

# Neural dissociation of the acoustic and cognitive representation of voice identity

Bestelmeyer, Patricia; Mühl, Constanze

**Neuroimage**

*DOI:*
https://doi.org/10.1016/j.neuroimage.2022.119647

Published: 01/11/2022

Publisher's PDF, also known as Version of record

NeuroImage 263 (2022) 119647

Contents lists available at ScienceDirect

NeuroImage

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

# Neural dissociation of the acoustic and cognitive representation of voice identity

Patricia E.G. Bestelmeyer*, Constanze Mühl

*Institute of Cognitive Neuroscience, Bangor University, UK*

ARTICLE INFO

*Keywords:*
Carry-over design
fMRI adaptation
Individual differences
Repetition suppression
Voice identity

ABSTRACT

Recognising a speaker's identity by the sound of their voice is important for successful interaction. The skill depends on our ability to discriminate minute variations in the acoustics of the vocal signal. Performance on voice identity assessments varies widely across the population. The neural underpinnings of this ability and its individual differences, however, remain poorly understood. Here we provide critical tests of a theoretical framework for the neural processing stages of voice identity and address how individual differences in identity discrimination mediate activation in this neural network. We scanned 40 individuals on an fMRI adaptation task involving voices drawn from morphed continua between two personally familiar identities. Analyses dissociated neuronal effects induced by repetition of acoustically similar morphs from those induced by a switch in perceived identity. Activation in temporal voice-sensitive areas decreased with acoustic similarity between consecutive stimuli. This repetition suppression effect was mediated by the performance on an independent voice assessment and this result highlights an important functional role of adaptive coding in voice expertise. Bilateral anterior insulae and medial frontal gyri responded to a switch in perceived voice identity compared to an acoustically equidistant switch within identity. Our results support a multistep model of voice identity perception.

## 1. Introduction

Effective communication and interaction rely on our ability to discriminate small acoustic variations in the vocal signal. This skill allows us to differentiate between subtly different speech sounds over time and to make sense of the linguistic message. Independently of deciphering the speech content, this ability also helps us interpret a myriad of socially impactful information carried by the speaker's voice such as the speaker's identity, geographical origin, and emotional state. The ability to recognise the identity of a familiar speaker is present shortly after birth (e.g., Beauchemin et al., 2011) and has a long evolutionary history (Belin, 2006). The fact that many species have developed the ability to recognise an individual's identity purely based on the sound of a conspecific's voice, underlines the biological importance of this skill. Despite its importance, the ability of identifying a person from their voice is characterised by large individual variability (e.g., Mühl et al., 2018; Lavan et al., 2019). The neural underpinnings of this behaviour and its variability between individuals are thus far poorly understood and are the aims of the current study.

Much of what we know about the neural correlates of person identification is based on the systematic investigations of the ability to recognise faces and when this ability goes wrong. A face recognition impairment, called prosopagnosia, is predominantly due to right hemisphere damage (e.g., Bodamer, 1947; Hecaen et al., 1952; De Renzi and Spinnler, 1966; Warrington and James, 1967; Yin, 1970; Benton and Van Allen, 1972; De Renzi et al., 1994). While voice recognition is much less well explored, some advances regarding the neural underpinnings have been made in understanding voice recognition. The first studies to describe a selective voice recognition impairment, coined phonagnosia, showed that just as with face recognition disorders, phonagnosia is more common after damage to the right hemisphere (Assal and Aubert, 1979; Van Lancker and Canter, 1982). This observation has been confirmed subsequently and is independent of a face recognition impairment (Van Lancker and Kreiman, 1987; Neuner and Schweinberger, 2000). Phonagnosia has been observed despite intact musical and language skills (Luzzi et al., 2018) adding further support to the notion that voice identity recognition is subserved by a dedicated neural network. Roswandowitz et al. (2018) conclude that the most likely area to mediate voice identity recognition, based on the overlap of lesions in most patients with acquired phonagnosia, is the right inferior parietal lobule.

In contrast to these clinical reports, neuroimaging research suggests that the temporal voice areas (TVAs), located bilaterally along the upper banks of the superior temporal gyri and sulci, are crucial for voice processing (Belin et al., 2000; Kriegstein and Giraud, 2004; Lewis et al., 2009; Bestelmeyer et al., 2011; Pernet et al., 2015; Agus et al., 2017).

* Corresponding author at: Institute of Cognitive Neuroscience, Brigantia Building, Bangor University, Gwynedd LL57 2AS, UK.
*E-mail address:* p.bestelmeyer@bangor.ac.uk (P.E.G. Bestelmeyer).

https://doi.org/10.1016/j.neuroimage.2022.119647.
Received 12 May 2022; Received in revised form 18 August 2022; Accepted 22 September 2022
Available online 24 September 2022.

The TVAs consist of an interconnected bilateral system of "voice patches", which reliably emerge when contrasting vocal with non-vocal sounds (Pernet et al., 2015). Bestelmeyer et al. (2011) demonstrated behavioural importance of the right TVA in voice/non-voice discrimination using repetitive TMS. The precise role of the temporal voice patches, specifically for voice identity perception and recognition is still unclear.

In one of the first fMRI studies on speaker identity, von Kriegstein et al. (2003) measured brain responses during two identification tasks that directed attention either towards determining the speaker or the verbal content of the sentences. The right anterior STS (a part of the TVA) and part of the right precuneus showed increased activations during the speaker identification task compared to the linguistic task, whereas the left middle STS was more active in the reverse contrast (see also Imaizumi et al. (1997) and Belin and Zatorre (2003) for convergent findings). In later studies that manipulate the acoustic content of voices, the TVAs have been shown to code their acoustic representations (e.g., Andics et al., 2010; Latinus et al., 2011; Bestelmeyer et al., 2012; Latinus et al., 2013; Bestelmeyer et al., 2014). Aglieri et al. (2018) then established that the degree of functional connectivity between the TVAs and voice-sensitive regions in prefrontal cortex covaries with voice memory performance. However, voice-induced BOLD activation in the TVAs did not predict voice memory abilities directly (Watson et al., 2012). A missing link in the current literature is a clear connection between the magnitude of the neural activation in the TVAs and behavioural performance on an unrelated identity assessment (i.e., unrelated to the task in the scanner). Such a link might allow us to paint a clearer picture regarding the function of these voice patches.

Theoretically, the processing of the acoustic properties of a voice and the processing of the identity representation of that voice are two distinct stages of identity perception (Belin et al., 2004). Using adaptation paradigms and voice morphs of learned identities, Andics et al. (2010) and Latinus et al. (2011) attempted to neurally dissociate these two processing stages. Andics et al. (2011) showed that increasing acoustic similarity to the preceding voice led to an adaptation effect, i.e., a reduction in signal, in bilateral middle and posterior STS and in right ventrolateral prefrontal regions. Voice identity processing on the other hand, involved bilateral STS, anterior temporal pole and left amygdala. Thus, bilateral STS, as part of the core voice perception network, showed sensitivity to *both* acoustic and identity processing. However, as Latinus et al. (2011) point out, the morphs included in the identity contrast also differed acoustically in Andics et al.'s experiment. In contrast, a voice learning study by Latinus et al. (2011) demonstrated that the right middle and superior STS responded to acoustic change, but only right posterior inferior frontal gyrus/precentral gyrus and left cingulate gyrus were sensitive to changes in perceived identity. Latinus et al.'s finding suggests that the TVAs are not involved in the higher order perception of voice identity when acoustic distance is controlled. More recent neuroimaging studies have highlighted the importance of extratemporal regions, in addition to the core network, in the recognition of speaker identity, particularly, inferior frontal areas and anterior insula (e.g., Aglieri et al., 2021; McGettigan et al., 2013; Zäske et al., 2017).

Adaptation paradigms are a powerful tool to explore the perceptual representation of certain stimulus attributes and probe how these may be coded in the brain. Behaviourally, adaptation refers to a process during which continued stimulation results in biased perception toward opposite features of the adaptor (Grill-Spector et al., 1999). Adaptation can reveal neural populations tuned to respond to specific stimulus attributes by isolating and subsequently distorting the perception of these attributes (Grill-Spector et al., 1999; Winston et al., 2004). These attributes can range from simple features such as pitch to higher-level, abstract features such as the identity of a speaker. In the latter case, morphing techniques are often employed to demonstrate these aftereffects (e.g., Zäske et al., 2010; Latinus and Belin, 2011; Bestelmeyer and Mühl, 2021). Neurally, fMRI adaptation to a specific stimulus feature is typically accompanied by a decrease in the hemodynamic response (also referred to as repetition suppression). Again, fMRI adaptation paradigms are often used to probe the sensitivity of neural populations to a specific stimulus dimension (i.e., to reveal functional specificity of neural populations).

