NeuroImage 201 (2019) 116004

Contents lists available at ScienceDirect

NeuroImage

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

Faces and voices in the brain: A modality-general person-identity representation in superior temporal sulcus

Maria Tsantani [a,**], Nikolaus Kriegeskorte [b], Carolyn McGettigan [c], Lúcia Garrido [a,*]

[a] Division of Psychology, Department of Life Sciences, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, UK
[b] Zuckerman Mind Brain Behavior Institute, Columbia University, 3227 Broadway, New York, NY, 10027, USA
[c] Speech Hearing and Phonetic Sciences, University College London, 2 Wakefield St, Kings Cross, London, WC1N 1PJ, UK

ARTICLE INFO

Keywords:
Face recognition
Multisensory processing
Person-identity recognition
Representational similarity analysis
Voice recognition

ABSTRACT

Face-selective and voice-selective brain regions have been shown to represent face-identity and voice-identity, respectively. Here we investigated whether there are modality-general person-identity representations in the brain that can be driven by either a face or a voice, and that invariantly represent naturalistically varying face videos and voice recordings of the same identity. Models of face and voice integration suggest that such representations could exist in multimodal brain regions, and in unimodal regions via direct coupling between face- and voice-selective regions. Therefore, in this study we used fMRI to measure brain activity patterns elicited by the faces and voices of familiar people in face-selective, voice-selective, and person-selective multimodal brain regions. We used representational similarity analysis to (1) compare representational geometries (i.e. representational dissimilarity matrices) of face- and voice-elicited identities, and to (2) investigate the degree to which pattern discriminants for pairs of identities generalise from one modality to the other. We did not find any evidence of similar representational geometries across modalities in any of our regions of interest. However, our results showed that pattern discriminants that were trained to discriminate pairs of identities from their faces could also discriminate the respective voices (and vice-versa) in the right posterior superior temporal sulcus (rpSTS). Our findings suggest that the rpSTS is a person-selective multimodal region that shows a modality-general person-identity representation and integrates face and voice identity information.

1. Introduction

Looking at a familiar person's face or listening to their voice automatically grants us access to a wealth of information regarding the person's identity, such as their name, our relationship to them, and memories of previous encounters. Knowledge about how the brain processes faces and voices separately has advanced significantly over the past twenty years: functional magnetic resonance imaging (fMRI) revealed cortical regions that are face-selective (Kanwisher et al., 1997; McCarthy et al., 1997) and regions that are voice-selective (Belin et al., 2000). Recent advances using multivariate classification methods have further shown that some of these regions are important for identification. In particular, face-selective regions in the posterior occipitotemporal lobe, anterior temporal lobe, and posterior superior temporal sulcus (pSTS) can discriminate different face images (Kriegeskorte et al., 2007; Nestor et al., 2011; Goesaert and Op de Beeck, 2013; Verosky et al., 2013; Axelrod and Yovel, 2015; Collins et al., 2016; Visconti Di Oleggio Castello et al., 2017). Crucially, a number of studies also found representations in these regions that generalised across different images of the same person (Anzelotti et al., 2014; Anzellotti and Caramazza, 2016; Guntupalli et al., 2017), i.e., were able to "tell people together" (Jenkins et al., 2011; Burton, 2013). Similarly for voices, Formisano et al. (2008) found voice-identity representations in the right STS and Heschl's gyrus that could both discriminate between speakers and generalise across different vowel sounds spoken by the same voice.

Despite these advances, we still have a limited understanding of how the brain combines and integrates face and voice information. Previous work has proposed two different models for face and voice integration (Campanella and Belin, 2007; Blank et al., 2011; Yovel & O'Toole, 2016). According to the first model, there are multimodal brain regions that process information about people and receive input from both face- and voice-responsive regions (Ellis et al., 1997; Campanella and Belin, 2007).

* Corresponding author.
** Corresponding author.
E-mail addresses: maria.tsantani@gmail.com (M. Tsantani), garridolucia@gmail.com (L. Garrido).

https://doi.org/10.1016/j.neuroimage.2019.07.017
Received 5 November 2018; Received in revised form 17 May 2019; Accepted 7 July 2019
Available online 9 July 2019

Patient (e.g. Ellis et al., 1989; Gainotti, 2011; Hailstone et al., 2011) and fMRI studies (e.g. Shah et al., 2001; Joassin et al., 2011; Hölig et al., 2017) suggest the anterior temporal lobe, the posterior cingulate cortex, the angular gyrus, the STS, and the hippocampus as candidate multimodal regions (for a meta-analysis, please see Blank et al., 2014). According to the second model, face and voice information is also integrated via the direct coupling between face- and voice-responsive brain regions. In support of this, fMRI studies have shown that voice recognition of familiar (or recently learned) people is associated with increased activation in face-responsive regions of the fusiform gyrus (von Kriegstein et al., 2005, 2006; 2008; von Kriegstein and Giraud, 2006). Some studies have also shown functional and structural connectivity between face-responsive and voice-responsive regions (e.g. von Kriegstein et al., 2005, 2006; Blank et al., 2011). Crucially, the integration mechanisms proposed by the two models are not mutually exclusive.

Based on these two proposed mechanisms, in this study we investigated whether there are modality-general person-identity representations in person-selective multimodal regions and/or in face- and voice-selective regions. We measured fMRI activation patterns in response to the faces and to the voices of 12 famous individuals and then used representational similarity analysis — RSA (Kriegeskorte et al., 2008a, 2008b) to identify regions with modality-general person-identity representations. RSA allows us to abstract from the units of measurement, and thus seems ideally suited to compare brain representations across different sensory modalities. Our first analysis compared the representational geometries of face- and voice-elicited brain response patterns, in which representational geometry refers to how the brain response patterns corresponding to the different identities are related to each other (Kriegeskorte and Kievit, 2013). In other words, for each brain region, we compared the dissimilarities between the brain patterns measured in response to the face-identities with the dissimilarities of the brain patterns measured in response to the corresponding voice-identities. We expected that if a region showed a modality-general person-identity representation, the representational geometries of face and voice identities would be highly correlated in this region. Our second analysis investigated the degree to which pattern discriminants for pairs of identities generalised from one modality to the other. In other words, we used a linear discriminant computed in one modality to test discriminability of the same pair of identities in the other modality, in a similar way to traditional pattern classification methods (Nili et al., 2014; Walther et al., 2016; Carlin and Kriegeskorte, 2017). We expected that if a region showed a modality-general person-identity representation, the pattern discriminants would generalise across faces and voices.

Two recent studies (using methods similar to the ones we employed in our second analysis) showed that multimodal regions in the STS and inferior frontal gyrus (Hasan et al., 2016; Anzellotti and Caramazza, 2017) could discriminate between the activation patterns of two face-identities based on voice information (and vice-versa). However, these studies presented very few identities (4 and 3 identities) and a limited number of face and voice tokens per identity (1 and 2 tokens). We think it is thus possible that the observed crossmodal decoding could be due to learned associations between specific face images and voice recordings, rather than decoding of identity per se (Lavan, 2017). Therefore, in our study, we included 12 different familiar identities and multiple, naturalistically varying face videos and voice recordings for each identity. While behavioural studies have shown the importance of within-person variability for familiar face and voice recognition (Jenkins et al., 2011; Burton, 2013; Burton et al., 2016; Lavan et al., 2018a, 2018b), this is rarely taken into account in neuroimaging experiments, which typically use highly similar or artificial stimuli for the same person. In contrast, we aimed to sample the variability of visual and auditory appearance that we are exposed to in everyday life, in order to better capture processes of person identification, which are distinct from image or sound recognition (Burton, 2013; Jenkins et al., 2011; Lavan et al., 2018a, 2018b).

2. Materials and methods

2.1. Overview of study

The study consisted of two MRI scanning sessions that took place on separate days, with each session taking approximately 90 min. In each MRI session, participants completed three functional runs (main experimental runs) in which they viewed the faces and listened to the voices of 12 famous people in an event-related design (Fig. 1). In addition, participants underwent two structural scans (one in each session) and functional localisers to independently define face-selective, voice-selective, and multimodal ROIs. The same participants also completed a behavioural testing session, which took place after the scanning sessions; however, these results are not included here.

From the experimental runs, we computed fMRI activation patterns in response to the faces and to the voices of the 12 famous individuals. It was important to use highly familiar individuals because we needed to guarantee that participants were well acquainted with the faces and voices of those individuals. Therefore, all participants were able to recognise at least 9 out of the 12 famous individuals used as stimuli, as demonstrated with a Recognition Task during recruitment, which was repeated in the first session of testing (please see Supplementary Information 1 — SI-1). Before entering the scanner at the start of the first MRI session, participants also completed a Familiarity Task in which they rated all face and voice stimuli on perceived familiarity (please see SI-2). This task also served the purpose of familiarising participants with all the stimuli that they would be presented with in the main experiment in the scanner, and participants were presented with the name of the person after responding to each face/voice.

To investigate the existence of modality-general person-identity representations in each of our ROIs, we used RSA (Kriegeskorte et al., 2008a, 2008b; Kriegeskorte and Kievit, 2013) to compare the representational geometry of face-identities with the representational geometry of voice-identities (Analysis A), and to investigate the degree to which pattern discriminants for each pair of identities generalise from one modality to the other (Analysis B). These two analyses complement each other and allowed us to test different predictions regarding the nature of modality-general person-identity representations. Specifically, Analysis A (RSA comparing representational geometries) is constrained by two assumptions regarding the nature of these representations. The first assumption is that there needs to be sufficient variability in the representational distances between different identities, i.e. different degrees of similarity between identities. If, however, all identities were equally distinct from each other, we could no longer expect to find correlations between geometries across the two modalities. The second assumption is that for any ROI representing modality-general information, it needs to primarily represent that type of information. If an ROI, however, also represents modality-specific information in addition to modality-general information, the representational geometry will be affected by information that will not be shared across modalities. This is particularly important if modality-general representations exist in multimodal regions, given that the voxels comprising the pattern estimates in these regions may contain both unimodal and multimodal neurons (Laurienti et al., 2005; Driver and Noesselt, 2008; Quiroga et al., 2009). In this case, the influence of modality-specific information on the representational distances between all identities could override the influence of modality-general information on the representational geometry, and could result in non-matching representational geometries across modalities.

To overcome these constraints with Analysis A, we also conducted Analysis B (RSA investigating identity discriminability), in which we expected that regions with modality-general person-identity representations would be able to discriminate between pairs of identities in one modality based on their representational distance in the other modality. This analysis focuses on one pair of identities at a time, and thus is not affected by the degree of variability in the representational distances

between all identities. In addition, this analysis is focused on pattern discriminants that generalise across modalities, and therefore we believe that it is more sensitive to detect modality-general person-identity representations even in the presence of modality-specific information.

