. Neurosci. 11, 1004–1006.

Van Lancker, D., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain Cogn. 1, 185–195.

von Kriegstein, K., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948–955.

von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biol. 4, e326.

Wagner, A.D., Koutstaal, W., Maril, A., Schacter, D.L., Buckner, R.L., 2000. Task-specific repetition priming in left inferior prefrontal cortex. Cereb. Cortex 10, 1176–1184.

Warren, J., Scott, S., Price, C., Griffiths, T., 2006. Human brain mechanisms for the early analysis of voices. Neuroimage 31, 1389–1397.

Wong, P.C., Nusbaum, H.C., Small, S.L., 2004. Neural bases of talker normalization. J. Cogn. Neurosci. 16, 1173–1184.