While adaptation is ubiquitous in perception and has been used experimentally for many decades, its functional role is still debated. A range of explanations have been offered, most of which involve coding efficiency (Wainwright, 1999; Clifford et al., 2007; Wark et al., 2007; Webster, 2011). For example, adaptation may continuously recalibrate perceptual norms to maintain a match between coding and environment (Webster, 2011). It may also enhance coding by reducing the sensitivity to continued stimulation, which in turn enhances sensitivity to change (Webster, 2011). Recently, Bestelmeyer and Mühl (2021) have demonstrated that adaptive coding of voices contributes to our ability to discriminate and recognise voices. In two experiments we showed that larger aftereffect sizes, measured at the most ambiguous morph between two familiar identities, were linked with better voice perception ability. We also showed that this effect was specific to voices and was not related to general auditory abilities or how much a person may adapt generally to other sound categories. The data clearly support a functional role of adaptive coding in voice expertise. Our results mirrored those from the face literature (Dennett et al., 2012; Rhodes et al., 2014; Rhodes et al., 2015; Engfors et al., 2016) in which face identity aftereffects positively correlated with face memory tests but not non-face, object memory tests. While there are some reports on the relationship between BOLD signal and the behavioural performance on a given voice task in the scanner (e.g., Andics et al., 2010; Aglieri et al., 2021), the relationship of the performance on an *independent* voice assessment with BOLD signal changes to voice identity have not been systematically investigated. Previous results in the face literature have suggested that measuring sensitivity with repetition suppression might be more suitable to uncover the relationship between face perception and its neural bases (e.g., Goh et al., 2010; Jiang et al., 2013; Hermann et al., 2017).

The aims of this study were twofold. First, we critically examined a theoretical framework which proposes the functional distinction between core and extratemporal regions for voice processing (Belin et al., 2004). To this end, we created continua consisting of seven morphs with equal physical (i.e., acoustic) distances between utterances of two familiar speakers. We embedded these short, morphed nonsense syllables in an unbroken, balanced sequence to study the adaptation effects due to acoustic similarity between morphs. We predicted a decrease in fMRI signal, or repetition suppression, with acoustically more similar morphs in bilateral TVAs. The design of the stimulus sequence also allowed us to explore activation patterns in response to equidistant changes in acoustics. These changes could result in a switch in perceived identity or could be within a perceived identity. In other words, we compared identical shifts in acoustics within and across the category boundary of the two identities. When comparing pairs of 30% physical change within and across the category boundary we predicted that pairs which crossed the category boundary will activate areas of the extended voice perception network, such as bilateral insulae and inferior frontal gyri, compared to pairs that consisted of an equal physical shift but stayed within the same identity (see Rotshtein et al. (2005) for an equivalent methodological approach in the face literature).

Second, we were interested in the relationship between individual differences in voice perception ability and repetition effects as reported in our recent behavioural study (Bestelmeyer and Mühl, 2021). We therefore administered the Bangor voice matching test (BVMT; Mühl et al., 2018) to our participants and used the scores as a covariate in our analysis of the repetition suppression effect. We predicted that individuals with better performance on the voice matching test will be more sensitive to subtle changes in the acoustics of the voice and will therefore show less repetition suppression with the repetition of increasingly similar voices compared to individuals at the lower end of the performance spectrum. Previous literature has found evidence that the TVAs might be involved in early identity-specific processes (e.g.,

Schall et al., 2015). In contrast, patient work and some neuroimaging research has shown that the TVAs are not involved in coding the representation of identity. This latter research reports areas belonging to the extended voice perception network. Patient work highlights the right inferior parietal lobule and neuroimaging research suggests the involvement of inferior frontal regions (including insulae) for processing of speaker identity. We therefore predicted that the repetition suppression effect mediated by the voice test performance could involve the voice patches or the extended voice network. We also tested whether this repetition suppression effect to identity is specifically mediated by voice discrimination ability or whether it could relate to individuals' general auditory discrimination skills (PROMS; Law and Zentner, 2012). To this end we administered a control assessment on general auditory abilities such as pitch and rhythm matching (PROMS; Law and Zentner, 2012). We did not expect the repetition suppression effect to voice identity to covary with the scores of this general auditory abilities test.

## 2. Methods

### 2.1. Participants

The experiment consisted of two sessions. In a first session, we screened 135 volunteers (93 females; mean age = 21.36; S.D. = 3.19) from our student cohort for their suitability to take part in the MRI experiment. Of these participants, based on MRI safety and task performance, 43 individuals met all criteria and responded to our invitation to be scanned (session 2). Three of these participants had to be excluded due to excessive movement or poor task performance in the scanner. The reported analyses are based on 40 participants (27 females; 4 left-handed; 1 ambidextrous; mean age = 20.98; S.D. = 2.40) with self-reported normal hearing, without history of neurological problems or regular intake of medication. All five adextral volunteers had previously participated in an unrelated MRI study on verbal fluency and were confirmed to be left-brain dominant for language (Johnstone, Karlsson and Carey, 2020). All participants were reimbursed with course credit for their time. Informed consent was obtained from all individuals and the study protocol was approved by the ethics committee at Bangor University.

### 2.2. Materials and procedure (session 1)

We first invited participants from specific year groups of the Psychology degree to participate in a screening session. Volunteers were MRI safety screened and completed a behavioural test battery consisting of a speaker familiarity task, a speaker categorisation task, the voice test (BVMT; Mühl et al., 2018) and a general auditory ability test (Brief-PROMS; Law and Zentner, 2012). The tasks were administered in that same order and the full session lasted 90 minutes (see SFig. 1 for a schematic illustration of the four tasks). Each one of the behavioural tasks took approximately 10-15 minutes to complete and were conducted with headphones (Beyerdynamic DT770 Pro (250 Ω)) in our lab. Sound levels were set to 75 dB SPL (C).

### 2.3. Familiarity task

First, we created a task that assessed the familiarity of our participants with the two identities used in the scanning protocol and two foils. We obtained recordings (16-bit, 44.1 kHz sampling rate, mono) of four female university lecturers uttering sixteen different nonsense syllables in a sound attenuated booth (e.g., "aba", "ada", "hod"; see Mühl et al. (2018) for details on how stimuli were recorded and created). We used nonsense syllables to support recognition from the sound of the speaker's voice alone (rather than additional paralinguistic features such as accent or speech rate). All 64 stimuli were edited in Cool Edit Pro and normalised in energy (root mean square; RMS). These Psychology lecturers were of similar age and regularly taught several hours per week across the year groups we recruited from.

During this task participants listened to these non-sense syllables and were asked to press one of four keys to indicate the correct identity. Identity labels were displayed at the top of the screen throughout the task. The 64 trials were presented randomly with a response window of four seconds. The participant received auditory feedback in the form of a bell sound for correct identification and a buzzer for incorrect identification. Once the participant had responded, the next syllable was presented within one second. To ensure that participants recognised the speakers, only participants with an overall success rate of above 75% on this four-alternative forced choice task were considered for the scanning session.

### 2.4. Categorisation task

For the categorisation task, which we administered in both experimental sessions, we morphed two female identities on the same five syllables ("aba", "aga", "hid", "hod", "udu"). Each of the five morphed continua consisted of seven morph steps that corresponded to 5%/95%, 20%/80%, 35%/65%, 50%/50%, 65%/35%, 80%/20%, 95%/5% of SpeakerA/SpeakerB and were created using Tandem-STRAIGHT (Kawahara et al., 2008) in MatlabR2013a (The Mathworks, Inc). The waveforms and spectra of a morphed continuum are illustrated in SFig. 2(A). Tandem-STRAIGHT performs an instantaneous pitch-adaptive spectral smoothing of each stimulus for separation of contributions to the voice signal arising from the glottal source (including f0) versus supralaryngeal filtering (distribution of spectral peaks, including the first formant frequency, F1). Voice stimuli were decomposed by Tandem-STRAIGHT into five parameters: fundamental frequency (f0; the perceived pitch of the voice), frequency, duration, spectrotemporal density and aperiodicity. Each parameter can be manipulated independently. For each recording of a specific syllable, we manually identified one time landmark with three frequency landmarks (corresponding to the first three formants) at the onset of phonation and the same number of landmarks at the offset of phonation. Morphed stimuli were then generated by re-synthesis based on the interpolation (linear for time; logarithmic for F0, frequency, and amplitude) of these time-frequency landmark templates (see also Schweinberger et al. (2014) for a discussion of the voice morphing technique). All stimuli were RMS normalised before and after morphing.