2.2. Participants

Participants were recruited at Royal Holloway, University of London and Brunel University London to take part in a behavioural and fMRI experiment. All participants were required to be native English speakers aged between 18 and 30, and to have been resident in the UK for a minimum of 10 years. These requirements were set to increase the likelihood of participants being familiar with the famous people whose faces and voices were presented in the experiment. In addition, participants completed an online Recognition Task (please see SI-1) as part of the screening procedure for the study, and were only invited for the full study if they were able to recognise at least 9 out of 12 famous people from both their face and their voice (we repeated this task again in the presence of the experimenter in the first session of testing).

Thirty-one healthy right-handed adult participants were recruited who matched all the above criteria. One participant was excluded from the study after the first MRI session due to excessive head movement in the scanner (more than 3 mm in any direction within one run). The final sample consisted of 30 participants (eight men) with mean age of 21.2 years (SD = 2.37, range = 19–27). All reported normal or corrected-to-normal vision and normal hearing, provided written informed consent, and were reimbursed for their participation. The study was approved by the Ethics Committee of Brunel University London.

2.3. Main experimental runs: stimuli, design, and procedure

2.3.1. Stimuli

Six silent, non-speaking video clips of moving faces, and six sound clips of voices for each of the 12 famous people (six women, six men) were obtained from videos on YouTube (in total, 72 stimuli per modality). These people had been identified in our pilot studies as having highly recognisable faces and voices within samples of native English speakers between the ages of 18–30 who have been resident in the UK for a minimum of 10 years. Given that familiar voice recognition is more challenging than familiar face recognition (Damjanovic and Hanley, 2007; Hanley and Damjanovic, 2009; Hanley et al., 1998), it was particularly important to select famous people with highly recognisable voices. This list of famous people included actors, pop stars, politicians, comedians, and TV personalities: Alan Carr, Beyonce Knowles, Daniel Radcliffe, Emma Watson, Arnold Schwarzenegger, Barack Obama, Sharon Osbourne, Kylie Minogue, Graham Norton, Cheryl Cole, Barbara Windsor, and Jonathan Ross.

The face stimuli were selected so that the background did not provide any cues to the identity of the person. Other than the absence of speech, there were no constraints on the type of face movement. Examples of face movements included nodding, smiling, and rotating the head. However, all stimuli were selected to be primarily front-facing. Face stimuli were edited using Final Cut Pro X (Apple, Inc.) so that they were 3 s long and centred on the bridge of the nose. Six video-clips of the face of the same person were obtained from different original videos set in a different background.

Voice stimuli were edited using Audacity® 2.0.5 recording and editing software (RRID:SCR_007198) so that they contained 3 s of speech after removing long periods of silence. Voice stimuli were converted to mono with a sampling rate of 44100, low-pass filtered at 10 KHz, and root-mean-square (RMS) normalised using Praat (version 5.3.80; Boersma and Weenink, 2014; www.praat.org). Six sound clips of the voice of the same person were obtained from different original videos. All of the voice stimuli had a different verbal content and were non-overlapping. The stimuli were selected so that the speakers' identity could not be determined based on the verbal content, conforming to the standards set by Van Lancker et al. (1985) and Schweinberger et al. (1997).

2.3.2. Design and procedure

Face and voice stimuli were presented using the Psychophysics Toolbox (version 3; RRID:SCR_002881; Brainard, 1997; Pelli, 1997) via a computer interface inside the scanner. In an event-related design, face and voice clips of all 12 identities were intermixed within each run (Fig. 1). A fixation point was always present and participants were asked to fixate, which guaranteed more similar conditions between the face and voice trials. The videos were 640 × 360 pixels. The screen resolution was 1024 × 768 pixels, and from a distance of 85 cm, the videos subtended 20.83 × 12.27 degrees of visual angle. Audio stimuli were presented via MR-compatible earbuds (S14; Sensimetrics Corp.), which participants used for each entire scanning session. The six face videos and the six voice recordings for each of the 12 identities were evenly distributed among three experimental runs so that each run contained two different videos of the face and two different recordings of the voice of each identity. In other words, each of the three experimental runs presented two unique face tokens and two unique voice tokens. Each individual stimulus was presented twice within each run. Therefore, in each run there were 96 experimental trials (48 face trials, 48 voice trials) in total.

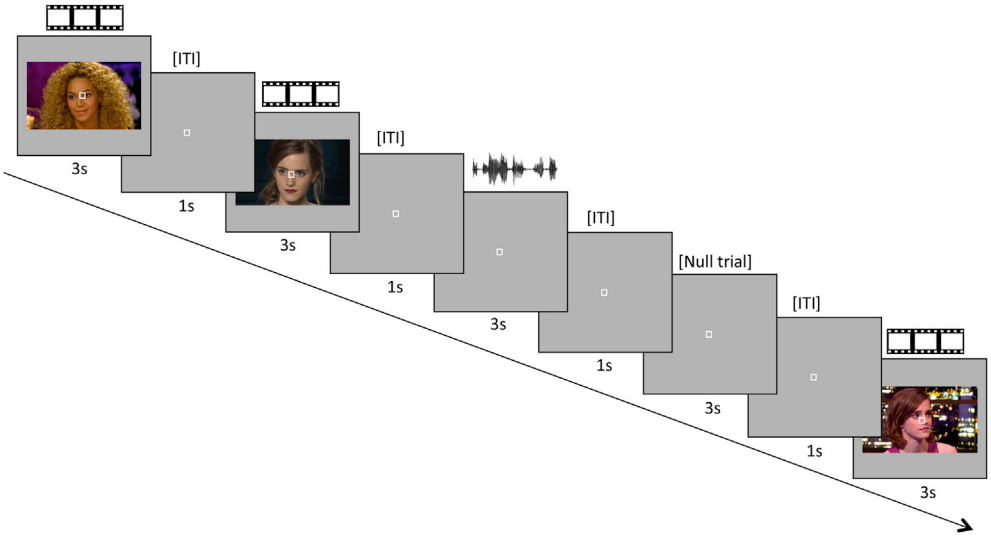


Fig. 1. Example trial sequence. Face videos, voice recordings, and null (fixation) trials were intermixed in the same run. Stimuli for each modality were sourced from different original videos on YouTube. All face videos had a different background, faces were mostly front-facing, and face movement was unconstrained, but did not feature any speech. Voice recordings included unconstrained natural speech, but verbal content did not reveal the identity of the speaker. The duration of face and voice stimuli was 3s, with a 1s inter-trial interval (ITI). Fixation points are enlarged for visualisation purposes.

The three runs were then repeated in a different order in the second session.

Participants performed an anomaly detection task that involved pressing a button when they saw or heard a novel famous person that was not part of the set of the 12 famous people that they had been familiarised with prior to entering the scanner. Therefore, each run also contained 12 task trials presenting six famous faces and six famous voices that were not part of the set of famous people that the participants had been familiarised with. An anomaly detection task was chosen to maintain attention to face and voice identity without confounding motor task responses and experimental trials.

Stimuli were presented in a pseudorandom order that ensured that, within each modality, each identity could not be preceded or succeeded by one of the other identities more than once, and that each stimulus could not be succeeded by a repetition of the exact same stimulus. Face and voice clips were presented for 3 s with a SOA of 4 s. Thirty-six null fixation trials were added to each run (~25% of the total number of trials). Thus, each run contained 144 trials in total and lasted approximately 10 min.

The presentation order of the three runs within each session was counterbalanced across participants. The same three runs with the same face videos and voice recordings that were presented in scanning session one were also presented in session two. However, the three runs were presented in different orders in both sessions (counterbalanced across participants) and stimuli within each run were presented in a new pseudorandom sequence. As an exception, the stimuli for the task trials were different in the two sessions in order to maintain their novelty.

2.4. Functional localiser runs: stimuli, design, and procedure

Across both sessions, participants completed at least one run (in most cases two) of (1) the temporal voice area (TVA) localiser (Belin et al., 2000), (2) a face localiser, (3) a person (face-voice) localiser, and (4) a voice localiser.

2.4.1. TVA localiser

We used the TVA localiser developed by Belin et al. (2000) which contains vocal and non-vocal auditory stimuli. Stimuli were presented in 40 blocks of 8 s each. Vocal stimuli were presented in 20 blocks and included speech and non-speech vocalisations obtained from 47 speakers (Pernet et al., 2015). Non-vocal stimuli were presented in 20 blocks and consisted of industrial sounds, environmental sounds, and animal vocalisations. Within each block, stimuli were presented in a random order that was fixed across participants. Participants were instructed to close their eyes and focus on the sounds. The TVA localiser was presented directly after the main experimental runs. The duration of a single run was approximately 10 min.

2.4.2. Face, voice, and person localisers

We created new face, person (face-voice), and voice localiser runs that shared the same experimental design and presented stimuli from comparable categories (people and objects/scenes). Importantly, we used videos and not static images of faces. Dynamic face stimuli have been shown to be more effective than static face stimuli for localising face-selective regions (Fox et al., 2009; Pitcher et al., 2011). Stimuli used for the face localiser were silent, non-speaking video clips of famous and non-famous (French celebrities unknown to our participants) moving faces, and silent video clips of moving large objects and natural or manmade visual scenes (such as videos of airplanes, trains, traffic, rainforests, waves on a beach) obtained from videos on YouTube. For the person localiser, the stimuli were audiovisual and included videos clips of the faces of famous and non-famous people speaking, and video clips of moving large objects and natural or manmade scenes (same categories as above). For the voice localiser, we presented voice clips of famous and non-famous people, and sound clips of manmade or natural environmental sounds (same categories as used in the other two types of localisers), with no video. The stimuli used in the localisers were different from the stimuli used in the experimental runs.

Videos (640 × 360 pixels) were presented at the centre of the screen. Each stimulus lasted 8 s and each run presented 48 stimuli in an event-related design. Stimuli were presented in pairs (24 pairs) showing the same person (such as two videos of Brad Pitt) or the same category of objects or scenes (such as two videos of trains). Eight pairs showed stimuli from famous people, eight pairs showed stimuli from non-famous people, and eight pairs showed object/scene stimuli. Participants were encouraged to always fixate at the centre of the screen. Participants performed a one-back task in which they had to detect the exact same stimulus repetition within each pair, which occurred in approximately 15% of the trials. A 16-s period of fixation was presented at the end of each run and twice in the middle of each run (every 16 trials).

The order of the face, voice, and person localisers was counterbalanced across participants. For participants who completed two runs of each localiser, different identities were presented on each run. The duration of each localiser run was approximately 8 min.

2.5. MRI data acquisition

Participants were scanned using a 3.0 T Tim Trio MRI scanner (Siemens, Erlangen) with a 32-channel head coil at the Combined Universities Brain Imaging Centre (CUBIC) at Royal Holloway, University of London. In each of the two scanning sessions, a whole-brain T1-weighted anatomical scan was acquired using magnetization-prepared rapid acquisition gradient echo (MPRAGE) [1.0 × 1.0 in-plane resolution; slice thickness, 1.0 mm; 176 axial interleaved slices; PAT, Factor 2; PAT mode, GRAPPA (Generalised Autocalibrating Partially Parallel Acquisitions); repetition time (TR), 1900 ms; echo time (TE), 3.03 ms; flip angle, 11°; matrix, 256 × 256; field of view (FOV), 256 mm].