Participants were asked to perform a two-alternative forced choice task by labelling each morph as belonging to Speaker A or B by pressing one of two corresponding buttons. The task consisted of five repetitions of each of the five morphed continua. The seven morph steps within each continuum were randomised. Each trial consisted of a response window of three seconds followed by a fixation cross for two seconds. The categorisation task in session 1 consisted of 175 trials.

### 2.5. Bangor voice matching test (BVMT)

The BVMT presents two different nonsense syllables per trial (Mühl et al., 2018). The two syllables can either be spoken by the same (50% of trials) or by two different speakers. The test consists of 40 trials for each speaker sex. During each trial participants make same/different speaker decisions.

### 2.6. Brief-PROMS

This sound perception test consists of four subtests that tap into general auditory skills such as rhythm perception (Law and Zentner, 2012). Each trial follows the same structure: participants listen to two identical melodies, followed by a third that can be the same or differ slightly in melody, tempo, rhythm, or tuning (depending on the subtest). Participants then make a same/different decision, including levels of confidence involving, "definitely same", "probably same",

"probably different", "definitely different" and "don't know". While data were collected in the lab, this test was delivered via an online platform. The test can be accessed via this link (https://www.uibk.ac.at/psychologie/fachbereiche/pdd/personality_assessment/proms/take-the-test/brief-proms/) and takes 30 minutes to complete.

### 2.7. Image acquisition and fMRI paradigm (session 2)

All scans were acquired with a Philips 3 Tesla Achieva MR scanner using a 32-channel SENSE head coil. The scanning session consisted of three experimental runs, a voice localiser, and an anatomical scan. For the experimental runs, we used T2*-weighted functional scans (TR = 2s, TE = 30ms), with interleaved ascending sequence of 35 slices, no slice gap. FOV was 240 × 240 × 105mm, with a voxel size of 3 $mm^3$, an acquisition matrix of 80 × 78, a flip angle of 77°, and 370 volumes per run. We acquired three of these functional runs, with each one lasting 12:34 minutes.

We employed a continuous carry-over design (CCO; Aguirre, 2007) to measure the effects of one stimulus upon the next using a first-order serially balanced sequence of stimuli known as type-1-index-1 (Nonyane and Theobald, 2007). In this sequence each stimulus is preceded and followed by itself and every other stimulus an equal number of times and was defined by eight stimulus types (seven morph steps plus one null event [silence]) totalling 65 stimuli. A sample CCO sequence can be seen in SFig. 2(B). We presented a CCO sequence of 65 stimuli five times per run (one CCO sequence per syllable). With every new presentation of a full CCO sequence, we randomised the assignment of stimulus type to the numbers 1 to 8 in the CCO sequence (i.e., the silent event was stimulus 8 during the first full sequence but then was assigned to stimulus 6 during the second presentation of the full sequence and so on). Each trial lasted one TR (2s) and each full CCO sequence of stimuli was divided by nine TRs of silences (18s). Participants were asked to perform a two-alternative forced choice task in which each morph had to be categorised as either Speaker A or B using an MRI compatible response box (fORP; Current Designs, Inc.). We asked participants to respond as quickly as possible (see SFig. 2(C) for illustration of the categorisation task in the scanner). This categorisation task was the same as in session 1 except for differences in total trial numbers and a shorter response interval (response within TR of 2 seconds).

Following the experimental runs, participants completed a "voice localiser" scan (Belin et al., 2000; Pernet et al., 2015). Imaging parameters were the same as for the experimental scans, with the exception that we acquired 310 volumes for the localiser (10:34 minutes). During this block design, participants passively listened to various sounds. Stimuli were presented in 60 blocks (each lasting 5 TRs), consisting of 20 vocal sound blocks (e.g., words, vocal expressions, humming), 20 environmental sound blocks (e.g., objects, instruments, animal sounds) and 20 silent blocks. This localiser allows identification of the temporal voice areas (TVAs) using the vocal versus non-vocal contrast. All stimuli were presented binaurally using the electrostatic NNL headphone system with passive noise attenuation of 30dB at 1kHz and enhanced hearing protection (NordicNeuroLab, Inc.). Sounds were presented at an intensity of 80dB SPL(C) while EPIs were acquired (see SFig. 2(C)). We asked participants to keep their eyes closed during all functional runs. All tasks, except the brief PROMS, were implemented in Psychtoolbox-3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) for Matlab.

We also acquired a whole brain anatomical scan. This T1-weighted scan had the following parameters: field of view (FOV): 240 × 240 × 175, voxel size: 1 $mm^3$; 175 slices, with an acquisition matrix of 240 × 224, TR = 12 s, TE = 3.5 ms and a scan duration of 5:38 minutes.

### 2.8. Image analysis

Analysis of all MRI data was conducted using SPM12 (The Wellcome Trust Centre for Human Neuroimaging, University College London; available at https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) with Matlab 2020b. Pre-processing of the data consisted of AC-PC alignment of the anatomical images (and application of the orientation change to all functional images acquired in the same session). Functional scans were corrected for head motion (trilinear interpolation) by aligning all scans to the first scan of the last run and a mean image was created. The anatomical scan was co-registered to the mean image. Functional and anatomical data were transformed to Montreal Neurological Institute (MNI) space after segmentation of the anatomical scan. Normalised data were spatially smoothed by applying a Gaussian kernel of 8mm full width at half maximum (FWHM).

The analysis of the carry-over, or repetition suppression, effect followed Aguirre (2007) and employed a parametric modulation analysis (Büchel et al., 1998) at the first level. For each morph we calculated the physical difference between that stimulus n and the one immediately following it (n+1). The physical difference was calculated as the absolute difference in morph step between n and n+1, e.g., if the 65% Speaker B morph step was followed by a 20% Speaker B morph step, the absolute physical difference between both was 45% (see Fig. 1A). The design matrix of this first-level analysis contained the voice onset events of all n+1 trials as the first regressor which was followed by the physical differences as parametric regressor. We also included the onsets of all first sounds of each CCO sequence, the silent null events within blocks (which were skipped in the calculation of the physical difference) and the six movement parameters in our model. Onsets of first sounds in each sequence were modelled as a separate regressor because no sound preceded it and therefore no carry-over effect could be computed. Second-level analysis of the physical difference regressor consisted of a one-sample t-test allowing for the evaluation of positive and negative correlations. We used a family-wise error (FWE) corrected voxelwise threshold of $p < .05$ for this whole brain analysis (height threshold of T = 5.17, extent threshold k = 20 voxels).

In an additional second level analysis of the carry-over effect, we entered the voice test score (BVMT % correct) for each participant as a covariate. We performed a region of interest analysis to investigate the effects of voice perception skill on the size of the carry-over effect. The regions of interest (ROI) were defined a priori and are based on previous literature as reviewed in our introduction (and schematically illustrated in SFig. 3). We have one functionally and two structurally defined ROIs. The first ROI involved bilateral TVAs based on our voice localiser obtained from our sample of participants. We used an FWE-corrected voxel-wise threshold of $p < .05$ for this whole brain analysis to generate the ROI (height threshold of T = 5.59, extent threshold k = 20 voxels; left TVA: -60 –22 –1; t(38) = 13.40, k = 397; right TVA: 60 -16 –1; t(38) = 14.10, k = 677). This thresholded contrast was saved as a binary image. The additional two masks were structural masks. These masks were made using the WFU Pickatlas (http://fmri.wfubmc.edu/software/pickatlas) and were based on the neuropsychological literature (bilateral inferior parietal lobule, including supramarginal and postcentral gyrus (Roswandowitz et al., 2018)), and previous neuroimaging literature (bilateral inferior frontal gyri and insulae (Andics et al. 2010; Latinus et al. 2011)). Statistical significance for the ROI analyses was set at a threshold of $p < .001$ with FWE-correction of $p < .05$ at the cluster level. We ran the same analysis with the control test score of the general auditory abilities assessment (PROMS).

To investigate the brain regions sensitive to a change in identity rather than acoustic representation of the voice, we ran another general linear model. This model included the onset regressors for trials in which the second voice differed 30% in acoustics from the first voice. This acoustic change was either within the same identity or meant a change in perceived identity by crossing the category boundary (see Fig. 1B for illustration of within and across identity trials). Trials included in this model were the 5% and 35% (%Speaker B) morphs ("within identity" trials) and the 35% and 65% (%Speaker B) morphs ("across identity" trials). We only included one speaker in the "within trials". This is because the carry-over sequence meant that there were fewer "across identity"

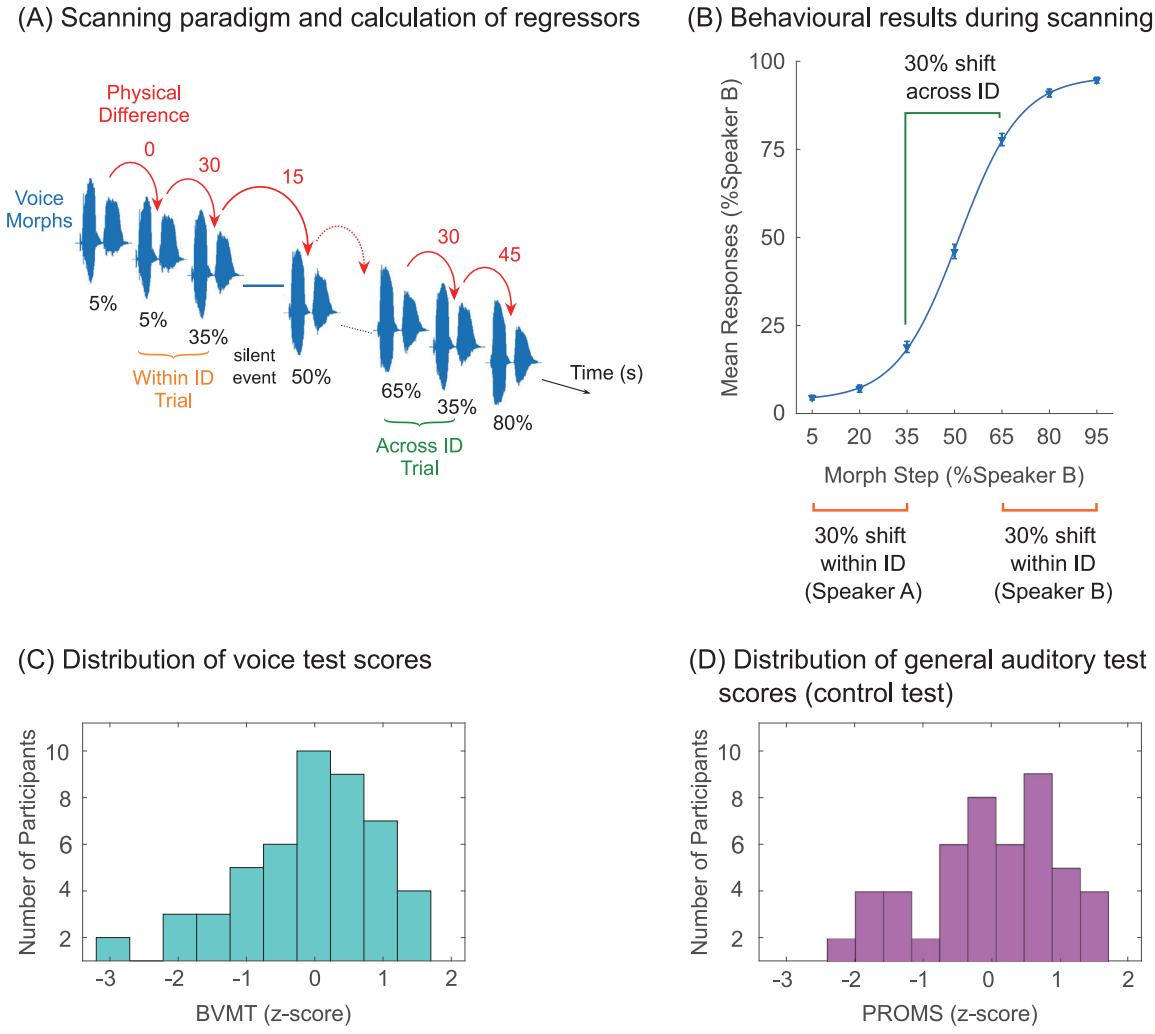


**Fig. 1.** Paradigm and behavioural data. (A) Illustration of the calculation of the "physical" difference regressor (or carry-over effect). The physical difference regressor is calculated based on the absolute difference in morph step between two consecutively presented stimuli. To investigate the effect of identity change we analysed "within identity" trials and "across identity" trials (both with identical acoustic shift of 30%). (B) Behavioural results of the identity categorisation task within the scanner with clear category boundary between two voice identities. Error bars represent standard error of the mean (SEM); see also SFig 4 for direct comparison of performance on the categorisation task in sessions 1 and 2. (C) Histogram illustrating the wide distribution of scores on the Bangor voice matching test in the scanned sample of participants. (D) Histogram illustrating the distribution of scores on the general auditory perception test in the scanned sample of participants.

trials than "within identity" trials. The design matrix included within and across identity trials as the first two onset regressors. We also included an additional 8 regressors in the model: the remaining sound onsets, the within-block silent events, and all movement regressors. At the second level we performed a one-sample t-test over the across > within identity contrast from the first level analyses to isolate the areas that responded to a switch in identity. Statistical significance for this whole brain analysis was set at a threshold of $p < .001$ with FWE-correction of $p < .05$ at the cluster level. Note that a separate analysis in which "same identity" trials of *both speakers* were used, and trial numbers down-sampled to match the number of "within identity" trials, resulted in very similar activations.

Lateralization indices (LI) were computed with the LI-Toolbox plugin for SPM12 (Wilke and Schmithorst, 2006) to quantify the asymmetry of functional activation. This toolbox implements a bootstrapping algorithm to avoid commonly cited issues with computing LIs (Wilke & Lidzba, 2007). It relies on the basic computation LI = (Left – Right) / (Left + Right) with Right and Left representing the effect sizes of the right and left clusters peak voxel. Six bilateral ROIs were created in WFU pickatlas with a dilation factor of 3 (bilateral superior temporal and supramarginal gyri, superior and middle orbitofrontal gyri, supplementary motor areas, putamen, precuneus, IFGs/insulae). A negative LI-value indicates right hemispheric dominance, and a positive value indicates left hemispheric dominance.

## 3. Results

### 3.1. Behavioural results

Results of the behavioural data acquired in the scanner is summarized in Fig. 1B and illustrated in comparison to the data acquired in the categorisation task in session 1 in SFig 4. Performance across the two sessions is identical despite the addition of scanner noise in session 2. In each session, the response data were averaged as a function of the seven morph steps, and a psychophysical curve (based on the hyperbolic tangent function) was fitted to the mean data (see Bestelmeyer and Mühl (2021) for similar behavioural results). Participants more frequently categorized the first three morph steps as belonging to Speaker A and the last three morph steps as belonging to Speaker B, while the 50% morph was perceived as the most ambiguous. Despite an equal acoustic change of 30%, behavioural responses are maximally different across the category boundary of the two identities.