For all functional runs, T2*-weighted whole-brain functional scans were acquired using echo-planar imaging (EPI) [3.0 × 3.0 in-plane resolution; slice thickness, 3.0 mm; PAT, Factor 2; PAT mode, GRAPPA (Generalised Autocalibrating Partially Parallel Acquisitions); 34 sequential (descending) slices; repetition time (TR), 2000 ms; echo time (TE), 30 ms; flip angle, 78°; matrix, 64 × 64; field of view (FOV), 192 mm]. For the majority of participants, slices covered all parts of the brain except for the most dorsal part of parietal cortex. In each experimental run we obtained 293 brain volumes, in the TVA localiser we obtained 251 brain volumes, and in each run of the face, voice, and person localiser runs we obtained 227 brain volumes.

2.6. fMRI data pre-processing and general linear models

Data were pre-processed using Statistical Parametric Mapping (SPM12; Wellcome Department of Imaging Science, London, UK; RRID:SCR_007037; http://www.fil.ion.ucl.ac.uk/spm) operating in Matlab (version R2013b; MathWorks; RRID:SCR_001622). Pre-processing was performed separately for each scanning session. All runs within each session (main experiment or localiser runs) were pre-processed together. The first three EPI images in each run (dummy scans) were discarded to allow for T1-equilibration effects. Images were slice-time corrected based on the middle slice in each volume and then realigned to correct for head movement based on the first image. The structural image in native space was then coregistered with the realigned mean functional image and segmented into grey matter, white matter, and cerebrospinal fluid. No smoothing was performed on the images from the experimental runs. Functional images from the localiser runs were smoothed with a 4-mm Gaussian kernel (full width at half maximum).

After separate pre-processing of the images in each session, images from the second scanning session were realigned to the structural image from the first session. Specifically, the structural image from session two was coregistered to the structural image from session one, and the transformation was then applied to all functional images from session

two. As a result, all functional images for each participant were in the same space.

For the analysis of data from both the main experimental runs and the functional localiser runs we computed mass univariate time-series models for each participant. Regressors modelled the BOLD response following the onset of the stimuli and were convolved with a canonical hemodynamic response function (HRF). We also used a high-pass filter cutoff of 128 s and autoregressive AR (1) model to account for serial correlations. Six head motion parameters computed during realignment were included as regressors of no interest.

2.7. Main experimental runs: data analysis

Models were defined separately for each scanning session and each experimental run (six runs in total). The 12 different identities in each modality were entered as separate regressors in the model (i.e. 24 regressors). Each of these regressors included the two different face videos and voice recordings of each identity that were presented in the run, as well as the two repetitions of each stimulus. Task trials were included as regressors of no interest. As part of the crossvalidation procedure used in the main analyses described below, separate models were estimated for each partition of each crossvalidation fold, thus resulting in parameter estimates and residual time courses for every possible independent partition. For partitions with two runs, data was concatenated before estimating the model. In the analyses described below, we used the beta estimates computed at each voxel of each ROI for each of the 24 experimental conditions (12 face-identities and 12 voice-identities).

2.7.1. Mean response to faces and voices in ROIs

We conducted an analysis to characterise the responses to the faces and voices presented in the main experimental runs in each ROI, and to confirm that each ROI showed the expected responsivity to faces and voices. For this analysis, we calculated the mean (across all voxels in each ROI, and across all runs) of the parameter estimates for the 12 face-identities and the mean of the parameter estimates for the 12 voice-identities. For each ROI, we tested whether the mean for faces and the mean for voices were significantly different from zero (across participants) using one-sample *t*-tests. P values were corrected for 24 comparisons (2 tests x 12 ROIs) controlling the false discovery rate (FDR), with $q < 0.05$. We also compared the mean for faces with the mean for voices in each ROI using paired *t*-tests. P values were corrected for multiple comparisons (12 comparisons) using FDR with $q < 0.05$.

2.7.2. Analysis A: RSA comparing representational geometries

For this analysis, we computed representational dissimilarity matrices (RDMs) for face-identities and for voice-identities (each RDM was 12×12) separately for each participant, each scanning session and each ROI. We then computed the correlations between pairs of these RDMs. These analyses were performed using in-house Matlab code and the RSA toolbox (Nili et al., 2014). To compute the RDMs we used the linear discriminant contrast (LDC), a crossvalidated distance measure (Nili et al., 2014; Walther et al., 2016). For each ROI, each modality (i.e. faces and voices separately), and each scanning session, we calculated the LDC between the pattern estimates (beta estimates across all voxels within an ROI) elicited by the different identities. The resulting 12×12 matrices were symmetric around a diagonal of zeros. Each cell in the RDMs showed the discriminability of the pattern estimates corresponding to a pair of identities in the chosen modality and ROI.

The procedure for calculation of the LDC is illustrated in Fig. 2. RDMs were computed using leave-one-run-out crossvalidation across the three runs in each session (each run presented the same identities with different stimuli). In each of three crossvalidation folds, the pattern estimates for each identity were computed with data from two runs (partition one) and separately from the pattern estimates from the remaining run (partition two). The pattern estimates from each pair of identities from partition one were used to obtain a linear discriminant, which was then applied to differentiate the activity patterns of the same identity pairs in partition two (Nili et al., 2014; Walther et al., 2016). We applied multivariate noise normalisation by computing a noise variance-covariance matrix based on the residual time courses obtained from the model that was estimated with data from partition one. More specifically, to compute the LDC for each pair of identities, we first multiplied the contrast between the patterns of a pair of identities in partition one (the discriminant weights) by the inverse of the noise variance-covariance matrix (after regularisation using the optimal shrinkage method: Ledoit and Wolf, 2004), and transformed the resulting weights to unit length. We then computed the dot product between the resulting vector and the vector with the contrast between the patterns of the same pair of identities from partition two (Carlin and Kriegeskorte, 2017), which resulted in a single value showing the discriminability of those identities. Under the null hypothesis, LDC values are symmetrically distributed around zero (Walther et al., 2016).

The resulting RDMs with LDC values from each crossvalidation fold were averaged to create one RDM per scanning session. This procedure resulted in four RDMs per participant per ROI: faces session 1, voices session 1, faces session 2, and voices session 2 (Fig. 5B). Crossvalidating across runs with different videos of the face and recordings of the voice of each identity (Fig. 2A) ensured that the resulting RDMs represented face- and voice-*identity*, rather than specific face videos and voice recordings.

In order to compare the representational geometries of the face- and voice-identities, the RDMs for each participant were compared across the two scanning sessions using Pearson's correlation coefficient (Fig. 5B). We also compared the representational geometries of face and voice-identities within modality across two scanning sessions in order to investigate the stability of the representational geometries across the two sessions. For the *crossmodal comparisons* we compared the face and voice RDMs from session one with the RDMs of the *other* modality in session two (i.e. faces session 1 vs. voices session 2, and voices session 1 vs. faces session 2). For the *unimodal comparisons* we compared the face and voice RDMs from session one with RDMs of the *same* modality in session two (i.e. faces session 1 vs. faces session 2 and voices session 1 vs. voices session 2). At the group level, for each ROI we compared the single-subject correlations for each of the four comparisons (two crossmodal, two unimodal) against zero using one-sample one-tailed Wilcoxon signed-rank tests (because correlations are not normally distributed). P values were corrected for multiple comparisons (48 comparisons: 4 tests x 12 ROIs) controlling for FDR with $q < 0.05$.

2.7.3. Analysis B: RSA investigating identity discriminability

For this analysis, we computed crossmodal RDMs separately for each participant, each scanning session and each ROI. We used the activity patterns of identity pairs in one modality to create a linear discriminant and then applied the discriminant to the activity patterns of the same identity pairs in the other modality (Fig. 2B). With this exception, the crossvalidation procedure was identical to the procedure for creating face and voice RDMs for the previous analysis. Two crossmodal RDMs for each ROI were computed using this method: one by applying a linear discriminant based on face data to voice data, and one by applying a linear discriminant based on voice data to face data. The LDC provides a continuous measure of discriminability for each pair of stimuli (Nili et al., 2014; Walther et al., 2016; Carlin and Kriegeskorte, 2017). Importantly, under the null hypothesis the LDC is symmetrically distributed around zero, and thus unbiased. By calculating the mean LDC value across all cells in an RDM for a certain ROI we can determine the overall ability of that ROI to discriminate between identities (Fig. 6B). Mean LDC values for all participants can then be subjected to random-effects inference comparing against zero. Therefore, we expected that crossmodal RDMs for regions with modality-general person-identity representations would show mean LDC values that are significantly greater than zero.

In addition to investigating identity discrimination *across modalities* using crossmodal RDMs, we also investigated the ability of each ROI to discriminate between identities *within modality,* using the face and voice

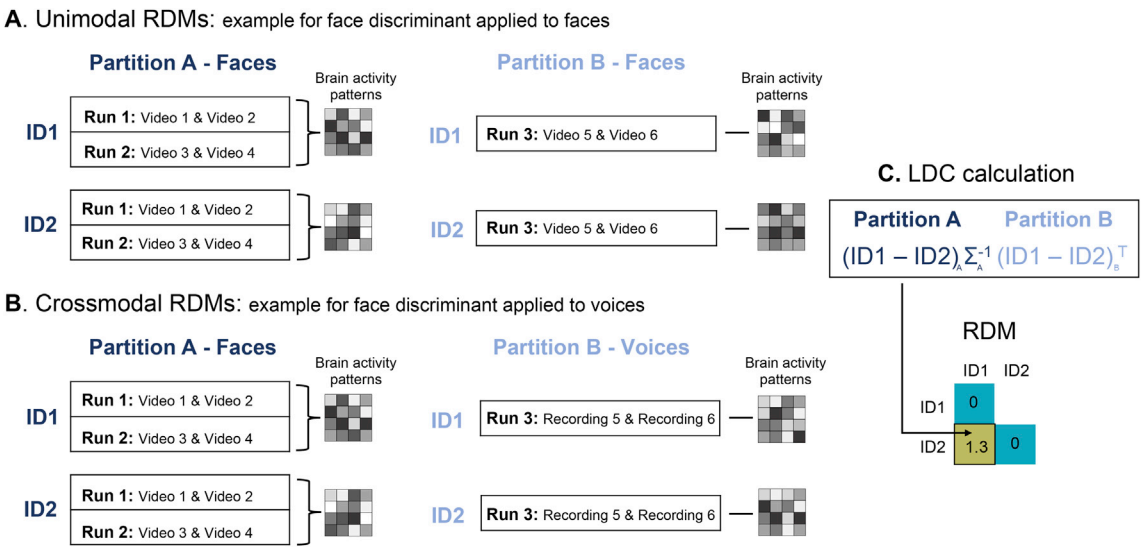


Fig. 2. Procedure for LDC calculation. A: Illustration of the crossvalidation procedure applied for the unimodal RDMs used in Analyses A & B. This example shows a single crossvalidation fold, in which runs 1 & 2 form data partition A and run 3 forms data partition B. Brain activity patterns are obtained for the videos of the face of identity 1 (ID1) and the face of identity 2 (ID2) in both partitions. **B**: Illustration of the crossvalidation procedure applied for the crossmodal RDMs used in Analysis B. This procedure is identical to A, with the exception that activity patterns in partition B are obtained for the recordings of the *voices* of ID1 and ID2. **C**: Illustration of the calculation of the LDC between two identities (ID1 and ID2). A discriminant is obtained by contrasting the activity patterns for ID1 and ID2, and then applied to the contrast between the activity patterns for the same identities in Partition B. Multivariate noise normalisation is applied using the noise variance-covariance matrix (Σ). The resulting value, which is entered in the corresponding cell of the RDM, shows the discriminability of the activity patterns for the two identities. For unimodal RDMs, the LDC represents identity discriminability within modality. For crossmodal RDMs, the LDC represents identity discriminability across modalities, generalising across faces and voices.

RDMs that were created in Analysis A. We expected that face or voice RDMs for regions that represent face or voice identity, respectively, would show mean LDC values that are significantly greater than zero. For this analysis, the corresponding RDMs for each scanning session (e.g. faces session 1 and faces session 2) were averaged across the two sessions, and then the mean LDC across the vectorised matrix was calculated (Fig. 6B). Thus, for each participant and each ROI we obtained four mean LDC values representing (1) face discriminability, (2) voice discriminability, (3a) crossmodal discriminability - face discriminant generalised to voices, and (3b) crossmodal discriminability - voice discriminant generalised to faces (Fig. 6B). For each ROI and each type of discriminability, we entered participants' LDC values into a one-sample one-tailed *t*-test comparing them against zero. P values were corrected for all comparisons (48 comparisons: 4 tests x 12 ROIs) controlling for FDR with $q < 0.05$.

2.7.4. Exploratory whole-brain searchlight analyses

Despite including a broad range of functionally defined ROIs, it is possible that modality-general person-identity representations may exist in brain regions not included in our ROIs. Specifically, these representations may exist in brain regions that are not face-selective or voice-selective. Therefore, we used an exploratory whole-brain searchlight analysis to identify potential brain regions with person-identity representations using the same methods as in our main ROI analyses. We note that we focused solely on modality-general person-identity representations in this exploratory analysis, as that was the main aim of this study.

For each participant, we created 6 mm radius spheres centred on each voxel within a grey-matter mask of their brain (obtained from the segmentation procedure) using the RSA toolbox (Nili et al., 2014) in Matlab. A 6 mm radius resulted in a searchlight sphere of 33 voxels, which matched our requirement for minimum ROI size of 30 voxels in the main analyses. For the analysis comparing representational geometries, we computed a face and a voice RDM in each searchlight sphere, averaging the RDMs from both scanning sessions, and then calculated the Pearson correlation between them. Correlations were Fisher z-transformed. The output of this analysis was a whole-brain map of Fisher-transformed correlation coefficients for each participant.

For the second analysis investigating identity discriminability, we computed a single crossmodal RDM in each searchlight sphere by averaging the crossmodal face-voice RDM with the crossmodal voice-face RDM, and then calculating the mean LDC across the resulting matrix in vector form. The output for each participant was a whole-brain map of mean LDC values.

The whole-brain searchlight maps from each analysis were normalised to MNI space using the normalisation parameters generated during the segmentation procedure and spatially smoothed with 9-mm Gaussian kernel (full width at half maximum) to correct for errors in intersubject alignment. For group-level analysis, all searchlight maps were entered into a one-sample *t*-test to determine whether the correlation coefficient/mean LDC value was significantly greater than zero at each voxel. We used the randomise tool (Winkler et al., 2014) in FSL (version 5.0.9; RRID:SCR_002823; Jenkinson et al., 2012) for inference on the resulting statistical maps (5000 sign-flips). Clusters were identified with threshold-free cluster enhancement (TFCE), and p-values were corrected for multiple comparisons (FWE < .05).

2.8. Functional localiser runs: data analysis and ROI definition

For the face, voice, and person localisers there were three experimental regressors in each localiser: (1) famous faces, (2) non-famous faces, and (3) visual objects and scenes in the face localiser; (1) famous voices, (2) non-famous voices, and (3) auditory objects and scenes in the voice localiser; (1) audiovisual famous people, (2) audiovisual non-famous people, and (3) audiovisual objects and scenes in the person localiser. For the TVA localiser there were two experimental regressors: (1) voices and (2) non-voices. Selectivity was defined with a *t*-test contrasting the responses to faces/voices/people (famous and non-famous) *versus* responses to the control stimuli.

To define the ROIs, we used a procedure similar to the Group-Constrained Subject-Specific method proposed by Fedorenko et al. (2010) and Julian et al. (2012). This method has the advantages of reducing experimenter bias and allowing for reproducible results, which

are particularly valuable in cases in which there is much variability across participants in the location, level of activity, and size of ROIs (for example, for face selective ROIs, see Rossion et al., 2012). The Group-Constrained Subject-Specific method starts by defining group functional masks that are consistently activated across participants (ideally defined with independent data), and then intersecting each participant's activation map with those group masks to define individual ROIs (Fedorenko et al., 2010; Julian et al., 2012). To define functional masks, we used probabilistic maps from previous studies that had used the same functional localisers, and then we intersected each participant's activation map with the respective functional masks (i.e. we extracted all selective voxels within each functional mask). SI-3 describes the full details of ROI definition.

Using this method, we attempted to define for each participant the following ROIs: (1) face-selective ROIs based on the face localiser: right fusiform face area (rFFA), right occipital face area (rOFA), and right posterior superior temporal sulcus (rpSTS); (2) voice-selective regions based on the voice localiser: right and left superior temporal sulcus and gyrus (rSTS/STG, lSTS/STG); (3) voice-selective regions based on the TVA localiser: right and left TVA (rTVA, lTVA); and (4) person-selective multimodal regions based on the person localiser: precuneus/posterior cingulate (Prec./P.Cing. —please note that this also included retrosplenial cortex), orbitofrontal cortex (OFC —please note that this region included a broad region of the ventromedial prefrontal cortex), frontal pole (FP — please note that this region was also large and also included part of the superior frontal gyrus), and right and left temporal pole with anterior inferior temporal cortex (rTP-aIT, lTP-aIT) — we considered the TP and aIT together as the peaks were difficult to separate in most participants.

We note that in the context of the present study, we define person-selective multimodal regions as regions that showed significantly higher responses to audiovisual clips of speaking faces than to audiovisual clips of scenes and objects, within regions that selectively responded to *both* faces and voices (i.e. using group functional masks of regions that responded both to faces and voices) — please see full details in SI-3. A limitation of this approach is that we were unable to define *a priori* person-selective regions in the STS, a region which has been shown to respond to both faces and voices (e.g. Deen et al., 2015; Watson et al., 2014a). This is because both voice- and face-selective voxels are spread throughout the STS, with many different peaks (e.g. Pernet et al., 2015; Deen et al., 2015). Most of these peaks are contiguous, and therefore it would be subjective to separate them, and also to assess which peaks are comparable across participants (i.e. which one of several peaks in one participant corresponds to which peak in another participant). Moreover, previous studies have shown a patchy organisation of the STS, in which multisensory responses are interspersed with unisensory responses (Beauchamp et al., 2004; Gentile et al., 2017). This patchy organisation makes it difficult to define ROIs of sufficient size to conduct multivariate analyses. In sum, we did not define person-selective multimodal ROIs in the STS *a priori* because we were unable to do so on an individual basis and using bias-free procedures. However, future studies could focus on developing procedures that better allow for definition of these ROIs.

2.9. Code and data availability statement

Data and code to reproduce the main analyses are available at https://doi.org/10.17633/rd.brunel.6429200.v1.

3. Results

In each of 30 participants, we computed beta parameter estimates at each voxel of each ROI for the 12 face-identities and 12 voice-identities, obtaining a response pattern for each identity in each ROI. For the main analyses, we then computed representational dissimilarity matrices (RDMs) for each of our ROIs using the LDC — Fig. 2 (Nili et al., 2014; Walther et al., 2016). To investigate the existence of modality-general person-identity representations in each of our ROIs, we performed two main analyses. In Analysis A, we computed RDMs for the faces and voices of the 12 identities and compared their representational geometries, i.e., we compared the RDMs across modalities. In Analysis B, we computed crossmodal RDMs and calculated the mean crossmodal discriminability across all identity pairs. Next, after briefly summarising the response profile of the ROIs and the behavioural results during the main experimental task, we describe the results for these two main analyses.

3.1. ROIs and mean responses to faces and voices

Using functional localisers, we defined for each participant (1) face-selective ROIs based on a face localiser: right fusiform face area (rFFA), right occipital face area (rOFA), and right posterior superior temporal sulcus (rpSTS), (2) voice-selective ROIs: right and the left superior temporal sulcus and gyrus (rSTS/STG, lSTS/STG) based on our own voice localiser, and right and left TVA (rTVA, lTVA) based on an established voice localiser (Belin et al., 2000; Pernet et al., 2015), and (3) person-selective multimodal ROIs based on a person localiser: precuneus/posterior cingulate (Prec./P.Cing.), orbitofrontal cortex (OFC), frontal pole (FP), and right and left temporal pole with anterior inferior temporal cortex (rTP-aIT, lTP-aIT). We were able to localise these ROIs with at least 30 voxels in all 30 participants, except for the rFFA (28 participants), the rOFA (29 participants), the Prec./P.Cing. (26 participants), and the OFC (21 participants) — please see SI-3 for full details. We note that the voice-selective ROIs in the right hemisphere (rTVA, rSTS/STG) overlap with each other and with the face-selective rpSTS and the person-selective multimodal rTP-aIT ROIs. In addition, the voice-selective ROIs in the left hemisphere (lTVA, lSTS/STG) overlap with each other and with the person-selective multimodal lTP-aIT ROI. For visualisation purposes only, Fig. 3 shows location of all ROIs in standardised space.

In order to confirm that each ROI showed the expected responsiveness to faces and voices in the main experimental runs, we computed the regional mean of the parameter estimates for faces and for voices across participants for each ROI and modality (Fig. 4). All regions showed the expected pattern of responses to faces and voices. Full results are described in SI-4, but we summarise here two main findings. First, although the rpSTS was defined based on face-selectivity, this region responded to both faces and voices. In fact, the rpSTS showed significantly greater responses to voices compared with faces ($p = .0002$). Second, the frontal pole did not show significant responses to faces compared to baseline and thus, although we still included this ROI in the main analysis, we do not think it displays person-selective responses.