**Table 1**
Significant clusters for the analyses illustrated in Fig. 2.

| Analysis | Anatomical definition | Peak voxel coordinates | | | T-value | Cluster size |
|---|---|---|---|---|---|---|
| | | x | y | z | | |
| **(A) Physical Difference** | | | | | | |
| (carry-over effect) | **Left middle temporal gyrus** | **-63** | **-40** | **5** | **10.31** | 875 |
| | Left superior temporal gyrus | -51 | -16 | 2 | 10.06 | |
| | Left middle temporal gyrus | -57 | -13 | -4 | 9.99 | |
| | **Right middle temporal gyrus** | **60** | **-4** | **-13** | **9.66** | 805 |
| | Right superior temporal gyrus | 63 | -19 | -1 | 8.62 | |
| | Right middle temporal gyrus | 66 | -25 | -7 | 8.27 | |
| | **Left putamen** | **-21** | **8** | **2** | **9.59** | 279 |
| | Left posterior orbitofrontal cortex | -15 | 8 | -19 | 5.74 | |
| | **Left supplementary motor area** | **-3** | **-7** | **53** | **7.44** | 166 |
| | Right supplementary motor area | 9 | -7 | 56 | 6.79 | |
| | **Left postcentral gyrus** | **-39** | **-40** | **53** | **7.27** | 301 |
| | Left postcentral gyrus | -30 | -34 | 41 | 6.95 | |
| | Left inferior parietal lobule | -51 | -28 | 35 | 6.21 | |
| | **Left superior frontal cortex** | **-24** | **-10** | **50** | **6.94** | 146 |
| | Left precentral cortex | -36 | -10 | 50 | 6.30 | |
| | Left precentral cortex | -36 | -16 | 56 | 6.28 | |
| | **Right precuneus** | **12** | **-52** | **38** | **6.91** | 235 |
| | Left precuneus | -6 | -55 | 44 | 6.60 | |
| | Left precuneus | -12 | -58 | 56 | 6.47 | |
| | **Right putamen** | **24** | **5** | **2** | **6.83** | 119 |
| | Right hippocampus | 33 | -10 | -16 | 6.76 | |
| | **Right medial orbitofrontal gyrus** | **6** | **44** | **-10** | **6.15** | 46 |
| | Left medial orbitofrontal gyrus | -6 | 38 | -10 | 5.37 | |
| **(B) Voice Test** | | | | | | |
| (covariate effect) | **Left inferior frontal gyrus** (within bilateral ROI) | **-51** | **2** | **20** | **4.10** | 49 |
| | **Right supramarginal gyrus** (within bilateral ROI) | **66** | **-22** | **23** | **5.77** | 62 |
| **(C) Identity representation** | | | | | | |
| Across ID > Within ID | **Medial superior frontal** | **9** | **17** | **47** | **5.26** | 222 |
| | **Right insula** | **33** | **26** | **-1** | **4.87** | 125 |

The distribution of the scanned participant's test scores on the voice matching test (BVMT) and the general auditory ability test (PROMS) is illustrated in Fig. 1C and D, respectively. Test scores did not significantly correlate (Pearson's r = -.099; p = .54) and the distributions of scores did not differ significantly from a normal distribution (Kolmogorov-Smirnov: both statistic < .14). The absence of a correlation between the two test scores suggests that each test taps into a distinct auditory ability.

### 3.2. fMRI results

Significant clusters for all analyses are illustrated in Fig. 2 A-C and are summarised in Table 1.

(A) Repetition suppression effect of the physical difference between consecutive morphs

The repetition suppression, or carry-over effect, of the physical difference between morphs is illustrated in Fig. 2A. Parametric modulation analysis of the carry-over regressor showed significant positive correlations between physical difference and BOLD signal in bilateral middle temporal gyri (left: t(39) = 10.31, k = 875; right: t(39) = 9.66, k = 805; lateralization index (LI) = -.01), bilateral supplementary motor area (with peak maximum in the left: t(39) = 7.44, k = 166; LI = –.06), bilateral precuneus (with peak maximum in the right: t(39) = 6.91, k = 235; LI = .01), bilateral medial orbitofrontal gyrus (with peak maximum in the right: t(39) = 6.15, k = 46; LI = -.18), left postcentral gyrus (t(39) = 7.27, k = 301) and left superior frontal gyrus (t(39) = 6.94, k = 146). Subcortical clusters of activation in bilateral putamen (not illustrated; left: t(39) = 9.59, k = 279; right: t(39) = 6.83, k = 119; LI = .04) covers parts of amygdalae and hippocampi.

(B) Effect of individual differences on the magnitude of the repetition suppression effect

The results of the covariate analysis (ANCOVA) involving three regions of interest ((1) bilateral TVAs, (2) bilateral inferior frontal gyri/insulae and (3) bilateral inferior parietal lobule) are illustrated in Fig. 2B. We found negative correlations, as illustrated by the scatterplots (blue), between the amount of repetition suppression and the voice test score (BVMT) in the left inferior frontal gyrus (t(38) = 7.20, k = 297; correlation within 6mm sphere around peak cluster: r = -0.52; p < .001) and right supramarginal gyrus (t(38) = 5.77, k = 62; correlation within 6mm sphere around peak cluster: r = -0.65; p < .0001). In other words, participants with lower scores on the BVMT showed more repetition suppression in these areas. No significant activations were observed in bilateral TVAs.

An identical ANCOVA with the general auditory ability test (PROMS) as covariate showed no significant clusters in any of the regions of interest (scatterplots in purple of Fig. 2B are illustrated for the same coordinates as the voice test).

(C) Regions showing sensitivity to a change in perceived identity

Results of this analysis are illustrated in Fig. 2C. Here we contrasted pairs of stimuli with a 30% change in the acoustic signal that meant a change in the identity of the speaker (i.e., across identity trials) against pairs with a 30% change in acoustics that were perceived as the same identity (i.e., within identity trials). This contrast revealed regions that are sensitive to identity change while keeping acoustic change constant. Significant activation was found in medial frontal gyrus (t(39) = 5.26, k = 222) and a cluster covering right insula (t(39) = 4.87, k = 125) and IFG. Activation in the left insula/IFG was marginally significant (-30 23 2; t(39) = 5.65, k = 80; p = .065). LI for bilateral insulae/IFGs was -.02.

## 4. Discussion

Our aim was to disentangle regions of the voice identity network that process acoustic change from regions that deal with changes in the