3.2. Behavioural results for main experimental runs

In the task that participants completed during the main experimental runs, participants had to identify 36 novel faces and 36 novel voices belonging to famous people that were not among the 12 people that participants had been familiarised with prior to scanning. On average, participants correctly identified 97% ($M = 35$, $SD = 1.91$) of the novel faces, and 74% ($M = 26.8$, $SD = 6.51$) of the novel voices. Significantly more faces were identified than voices ($t(29) = 8.3922$, $p < .0001$). This is consistent with previous findings showing that familiar faces are easier to recognise than familiar voices (Damjanovic and Hanley, 2007; Hanley and Damjanovic, 2009; Hanley et al., 1998).

3.3. Analysis A: RSA comparing representational geometries

Analysis A focused on the representational geometry of all identities, i.e. the entire structure of pairwise distances between the activity patterns elicited by these identities in each modality, and compared geometries across modalities (Kriegeskorte et al., 2008a, 2008b; Kriegeskorte and Kievit, 2013). For this analysis, we computed four RDMs per participant per ROI: faces session 1, voices session 1, faces session 2, and

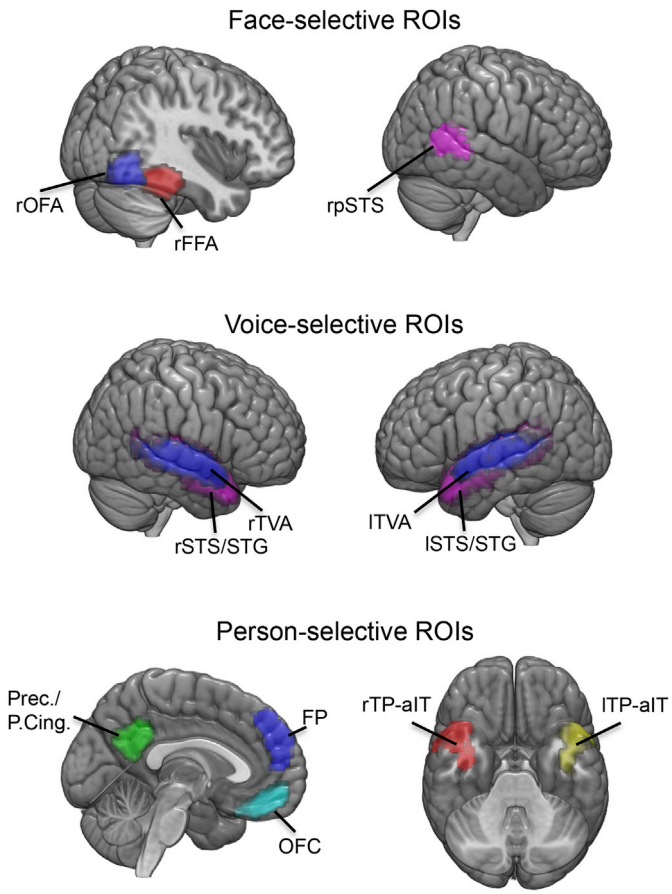


Fig. 3. Face-selective, voice-selective, and person-selective multimodal ROIs. Location of ROIs that resulted from the face, voice, and person localisers in MNI space. For illustration purposes only, we created probabilistic maps of all ROIs by normalising the single subject ROIs to MNI space and summing them across participants. Then, we thresholded those maps to display all voxels that were present in at least 20% of the participants. Please note that these ROIs were not used in the analyses, which used participant-specific ROIs in native space. Different colours are used for illustration purposes only. r = right, l = left, FFA = fusiform face area, OFA = occipital face area, pSTS = posterior superior temporal sulcus, STS/STG = superior temporal sulcus/superior temporal gyrus, TVA = temporal voice area, OFC = orbitofrontal cortex, FP = frontal pole, TP = temporal pole, aIT = anterior inferior temporal cortex, Prec. = precuneus, P. Cing. = posterior cingulate.

voices session 2 (Fig. 5B). We then compared the RDMs across modalities and predicted that correlations between face and voice RDMs would be significantly greater than zero in ROIs that represent person-identity independently from modality. However, our results showed no significant correlations between face and voice RDMs in face-selective, voice-selective, or person-selective multimodal ROIs (Fig. 5 and SI-5). It is possible that comparing RDMs across different scanning sessions taking place on separate days did not allow us to detect subtle consistencies in the representational geometry for face-identities and voice-identities. To address this concern, we also compared face and voice RDMs within the same scanning session (i.e. we correlated faces 1 with voices 1, and also faces 2 with voices 2). However, we still found no significant correlations between face and voice RDMs. Therefore, using this method we found no evidence of modality-general person-identity representations in our ROIs.

We also predicted that there would be correlations between RDMs within the same modality across sessions in regions that represent only face-identity or only voice-identity. No correlations between face RDMs or between voice RDMs in any ROI were significant after correction for multiple comparisons (Fig. 5A and SI-5).

3.4. Analysis B: RSA investigating identity discriminability

Our second main analysis tested the generalisation of pattern discriminants from one modality to the other. For this analysis, we additionally computed two crossmodal RDMs for each participant and ROI, which showed whether linear discriminants computed on pairs of faces could be used to discriminate between pairs of voices, and vice-versa. We then computed the mean LDC distance across all cells in each crossmodal RDM as an overall measure of the ability of each ROI to discriminate between identities using crossmodal information (Fig. 6B). We predicted that, in brain regions with modality-general person identity representations, the mean LDC values for crossmodal RDMs would be significantly greater than zero. Our results showed that mean LDC values in these RDMs were significantly greater than zero in the rpSTS (Fig. 6A and Table 1). These results show that the rpSTS could discriminate pairs of face-identities based on pattern discriminants computed from pairs of voice-identities (and vice-versa), and therefore appears to form modality-independent person-identity representations. We note that we had defined the rpSTS using the face localiser, but that this region also showed substantial responses to voices, as shown in Fig. 4. To further probe the response properties of the rpSTS, we investigated responses in this region during the TVA, voice, and person localisers. These supplementary analyses (SI-6) showed that the rpSTS also demonstrates voice-selectivity and person-selectivity in these localisers (for similar results, see Deen et al., 2015).

We note that while the mean LDC values for crossmodal RDMs in the lSTS/STG were also significant, the mean LDC value for face RDMs was not. While this result suggests that this region was able to discriminate identities based on crossmodal information, it is unlikely that a modality-general representation could exist without face-identity discrimination. Therefore, this result should be interpreted with caution. It is possible that in addition to the rpSTS, the lpSTS also contains a modality-general person-identity representation that could be driving the positive result in the lSTS/STG. However, we were not able to test this because we could not consistently localise the lpSTS using our localisers.

We also tested whether each ROI could discriminate between pairs of stimuli within the same modality (Fig. 6B). We predicted that mean LDC values for face RDMs and voice RDMs would be significantly greater than zero in ROIs that represent face-identity and voice-identity, respectively. We found that mean LDC values in face RDMs were significantly greater than zero in all ROIs originally defined as face-selective (rFFA, rOFA, rpSTS), in the TVAs, and in the person-selective multimodal Prec./P.Cing (Fig. 6A and SI-7). These results show that all these regions could discriminate between face-identities. A follow up analysis in which all overlapping rpSTS voxels were removed from the rTVA showed that the significant result for faces in rTVA was driven by the rpSTS. Mean LDC values in voice RDMs were significantly greater than zero in all voice-selective ROIs (TVAs, STS/STG), in the rpSTS, and in the person-selective multimodal OFC, FP, rTP-aIT and Prec./P.Cing. (Fig. 6A and SI-7).

It is possible that the discrimination of identities in our ROIs was driven by different-gender identity pairs (female-male). To investigate this possibility, for each ROI and condition that showed mean LDC values significantly greater than zero, we compared the mean LDC values for different-gender identity pairs (calculated across 36 pairs: male-female) with the mean LDC values for same-gender identity pairs (calculated across 30 pairs: female-female & male-male) in each RDM (we used paired *t*-tests, and used FDR correction for all 19 comparisons). Results for the rpSTS showed no significant difference between the discriminability of different-gender and same-gender identity pairs for face, voice, or crossmodal RDMs (all $p > .0533$), demonstrating that person-identity discrimination in this region was not driven by discriminating gender. In contrast, mean LDC values for different-gender identity pairs were significantly higher than mean LDC values for same-gender identity pairs for face RDMs in the rFFA and rOFA (both $p \leq .0010$), and for voice RDMs in the bilateral TVAs and STS/STG (all $p \leq .0005$), suggesting that gender

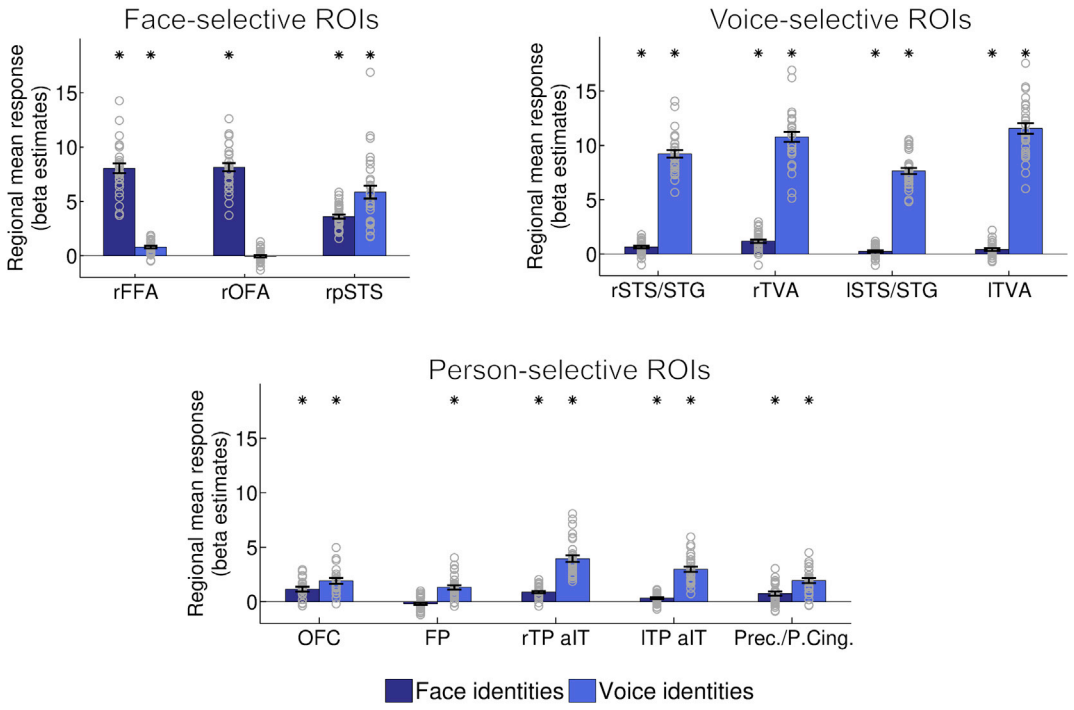


Fig. 4. Regional mean responses to faces and voices in ROIs. Regional mean responses for all faces and for all voices in face-selective, voice-selective, and person-selective multimodal ROIs (mean beta estimates across all voxels of each ROI, and across all runs). Bars show mean responses across participants, error bars show standard error, and grey circles show individual participants. We tested whether mean responses were significantly greater than zero using one-sample *t*-tests across all 30 participants, and stars show significant results at $p \leq .0209$ (FDR corrected for all 24 comparisons). We also tested whether mean beta values for faces were significantly different from mean beta values for voices in each ROI using paired *t*-tests across all participants. In all ROIs mean beta values for faces and voices were significantly different at $p \leq .0011$ (FDR corrected for all 12 ROIs).

contributed to the discrimination in these regions. However, mean LDC values for same-gender identity pairs were still significantly greater than zero (one-sample *t*-tests) for face RDMs in the rFFA and rOFA (both $p < .0001$) and for voice RDMs in the bilateral TVAs and STS/STG (all $p \leq .0239$), suggesting that identity discrimination in these regions is not solely driven by differences in gender.