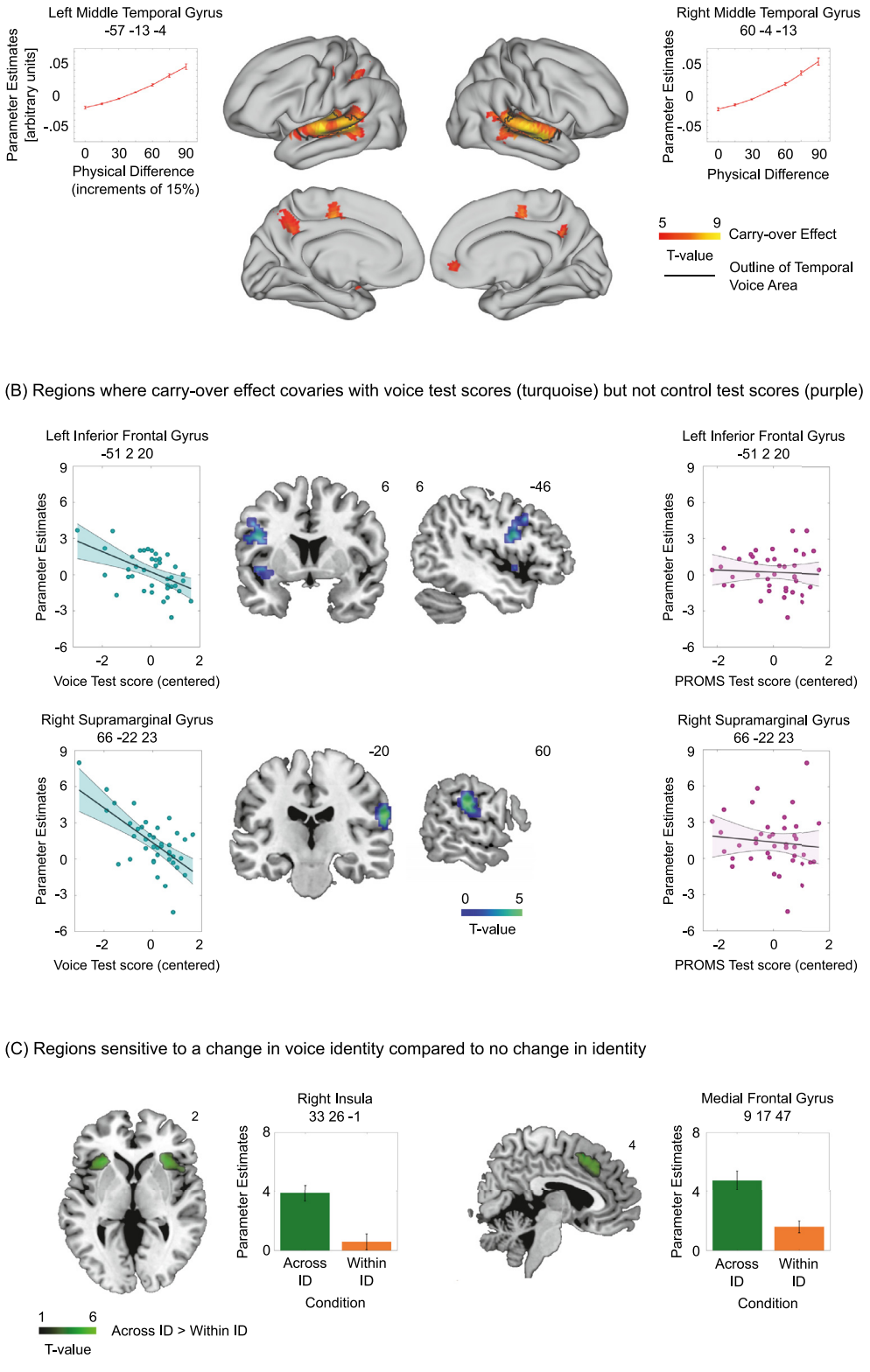


**Fig. 2.** Evidence for regional dissociation of responses to physical changes and changes in the representation of identity. (A) Results of a whole brain analysis showing activation maps of significant correlations between the physical difference between consecutive stimuli and BOLD signal illustrated on an inflated cortical surface. Positive correlations (i.e., increased repetition suppression for more similar consecutive stimuli) are evident in areas that overlap with bilateral voice-sensitive areas (shown with black outline along the bilateral upper banks of the superior temporal gyri). Plots of parameter estimates (arbitrary units) are illustrated and extracted from a 6mm sphere around each of four peak clusters in the bilateral temporal voice areas, precuneus and right orbitofrontal cortex. (B) Illustration of areas in the extended voice perception network where amount of repetition suppression based on physical difference between morphs covaries with test scores on an independent and standardised voice assessment (ANCOVA; region of interest analysis). Scatterplots and regression lines between estimates of BOLD signal and the voice test scores are shown in turquoise. Separate ANCOVA with general auditory ability score (PROMS) did not reveal any significant clusters in the regions of interest (scatterplots in purple). Shaded area represents 95% confidence intervals. (C) Results of a separate whole brain analysis showing regions that are sensitive to a change in identity (across ID > within ID trials; green). Parameter estimates for "Across ID" and "Within ID" trials are illustrated for significant clusters. Results are overlaid on a T1 averaged template for Fig. 2B and 2C. Error bars in Fig. 2A and 2C represent SEM.

cognitive representation of voice identity. To this end, we created continua consisting of seven morph steps with equal physical distances between utterances of two familiar speakers. We examined correlations between MRI signal and physical distance between morph steps and found that MRI signal linearly decreased with linearly increasing physical similarity between two consecutive stimuli. This repetition suppression effect was particularly strong in bilateral temporal voice-sensitive areas (TVAs), precuneus and putamen. This effect was mediated by an individual's voice perception performance on an independent voice assessment performed outside the scanner. Individuals with better voice matching

skills showed less repetition suppression than individuals with lower scores. In other words, individuals with greater sensitivity to physical differences between voices showed less repetition suppression. This covariation was specific to individual differences in voice identity discrimination as measured by the voice matching test, rather than differences in our participants' general auditory skills (as measured with a control assessment). This result highlights the important functional role of repetition suppression in voice expertise and further suggests that human voice perception is a highly specialised auditory ability. Our balanced stimulus sequence also allowed us to explore the activation patterns in response to changes in perceived identity while controlling for acoustic distance. Here we contrasted pairs of stimuli that differed physically by 30%. This 30% in physical difference could either reach across the category boundary and result in a change in perceived identity or it could remain within the category boundary and be perceived as the same identity. In response to a change in perceived speaker identity, while the acoustic change was controlled, we found effects in bilateral insulae and medial frontal gyrus. Taken together, our findings support a multi-step process of speaker identification, whereby the bilateral TVAs may deal with "structural" encoding of the voice and bilateral insulae and medial frontal gyrus process the higher-level representation of identity.

Using a carry-over design, we have shown that voice identity perception of familiar voices involves separable stages that do not engage anatomically overlapping regions. One stage involves general acoustic processing, and another stage deals with the high-level, cognitive representation of voice identity. This finding is line with neuroscientific and cognitive models of voice perception that predict low-level acoustic features to be handled by belt and parabelt of auditory cortex while higher-level representations of identity are dealt with separately (Belin et al. (2004); see also Schirmer and Kotz (2006) for a model on speech prosody). Schirmer and Kotz (2006) propose a multi-step model of speech prosody, another prominent paralinguistic aspect of the voice. In this case, higher-level representations of vocal emotion are processed by right inferior frontal regions. Our findings support and extend this model by reporting that additional regions are involved in the processing of acoustic information including a large cluster along the superior temporal gyri ranging towards the temporal pole and involving subcortical structures (e.g., cluster covering putamen and hippocampus/amygdala).

However, other neuroimaging studies on voice identity have not shown as clear a separation between processing stages as our study. Both Andics et al. (2010) and Latinus et al. (2011) report at least some overlap in regions dealing with the acoustic-based and higher-level identity representation of voices. These two studies diverge on the location of the overlap between acoustic and identity-based representations. In Andics et al. (2010), bilateral STS showed sensitivity to both acoustic and identity processing while Latinus et al. (2011) showed minor overlap between both stages in the right posterior inferior frontal cortex. All studies differ in how participants learned the voices (e.g., bimodally or just the voice). Our study is the only one with *personally* familiar speakers and therefore reports areas based on naturally acquired identity representations. While both familiar and unfamiliar voice recognition relies on the extraction of identity-specific features, the underlying mechanisms for unfamiliar and familiar speaker recognition are thought to diverge (e.g., Van Lancker and Kreiman (1987); Kreiman and Sidtis (2013); Maguinness, Roswandowitz, von Kriegstein (2018)). Our use of familiar speakers was also accompanied by better speaker recognition scores. Taken together, these differences may have contributed to us observing a clear separation of these processing stages.

Our finding is in line with previous research on voice emotion (Bestelmeyer et al., 2014) and gender (Charest et al., 2013) perception whereby the higher-level representation of emotion or gender involves very similar regions as reported in the current study. This involvement of the insulae and inferior frontal gyri in higher-level representations of identity, emotion and gender is apparent despite differences in how the contrast was computed. We employed a similar rationale, in terms of the morphing technology and analysis strategy, to Rotshtein et al. (2005) who showed a dissociation of physical and cognitive representations of face identity. While the in-scanner task and stimulus presentation were different to ours, Rotshtein et al. also report the involvement of bilateral IFGs for identity switch trials. These areas may be part of a supramodal, or modality independent, network for the abstract representation of person identity. That said, detailed cytoarchitectonic mapping of the human anterior insula is still missing and the precise role of the insula in our study and the aforementioned studies cannot be ascertained. We know that the insula is an integration hub which receives sensory input from all modalities (Nieuwenhuys, 2012). There is also neuroimaging evidence for considerable functional heterogeneity of the anterior insula with functions ranging from sensory and affective processing to high-level cognition (e.g., Alain et al., 2018). This heterogeneity includes a response to a wide range of tasks and stimulus types (e.g., emotional awareness, error awareness, attention to pain) but without consistent coactivations of other areas (Craig, 2009). It is therefore plausible that the anterior insulae along with the medial superior frontal region support person identity recognition via domain-general mechanisms and may respond whenever a categorical response to a stimulus is required. In fact, some researchers propose that the anterior insulae are a correlate of stimulus awareness or consciousness (Craig, 2009).We report strong repetition suppression effects in response to acoustic similarities in areas that overlapped with the TVAs, but we also found repetition suppression effects in areas outside the core voice perception network. These regions were right medial orbitofrontal cortex, bilateral precuneus and putamen. We know that the putamen is extensively and reciprocally connected to the superior temporal gyrus (Yeterian and Pandya, 1998; Cho et al., 2013) and is thought to play a role in sequencing, not only of motor behaviour but also of temporal changes in sound (Kotz et al., 2009; Geiser et al., 2012). The putamen seems to be an important structure for the extraction of regular auditory patterns (Kotz et al., 2009) and has been shown to be active in speech sound categorisation tasks (e.g., Feng et al., 2019). Adaptation in subcortical areas, such as the putamen, may be due to top-down feedback from higher-level areas (Friston, 2012). The activation in putamen and precuneus, as reported here and in Bestelmeyer et al. (2014), points to the existence of an extended network of both cortical and subcortical structures being involved in the processing of acoustic features in voices.