3.5. Exploratory whole-brain searchlight analyses

Finally, we conducted additional exploratory searchlight analyses across the whole brain to determine whether there are brain regions with modality-general person-identity representations that were not included in our ROIs. The first searchlight analysis investigated correlations between face and voice RDMs across the whole brain, and we did not find any regions showing such correlations between face and voice representational geometries.

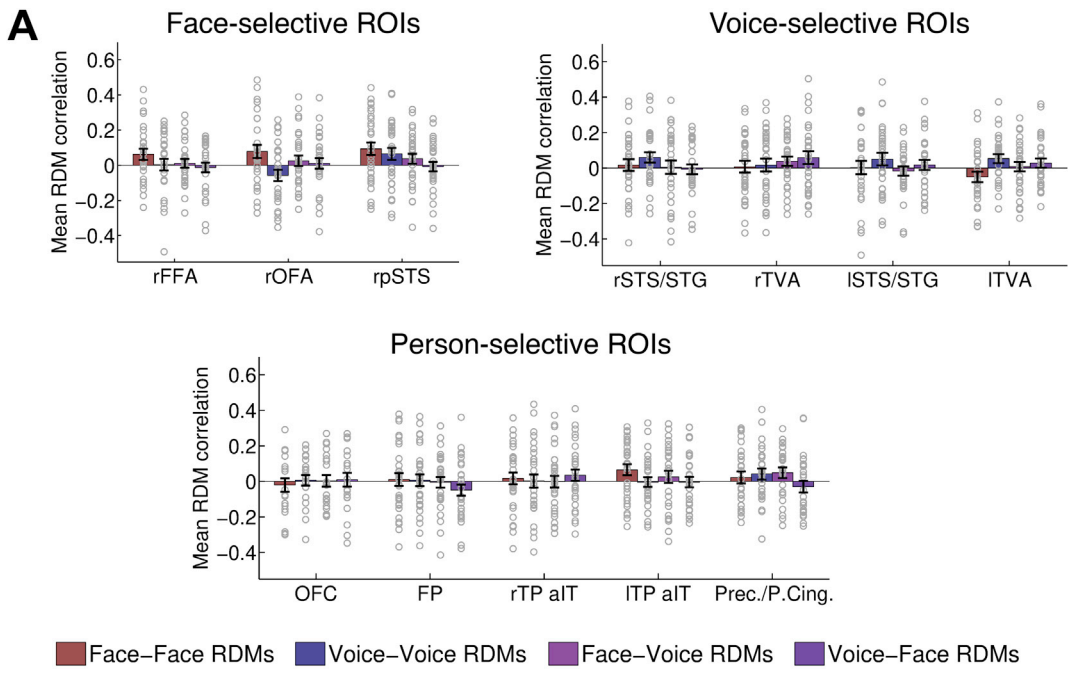
The second searchlight analysis investigated crossmodal generalisation of discriminants for pairs of identities across the whole brain. We found a number of clusters in which the mean LDC in crossmodal RDMs was significantly greater than zero (FWE corrected threshold $p \leq .05$), and below we report t-values and MNI coordinates for the peak grey matter voxels in each cluster. Anatomical labels for peak voxels are based on the Harvard-Oxford cortical and subcortical structural atlases. The results showed a large cluster ($k = 1927$, $p = .007$) with peaks in the right putamen ($t = 4.33$, x = 21, y = 20, z = −1), the left posterior middle temporal gyrus ($t = 4.04$,x = −57, y = −19, z = −7), and the right precentral gyrus ($t = 3.89$, x = 54, y = 8, z = 32). Significant clusters were also found in the right paracingulate gyrus (k = 1340, $p = .003$, $t = 4.34$, x = 6, y = 47, z = 23), in the left hippocampus (k = 160, $p = .017$, $t = 4.45$, x = −24, y = −37, z = 2), in the right anterior supramarginal gyrus (k = 84, $p = .006$, $t = 6.18$, x = 48, y = −22, z = 38), in the left cuneal cortex (k = 48, $p = .036$, $t = 3.99$, x = −18, y = −76, z = 29), and a cluster ($k = 100$, $p = .039$) with peaks in the left temporooccipital middle temporal gyrus ($t = 3.58$, x = −48, y = −46, z = 5) and inferior lateral occipital cortex ($t = 3.45$, x = −48, y = −67, z = 8). Finally, we also found a significant cluster in the rpSTS at an uncorrected threshold of $p \leq .005$ (k = 592, $p = .001$, $t = 4.05$, x = 48, y = −49, z = 11) that overlapped with our rpSTS ROI.

4. Discussion

We show evidence of a modality-general person-identity representation in a face-selective, voice-selective and person-selective region of the rpSTS, demonstrating that this region was able to discriminate familiar identities based on modality-general information in faces and voices. More specifically, the rpSTS could discriminate response patterns for pairs of face-identities based on linear discriminants computed from response patterns for pairs of voice-identities, and vice-versa. A crucial and novel aspect of our study is that we showed that the rpSTS not only discriminates between identities, but also generalises across multiple naturalistically varying face videos and voice recordings of the same identity. By always comparing response patterns across independent runs with different face and voice tokens for the same identities, we showed that the face- and voice-elicited person-identity representations in the rpSTS are stimuli-invariant. Invariant identity representations were also found for face-identities in face-selective regions (rFFA and rOFA) and for voice-identities in voice-selective regions (bilateral TVA and STS/STG). Finally, we did not find evidence of matching representational geometries for faces and voices, across or within modalities.

4.1. A modality-general and invariant person-identity representation in the rpSTS

Although the rpSTS region was defined using a face localiser, we showed that it responded to both faces and voices, and also showed voice-selectivity in the voice localisers and person-selectivity in the



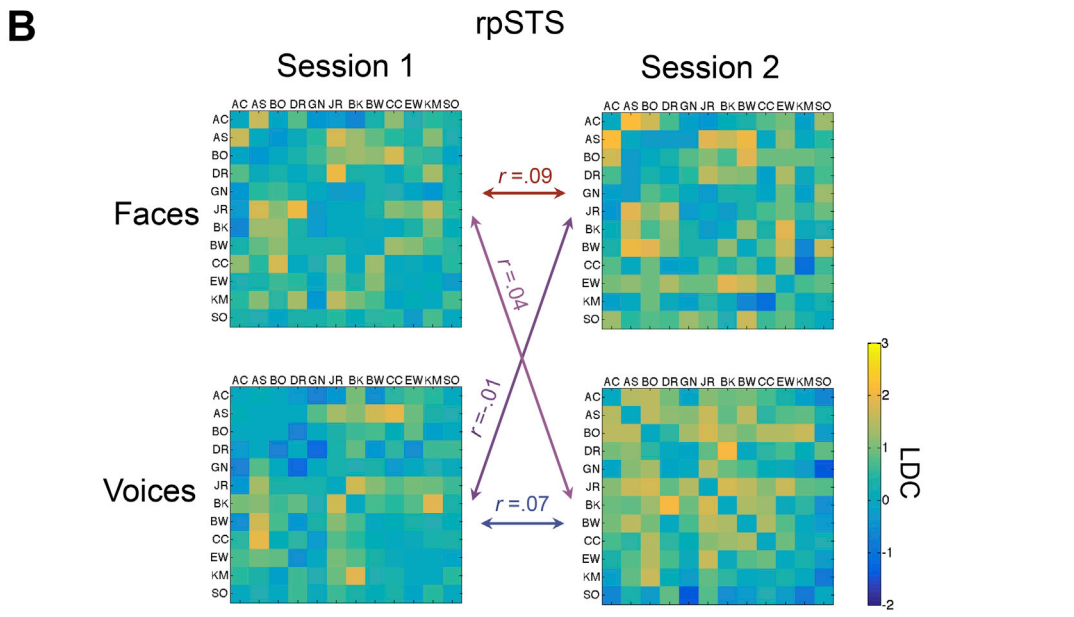
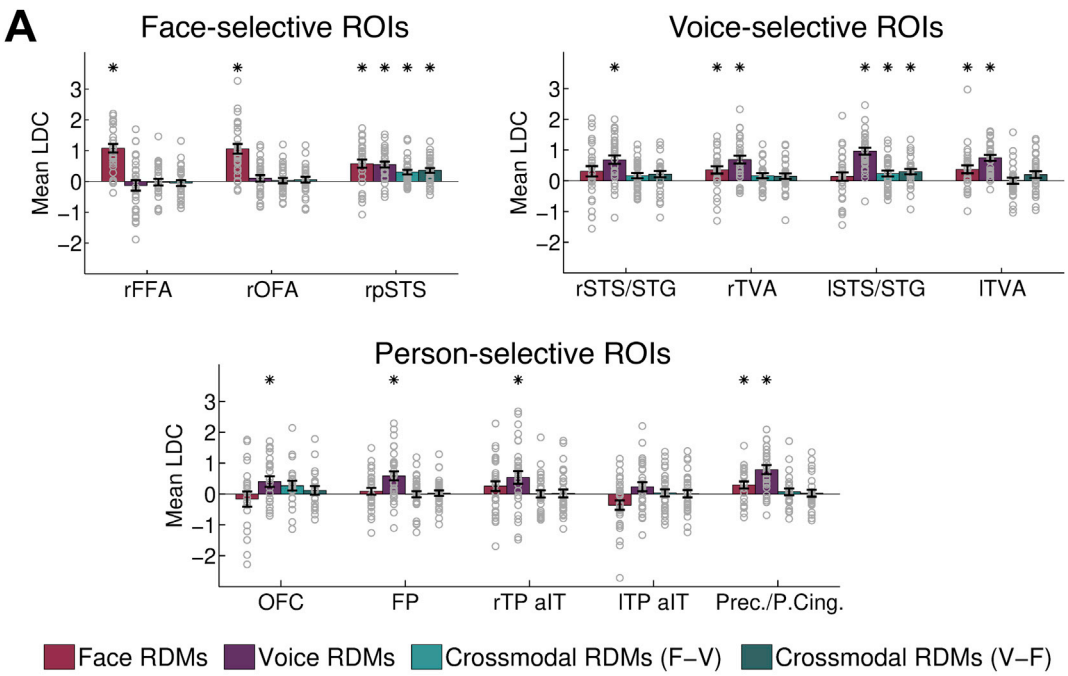