The degree of repetition suppression showed a similar pattern across the aforementioned brain regions and depended on the degree of similarity between two consecutive stimuli. Repetitions of identical stimuli caused the largest amount of suppression with increasingly dissimilar morphs resulting in increasing release from adaptation. This physical difference contrast covaried significantly with voice perception skill, obtained with an independent voice assessment, in left inferior frontal gyrus and right supramarginal gyrus. This coupling of behavioural results and functional activation patterns has not been reported before in an individual differences approach with neurotypical participants, using a standardised voice test (Mühl et al., 2018). Here participants who were better at discriminating between unfamiliar voices showed less repetition suppression to voice identity. In other words, participants who were more sensitive to subtle acoustic changes between voices showed overall less repetition suppression. Thus, the voice sensitive neurons of poorer voice recognisers may be less selective or tuned more broadly, and thereby showing increased repetition suppression to similar voices. Our finding is in line with Goh et al. (2010) who also reported a link between neural selectivity as approximated by fMRI adaptation and behavioural face discrimination performance in the right fusiform face area. This link is predicted by a computational model of face processing (Jiang et al., 2006; see also Jiang et al., 2013) and underlines the functional importance of adaptation.

Schirmer and Kotz's (2006) model does not directly predict an effect of individual differences in voice or general auditory perception ability on prosody perception. It does however acknowledge the likely contri-

bution of additional factors (e.g., context, individual significance) on all three stages of the processing hierarchy. Our study hypothesised and supports an influence of individual differences in voice discrimination ability at the lower end of the processing hierarchy. Whether individual differences affect the model's "cognition stage" remains to be tested. While the model incorporates these, potentially complex and heterogeneous, contributions they have not yet been extensively studied.

The core idea of face and voice perception models is that person identity is processed along a hierarchical pathway (e.g., Belin et al., 2004). Our results are in line with this notion. We found that the physical differences of a voice are processed in separate areas from the higher-level representation of voice identity in an analogous way to that reported in the face literature (e.g., Winston et al., 2004; Rotshtein et al., 2005). We also showed that individual differences in voice perception ability, as measured by an independent task to that performed in the scanner, is linked to the level of repetition suppression to physical differences in voices. Individuals with better voice discrimination scores on an independent voice assessment showed less repetition suppression. These results highlight an important functional role of adaptive coding in voice expertise.

## Ethics

The study protocol was vetted and approved by the local ethics committee at the Bangor Imaging Unit, Bangor University.

## Data availability

Processed EPI data will be publicly available via Mendeley along with the behavioural data to allow reproducibility of the results reported in the article.

## Data Availability

Data will be made available on request.

## Credit authorship contribution statement

**Patricia E.G. Bestelmeyer:** Conceptualization, Supervision, Methodology, Formal analysis, Visualization, Writing – original draft. **Constanze Mühl:** Conceptualization, Data curation, Formal analysis, Writing – review & editing.

## Acknowledgement

Constanze Mühl was funded by a PhD studentship sponsored by the School of Psychology at Bangor University.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119647.

## References

Aglieri, V., Cagna, B., Velly, L., Takerkart, S., Belin, P., 2021. FMRI-based identity classification accuracy in left temporal and frontal regions predicts speaker recognition performance. Sci. Rep. 11.
Aglieri, V., Chaminade, T., Takerkart, S., Belin, P., 2018. Functional connectivity within the voice perception network and its behavioural relevance. Neuroimage 183, 356–365.
Aguirre, G.K., 2007. Continuous carry-over designs for fMRI. Neuroimage 35, 1480–1494.
Agus, T.R., Paquette, S., Suied, C., Pressnitzer, D., Belin, P., 2017. Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci. Rep. 7.
Alain, C., Du, Y., Bernstein, L.J., Barten, T., Banai, K., 2018. Listening under difficult conditions: an activation likelihood estimation meta-analysis. Hum. Brain Mapp..
Andics, A., McQueen, J.M., Petersson, K.M., Gal, V., Rudas, G., Vidnyanszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage 52, 1528–1540.
Assal, G., Aubert, C., 1979. La-reconnaissance des onomatopées et des cris d'animaux lors de lésions focalisées du cortex cérébral. Rev. Neurol. (Paris) 135, 65–73.
Beauchemin, M., Gonzalez-Frankenberger, B., Tremblay, J., Vannasing, P., Martinez–Montes, E., Belin, P., Beland, R., Francoeur, D., Carceller, A.-M., Wallois, F., Lassonde, M., 2011. Mother and stranger: an electrophysiological study of voice processing in newborns. Cereb. Cortex 21, 1705–1711.
Belin, P., 2006. Voice processing in human and non-human primates. Philos. Trans. R. Soc. B-Biol. Sci. 361, 2091–2107.
Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. 8, 129–135.
Belin, P., Zatorre, RJ., 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.
Benton, A.L, Van Allen, M.W., 1972. Prosopagnosia and facial discrimination. J. Neurol. Sci. 15, 167–172.
Bestelmeyer, P.E.G., Belin, P., Grosbras, M.-H., 2011. Right temporal TMS impairs voice detection. Curr. Biol. 21, R838–R839.
Bestelmeyer, P.E.G., Latinus, M., Bruckert, L., Crabbe, F., Belin, P., 2012. Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. Cereb. Cortex 22, 1263–1270.
Bestelmeyer, P.E.G., Maurage, P., Rouger, J., Latinus, M., Belin, P., 2014. Adaptation to vocal expressions reveals multistep perception of auditory emotion. J. Neurosci. 34, 8098–8105.
Bestelmeyer, P.E.G., Mühl, C., 2021. Individual differences in voice adaptability are specifically linked to voice perception skill. Cognition 210.
Bodamer, J., 1947. Die Prosop-Agnosie; die Agnosie des Physiognomieerkennens. Archiv für Psychiatrie und Nervenkrankheiten 118, 6–53.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.
Büchel, C., Holmes, A.P., Rees, G., Friston, K.J., 1998. Characterizing stimulus–response functions using nonlinear regressors in parametric fMRI experiments. Neuroimage 8, 140–148.
Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2013. Cerebral porcessing of voice gender studied using a continuous carryover fMRI design. Cereb. Cortex 23, 116–120.
Cho, Y.T., Ernst, M., Fudge, JL., 2013. Cortico-Amygdala-Striatal Circuits Are Organized as Hierarchical Subsystems through the Primate Amygdala. J. Neurosci. 33, 14017–14030.
Clifford, C.W.G., Webster, M.A., Stanley, G.B., Stocker, A.A., Kohn, A., Sharpee, T.O., Schwartz, O., 2007. Visual adaptation: neural, psychological and computational aspects. Vision Res. 47, 3125–3131.
Craig, A.D., 2009. How do you feel—now? The anterior insula and human awareness. Nat. Rev. Neurosci. 10, 59–70.
De Renzi, E., Perani, D., Carlesimo, G.A., Silveri, M.C., Fazio, F., 1994. Prosopagnosia can be associated with damage confined to the right hemisphere–an MRI and PET study and a review of the literature. Neuropsychologia 32, 893–902.
De Renzi, E., Spinnler, H., 1966. Facial recognition in brain-damaged patients. An experimental approach. Neurology 16, 145–152.
Dennett, H.W., McKone, E., Edwards, M., Susilo, T., 2012. Face aftereffects predict individual differences in face recognition ability. Psychol. Sci. 23, 1279–1287.
Engfors, L.M., Jeffery, L., Gignac, G.E., Palermo, R., 2016. Individual differences in adaptive norm-based coding and holistic coding are associated yet each contributes uniquely to unfamiliar face recognition ability. J. Exp. Psychol. Hum. Percept. Perform. 43, 281–293.
Feng, G., Yi, H.G., Chandrasekaran, B., 2019. The role of the human auditory corticostriatal network in speech learning. Cereb. Cortex 29, 4077–4089.
Friston, K., 2012. Predictive coding, precision and synchrony. Cogn. Neurosci. 3, 238–239.
Geiser, E., Notter, M., Gabrieli, J.D.E., 2012. A corticostriatal neural system enhances auditory perception through temporal context processing. J. Neurosci. 32, 6177–6182.
Goh, J.O., Suzuki, A., Park, D.C., 2010. Reduced neural selectivity increases fMRI adaptation with age during face discrimination. Neuroimage 51, 336–344.
Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., Malach, R., 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24, 187–203.
Hecaen, H., de Ajuriaguerra, J., Magis, C., Angelergues, R., 1952. Le problème de l'agnosie des physionomies. Encephale 41, 322–355.
Hermann, P., Grotheer, M., Kovacs, G., Vidnyanszky, Z., 2017. The relationship between repetition suppression and face perception. Brain Imaging Behav. 11, 1018–1028.
Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K., Kojima, S., Nakamura, K., 1997. Vocal identification of speaker and emotion activates different brain regions. Neuroreport 8, 2809–2812.
Jiang, X., Bollich, A., Cox, P., Hyder, E., James, J., Gowani, S.A., Hadjikhani, N., Blanz, V., Manoach, D.S., Barton, J.J.S., Gaillard, W.D., Riesenhuber, M., 2013. A quantitative link between face discrimination deficits and neuronal selectivity for faces in autism. Neuroimage Clin. 2, 320–331.
Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., Riesenhuber, M., 2006. Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. Neuron 50, 159–172.
Johnstone, L.T., Karlsson, E.M., Carey, DP., 2020. The validity and reliability of quantifying hemispheric specialisation using fMRI: Evidence from left and right handers on three different cerebral asymmetries. Neuropsychologia 138, 107331.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. ICASSP Las Vegas 3933–3936.

Kleiner, M., Brainard, D., Pelli, D. 2007. What's new in Psychtoolbox-3? (Paper presented at the ECVP).
Kotz, S.A., Schwartze, M., Schmidt-Kassow, M., 2009. Non-motor basal ganglia functions: a review and proposal for a model of sensory predictability in auditory language perception. Cortex 45, 982–990.
Kreiman, J., Sidtis, D., 2013. Foundations of voice studies: An interdisciplinary approach to voice production and perception. Wiley-Blackwell, Malden, MA.
Kriegstein, K.V., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948–955.
Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Front Psychol 2, 175 -175.
Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb. Cortex 21, 2820–2828.
Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P., 2013. Norm-based coding of voice identity in human auditory cortex. Curr. Biol. 23, 1075–1080.
Lavan, N., Knight, S., Hazan, V., McGettigan, C., 2019. The effects of high variability training on voice identity learning. Cognition 193, 104026.
Law, L.N.C., Zentner, M., 2012. Assessing musical abilities objectively: construction and validation of the profile of music perception skills. PLoS One 7.
Lewis, J.W., Talkington, W.J., Walker, N.A., Spirou, G.A., Jajosky, A., Frum, C., Brefczynski-Lewis, J.A., 2009. Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J. Neurosci. 29, 2283–2296.
Luzzi, S., Coccia, M., Polonara, G., Reverberi, C., Ceravolo, G., Silvestrini, M., Fringuelli, F., Baldinelli, S., Provinciali, L., Gainotti, G., 2018. Selective associative phonagnosia after right anterior temporal stroke. Neuropsychologia 116, 154–161.
Maguinness, C., Roswandowitz, C., von Kriegstein, K., 2018. Understanding the mechanisms of familiar voice-identity recognition in the human brain. Neuropsychologia 116, 179–193.
McGettigan, C., Eisner, F., Agnew, Z.K., Manly, T., Wisbey, D., Scott, S.K., 2013. T'ain't what you say, it's the way that you say it - left insula and inferior frontal cortex work in interaction with superior temporal regions to control the performance of vocal impersonations. J. Cogn. Neurosci. 25, 1875–1886.
Mühl, C., Sheil, O., Jarutyte, L., Bestelmeyer, PEG., 2018. The Bangor voice matching test: a standardized test for the assessment of voice perception ability. Behav. Res. Methods 50, 2184–2192.
Neuner, F., Schweinberger, S.R., 2000. Neuropsychological impairments in the recognition of faces, voices, and personal names. Brain Cogn. 44, 342–366.
Nieuwenhuys, R., 2012. The insular cortex: a review. Prog. Brain Res. 195, 123–163.
Nonyane, B.A.S., Theobald, C.M., 2007. Design sequences for sensory studies: achieving balance for carry-over and position effects. Br. J. Math. Stat. Psychol. 60, 339–349.
Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E.G., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 119, 164–174.
Rhodes, G., Jeffery, L., Taylor, L., Hayward, W.G., Ewing, L., 2014. Individual differences in adaptive coding of face identity are linked to individual differences in face recognition ability. J. Exp. Psychol.-Hum. Perception Performance 40, 897–903.
Rhodes, G., Pond, S., Burton, N., Kloth, N., Jeffery, L., Bell, J., Ewing, L., Calder, A.J., Palermo, R., 2015. How distinct is the coding of face identity and expression? Evidence for some common dimensions in face space. Cognition 142, 123–137.
Roswandowitz, C., Kappes, C., Obrig, H., von Kriegstein, K., 2018. Obligatory and facultative brain regions for voice-identity recognition. Brain 141, 234–247.
Rotshtein, P., Henson, R.N.A., Treves, A., Driver, J., Dolan, RJ., 2005. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat. Neurosci. 8, 107–113.
Schall, S., Kiebel, S.J., Maess, B., von Kriegstein, K., 2015. Voice identity recognition: functional division of the right STS and its behavioral relevance. J. Cogn. Neurosci. 27, 280–291.
Schirmer, A., Kotz, SA., 2006. Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. Trends Cogn. Sci. 10, 24–30.
Schweinberger, S.R., Kawahara, H., Simpson, A.P., Skuk, V.G., Zäske, R., 2014. Speaker perception. Wiley Interdiscip. Rev. Cogn. Sci. 5, 15–25.
Van Lancker, D., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain Cogn. 1, 185–195.
Van Lancker, D., Kreiman, J., 1987. Voice discrimination and recognition are separate abilities. Neuropsychologia 25, 829–834.
von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, AL., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res. Cogn. Brain Res. 17, 48–55.
Wainwright, M.J., 1999. Visual adaptation as optimal information transmission. Vision Res. 39, 3960–3974.
Wark, B., Lundstrom, B.N., Fairhall, A., 2007. Sensory adaptation. Curr. Opin. Neurobiol. 17, 423–429.
Watson, R., Latinus, M., Bestelmeyer, P.E., Crabbe, F., Belin, P., 2012. Sound-induced activity in voice-sensitive cortex predicts voice memory ability. Front. Psychol. 3.
Warrington, E.K., James, M., 1967. An experimental investigation of facial recognition in patients with unilateral cerebral lesions. Cortex 3, 317–326.
Webster, M.A., 2011. Adaptation and visual coding. J. Vis. 11.
Wilke, M., Lidzba, K., 2007. LI-tool: a new toolbox to assess lateralization in functional MR-data. J. Neurosci. Methods 163, 128–136.
Wilke, M., Schmithorst, V.J., 2006. A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data. Neuroimage 33, 522–530.
Winston, J.S., Henson, R.N.A., Fine-Goulden, M.R., Dolan, R.J., 2004. fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J. Neurophysiol. 92, 1830–1839.
Yeterian, E.H., Pandya, D.N., 1998. Corticostriatal connections of the superior temporal region in rhesus monkeys. J. Comp. Neurol. 399, 384–402.
Yin, R., 1970. Face recognition by brain-injured patients: a dissociable ability? Neuropsychologia 8, 395–402.
Zäske, R., Schweinberger, S.R., Kawahara, H., 2010. Voice aftereffects of adaptation to speaker identity. Hear. Res. 268, 38–45.
Zäske, R., Hasan, A.S.B., Belin, P., 2017. It doesn't matter what you say: fMRI correlates of voice learning and recognition independent of speech content. Cortex 94, 100–112.