Fig. 5. Results of RSA comparing representational geometries. **A:** Comparisons between the representational dissimilarity matrices (RDMs) from two scanning sessions using Pearson's correlation coefficient. Bars show mean correlations across participants, error bars show standard error, and grey circles show the correlations of individual participants. Correlations were calculated across scanning sessions and compared face RDMs, voice RDMs, face and voice RDMs, and voice and face RDMs in face-selective, voice-selective, and person-selective multimodal ROIs. We tested whether correlations were significantly greater than zero using Wilcoxon signed-rank tests across all 30 participants. No correlations were significant after correction for multiple comparisons at $p \leq .001$ (FDR corrected for all 48 comparisons). **B:** Example of RDM comparisons across sessions 1 and 2 in the rpSTS. Face and voice RDMs for the rpSTS were averaged across all 30 participants for illustration purposes. Each cell shows the discriminability of the brain activity patterns corresponding to a pair of identities (12 identities in total) computed using the linear discriminant contrast (LDC) and crossvalidating across data from three runs. Each matrix is symmetric around a diagonal of zeros. A value of zero or lower indicates no discriminability. For each participant we compared the representational geometry of the face and voice RDMs with the representational geometry in the RDM of the *other* modality (crossmodal comparisons) and in the RDM of the *same* modality (unimodal comparisons) using Pearson's correlation. The figure shows Pearson's correlations for all comparisons averaged across participants.

person localiser. These findings are in line with a number of previous studies showing overlap between face-selective and voice-selective regions in the rpSTS (Watson et al., 2014a; Deen et al., 2015; Anzellotti and Caramazza, 2017), and stronger responses to audiovisual face-voice stimuli than to control audiovisual stimuli (Watson et al., 2014a). Here we extended these findings by demonstrating that the rpSTS not only shows selective responses to faces, voices, and people, but is also able to discriminate between different face-identities and voice-identities within modalities, and, crucially, to discriminate between person-identities using crossmodal information. Overall, our findings suggest that rpSTS



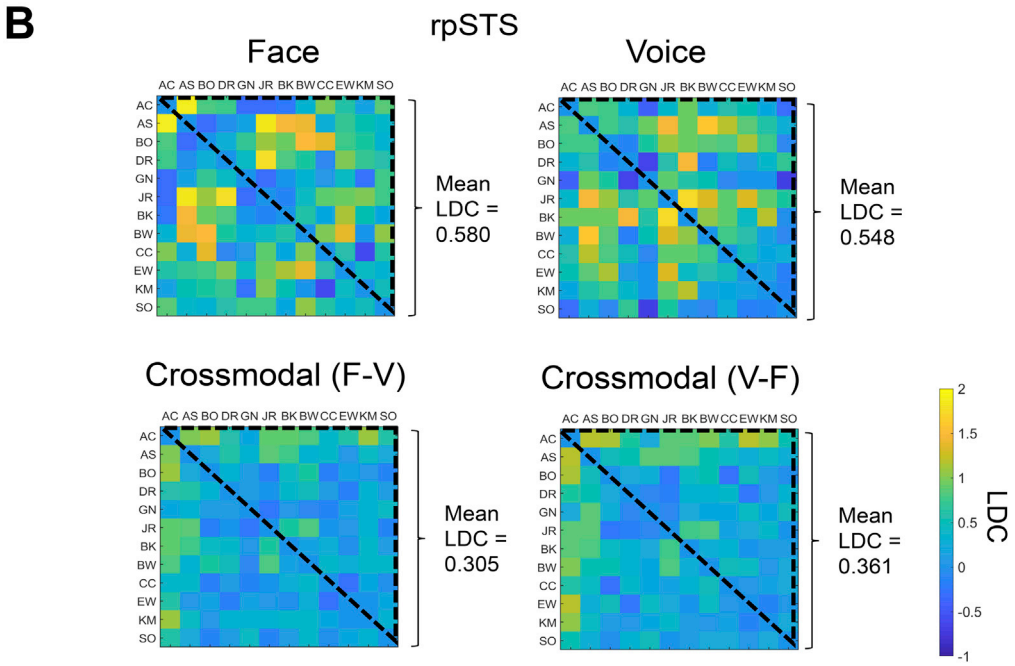


Fig. 6. Results of RSA investigating identity discriminability. A: Mean LDC between identities in face RDMs, voice RDMs, and crossmodal RDMs in face-selective, voice-selective, and person-selective multimodal ROIs. There are two types of crossmodal RDMs: (a) face discriminant applied to voices (F–V), and (b) voice discriminant applied to faces (V–F). Bars show mean LDC values averaged across participants, error bars show standard error, and grey circles show mean LDC values for individual participants. We tested whether the mean LDC values were significantly greater than zero using one-sample *t*-tests across all 30 participants. Stars represent significant tests at $p \leq .0150$ (FDR corrected for all 48 comparisons). These results show generalisation of the pattern discriminants from one modality to the other in the rpSTS and in the lSTS/STG. In addition, face-selective ROIs discriminate between face-identities, and voice-selective ROIs discriminate between voice-identities. **B:** Example of the calculation of the mean LDC for a face RDMs, a voice RDM, and two crossmodal RDMs in the rpSTS. Note that, in contrast to Analysis A, these RDMs were averaged across the two scanning sessions. For each participant we calculated the mean LDC across all cells on one side of the diagonal of each RDM (RDMs are symmetric around a diagonal of zeros). This value represents the mean discriminability of all pairs of identities in the RDM. The RDMs shown in the figure were averaged across all 30 participants for illustration purposes, and the mean LDC values have been averaged across participants.

is a person-selective multimodal brain region that represents crossmodal information that can be used to discriminate person-identities independently of the modality of the stimuli. This finding supports a model of face and voice integration whereby face and voice identity information is integrated in multimodal brain regions (Ellis et al., 1997; Campanella and Belin, 2007).

Our results do not mean that modality-general person-identity representations only exist in multimodal regions. In fact, our results show

Table 1
One-sample *t*-test results for mean LDC values in crossmodal RDMs.

		Crossmodal RDMs (face-voice)			Crossmodal RDMs (voice-face)		
	df	t	Sig. (1-tailed)	d	t	Sig. (1-tailed)	d
Face-selective ROIs							
rFFA	27	−0.198	.5779	0.04	−0.529	.6993	0.10
rOFA	28	0.374	.3557	0.07	0.624	.2689	0.12
rpSTS	29	4.091	.0002*	0.75	4.582	.0001*	0.84
Voice-selective ROIs							
rSTS/STG	29	1.928	.0319	0.35	2.093	.0226	0.38
rTVA	29	2.064	.0240	0.38	1.662	.0537	0.30
lSTS/STG	29	2.443	.0104*	0.45	3.543	.0007*	0.65
lTVA	29	0.062	.4755	0.01	1.891	.0343	0.35
Person-selective ROIs							
OFC	20	1.698	.0525	0.37	0.841	.0250	0.18
FP	29	−0.062	.5244	0.01	0.285	.3888	0.05
rTP-aIT	29	0.023	.4910	0.00	0.153	.4398	0.03
lTP-aIT	29	0.301	.3830	0.05	0.075	.4703	0.01
Prec./P.Cing.	25	0.660	.2577	0.13	0.220	.4138	0.04

Note: Stars represent statistical significance at $p \leq .0150$ (FDR corrected for all 48 comparisons in face, voice, and crossmodal RDMs).

some evidence that the voice-selective left STS/STG could also integrate face and voice information, and future studies could explore this further. The choice of task could also affect the type of integration mechanism that is recruited. For example, explicit voice recognition tasks have been shown to activate face-responsive regions (e.g. von Kriegstein et al., 2005, 2006), and therefore it is possible that face and voice integration through the coupling of face and voice-responsive regions is contingent on an explicit identity recognition task. Moreover, future studies should focus on defining a face- or person-selective lpSTS region, in addition to the rpSTS, and investigating the functional properties of this region. Furthermore, as we explained earlier, our study had the limitation of not being able to identify person-selective multimodal regions along the STS (bilaterally) *a priori*, and future studies could focus on defining these regions reliably across participants.

Our finding of a modality-general identity representation in the rpSTS is in agreement with two recent studies showing across-modality classification of pattern estimates for familiar faces and voices in the rpSTS (Anzellotti and Caramazza, 2017) and a more anterior part of the STS (Hasan et al., 2016). However, in contrast to these previous studies, we used a larger set of identities and multiple naturalistically varying tokens for each identity, and we additionally showed that face- and voice-elicited representations in the rpSTS were also invariant to different tokens of the same face and voice. Hasan et al. (2016) were unable to investigate invariant representations within each modality because they used a single face image and a single voice recording for each identity, which in turn were derived from the same original audiovisual stimulus, making interpretation of their results difficult (Lavan, 2017). Anzellotti and Caramazza (2017) used two face and voice tokens for each identity but did not train and test their classifier on different tokens of the same modality, and therefore did not demonstrate representations that were invariant to different tokens of the same face or voice in their study. We think that in our study, by showing that representations in rpSTS were invariant to multiple different face and voice tokens of the same identity, we can make a stronger case that the rpSTS may be coding for person-identity related information instead of general and recently learned associations between stimuli (Lavan, 2017).

The rpSTS has also been previously associated with crossmodal representations of emotion from faces and voices (Watson et al., 2014b), and with crossmodal representations of person-identity, in that it responded more to identity-incongruent face and voice pairs than to identity-congruent face and voice pairs (Hölig et al., 2017). In addition, multiple studies have shown the importance of the bilateral pSTS for audiovisual integration of speech information extracted from faces and voices, by demonstrating that these regions responded more to audiovisual speech than to unimodal visual and auditory speech (e.g. Calvert et al., 2000; Wright et al., 2003; Watson et al., 2014a), and by showing that sub-regions of the pSTS are engaged in audiovisual integration (Beauchamp et al., 2004; Rennig and Beauchamp, 2018). Given that the face videos presented in our study did not contain speech, we note that it is unlikely that our finding of a modality-general person-identity representation in the pSTS could be explained by audiovisual integration of speech.

Future work could further characterise the crossmodal information represented in the rpSTS. One possibility is that the rpSTS represents information about a person's idiosyncratic facial movements and, in line with this view, Yovel and O'Toole (2016) proposed that the STS integrates person-specific patterns of movement from faces, voices, and bodies to assist in person-identity recognition. To further test this, it would be interesting to build candidate models of types of information that could be represented in the rpSTS, including models of the patterns of movement, but also models of visual and auditory properties of the stimuli, and even models of the social information associated with people, such as social distance (Parkinson et al. 2014, 2017) to correlate with the brain representations and thus shed light on what type of information is represented in rpSTS.

4.2. Naturalistically varying stimuli and invariant representations of face-identity and voice-identity

A crucial aspect of our study is that we used a large set of familiar identities and multiple naturalistically varying tokens in order to better capture the level of robust invariant recognition required in everyday life (Jenkins et al., 2011; Burton, 2013; Burton et al., 2016; Lavan et al., 2018a, 2018b). The ability to "tell people together" by identifying different tokens of a face and voice as belonging to the same person is as important as the ability to "tell people apart" (i.e. discriminate between different people) (Burton, 2013; Anzellotti and Caramazza, 2014; Lavan et al., 2018a). In line with this, although we used these highly variable stimuli, we showed that the representations in a number of our ROIs generalised across different tokens of the same modality.

The face-selective rFFA and rOFA were able to discriminate between the faces of different people while also showing invariance to the different videos of each person's face. This finding is in agreement with Anzellotti et al. (2014) and Guntupalli et al. (2017), who showed representations of face-identity in the FFA (and OFA, in Anzelotti et al., 2014) that generalise across different viewpoints of the face. However, in contrast with these studies, which used stimuli with low within-person variability, we show that representations in these regions generalise across highly variable face videos, and can thus discriminate between different face-*identities*, rather than between individual face *images*.

Voice-selective regions in STS/STG and the TVAs bilaterally could discriminate between different speakers while showing invariance to the different recordings of each voice. These findings are in line Formisano et al. (2008), who showed representations of speaker identity that generalise across utterances of different vowels in the lateral Heschl's gyrus/sulcus and in the right STS. We extend this finding by showing that generalisation across different recordings of the same voice is possible even when using short sentences with variable speech content that were recorded in different settings. It would be very interesting for future studies to investigate what type of information is being represented in these face- and voice-selective regions that can discriminate between identities.

We also found invariant discrimination of face- and voice-identity in a multimodal region in the precuneus/posterior cingulate. This region has been previously associated with the processing of familiar faces and voices (Shah et al., 2001), and has been found to discriminate between different face-identities (Visconti Di Oleggio Castello et al., 2017). Our results suggest that representations of faces and voices may be interspersed in this region, but are not shared across modalities. Finally, we

showed invariant representations of voice-identity, but not face-identity, in the frontal pole, a region that has been previously associated with the processing of familiar voices (Nakamura et al., 2001). It should be noted that, although we initially localised the frontal pole as a multimodal region, our results showed that it did not respond significantly to faces in the main experimental runs.

4.3. Representational geometries

We did not find matching representational geometries across faces and voices in rpSTS despite finding crossmodal generalisation of the pattern discriminants. One possible explanation is that, for the first analysis, all identities were equally distinct from each other within each modality. In other words, the nature of person-identity code in the rpSTS may not result in variable representational distances between identities, and therefore we cannot expect to find positive correlations between representational geometries across modalities. Another possibility is that, the rpSTS may represent both modality-specific and modality-general information (Laurienti et al., 2005; Driver and Noesselt, 2008), and the former may have had stronger influence on the representational geometries for this region. In line with this, Beauchamp et al. (2004) showed that the pSTS contains intermixed visual, auditory, and multisensory patches, and future studies could use higher-resolution neuroimaging methods to further probe person-identity representations in this region. Our second analysis used pattern discriminants and focused on a pair of identities at a time, and therefore we believe that it was more sensitive to detect modality-general representations, even in the presence of the constraints described above.

In all other ROIs, we also did not find any evidence of stable representational geometries for face-identities or voice-identities only. Again, it could be that identities were equally distinct across from each other within each modality. Finally, it could be that experimental conditions would need to be improved to obtain more reliable representational geometries. We think that it may be particularly important to do all the testing in one single session, if possible.

4.4. Anterior temporal lobe and searchlight results

We did not find evidence of face-, voice-, or person-identity representations in the anterior temporal lobe. This was surprising given that this region has been previously associated with the processing of person-identity (Ellis et al., 1989; Gainotti, 2011). The fact that our TP-aIT ROIs responded more to voices than to faces suggests that our multimodal region localiser was not optimal for detecting multimodal responses in the anterior temporal lobe. One possibility is that our sequences were not tailored to detect fMRI responses in this region (Axelrod and Yovel, 2013), and therefore more research using specialised scanning parameters for the localisation of this region is warranted. A second possibility for the lack of results in anterior temporal lobe could be related to a limitation in fMRI multivariate methods to decode identity information in this region, given the nature of the neural populations responsive to identity (Dubois et al., 2015). Dubois et al. (2015) demonstrated that, while they could decode face identity in the macaque anterior face patches when using single-unit data, they could no longer do so when using fMRI data, and they suggested that scattered units responding to identity in anterior face patches contributed to the lack of decoding of identity. However, we note that multiple studies using human fMRI have shown that the anterior temporal lobe can discriminate between face identities (e.g. Kriegeskorte et al., 2007; Nestor et al., 2011; Goesart and Op de Beeck, 2013; Anzelotti et al., 2014; Anzelotti and Caramazza, 2016).

It is possible that modality-general representations also exist outside our face-, voice-, and person-selective multimodal regions, and our exploratory searchlight results revealed person-identity representations in the paracingulate gyrus, right insular cortex, left nucleus accumbens, left anterior postcentral gyrus, and left hippocampus. Quiroga et al. (2005, 2009) found that cells in the hippocampus (and also amygdala and entorhinal cortex) were highly responsive to specific identities, and responded to both the face and name of that person. It will be interesting to further probe the role of the hippocampus (and the other regions found during the searchlight analyses) in person-identity recognition.

5. Conclusion

To conclude, we showed a modality-general person-identity representation that generalises across different, naturalistically varying face videos and voice recordings of the same person in a person-selective multimodal region of the rpSTS. We also found evidence of video-invariant face-identity representations in face-selective regions (rFFA, rOFA), and sound-invariant voice-identity representations in voice-selective regions (TVA, STS/STG). Future studies could focus on the nature and type of face and voice information that is represented in these different regions, and on how these representations are formed, both through development, and during the process of becoming familiar with someone.

Acknowledgments

This work was supported by a research grant by the Leverhulme Trust (RPG-2014-392). We thank Matthew Longo for comments on a previous version of the manuscript, and Tiana Rakotonombana, Roxanne Zamyadi, and Rasanat Nawaz for help with preparing and piloting the stimuli.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.07.017.

Conflicts of interest

The authors declare no competing interests.

References

Anzellotti, S., Caramazza, A., 2014. The neural mechanisms for the recognition of face identity in humans. Front. Psychol. 5, 1–6.
Anzellotti, S., Caramazza, A., 2016. From parts to identity: invariance and sensitivity of face representations to different face halves. Cerebr. Cortex 26, 1900–1909.
Anzellotti, S., Caramazza, A., 2017. Multimodal representations of person identity individuated with fMRI. Cortex 89, 85–97.
Anzellotti, S., Fairhall, S.L., Caramazza, A., 2014. Decoding representations of face identity that are tolerant to rotation. Cerebr. Cortex 24, 1988–1995.
Axelrod, V., Yovel, G., 2013. The challenge of localizing the anterior temporal face area: a possible solution. Neuroimage 81, 371–380.
Axelrod, V., Yovel, G., 2015. Successful decoding of famous faces in the fusiform face area. PLoS One 10 e0117126.
Blank, H., Anwander, A., von Kriegstein, K., 2011. Direct structural connections between voice-and face-recognition areas. J. Neurosci. 31 (36), 12906–12915.
Blank, H., Wieland, N., von Kriegstein, K., 2014. Person recognition and the brain: merging evidence from patients and healthy individuals. Neurosci. Biobehav. Rev. 47, 717–734.
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., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.
Boersma, P., Weenink, D., 2014. Praat: Doing Phonetics by Computer [Computer Software]. Version 5.3.80.
Brainard, D.H., 1997. The psychophysics toolbox. Spat vis 10, 433–436.
Burton, A.M., 2013. Why has research in face recognition progressed so slowly? The importance of variability. Q. J. Exp. Psychol. 66, 1467–1485.
Burton, A.M., Kramer, R.S.S., Ritchie, K.L., Jenkins, R., 2016. Identity from variation: representations of faces derived from multiple instances. Cogn. Sci. 40, 202–223.
Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10, 649–657.
Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends Cognit. Sci. 11, 535–543.

Carlin, J.D., Kriegeskorte, N., 2017. Adjudicating between face-coding models with individual-face fMRI responses. PLoS Comput. Biol. 13 e1005604.
Collins, J.A., Koski, J.E., Olson, I.R., 2016. More than meets the eye: the merging of perceptual and conceptual knowledge in the anterior temporal face area. Front. Hum. Neurosci. 10, 1–11.
Damjanovic, L., Hanley, J.R., 2007. Recalling episodic and semantic information about famous faces and voices. Mem. Cogn. 35, 1205–1210.
Deen, B., Koldewyn, K., Kanwisher, N., Saxe, R., 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cerebr. Cortex 25, 4596–4609.
Driver, J., Noesselt, T., 2008. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. Neuron 57, 11–23.
Dubois, J., Otto de Berker, A., Tsao, D., 2015. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. J. Neurosci. 35, 2791–2802.
Ellis, A.W., Young, A.W., Critchley, E.M.R., 1989. Loss of memory for people following temporal lobe damage. Brain 112, 1469–1483.
Ellis, H.D., Jones, D.M., Mosdell, N., 1997. Intra- and inter-modal repetition priming of familiar faces and voices. Br. J. Psychol. 88, 143–156.
Fedorenko, E., Hsieh, P.J., Nieto-Castañon, A., Whitfield-Gabrieli, S., Kanwisher, N., 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104, 1177–1194.
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.
Fox, C.J., Iaria, G., Barton, J.J.S., 2009. Defining the face processing network: optimization of the functional localizer in fMRI. Hum. Brain Mapp. 30, 1637–1651.
Gentile, F., can Atteveldt, N., de Martino, Gobel, R., 2017. Approaching the ground truth: revealing the functional organization of human multisensory STC using ultra-high field fMRI. J. Neurosci. 37, 10104–10113.
Gainotti, G., 2011. What the study of voice recognition in normal subjects and brain-damaged patients tells us about models of familiar people recognition. Neuropsychologia 49, 2273–2282.
Goesaert, E., Op de Beeck, H.P., 2013. Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. J. Neurosci. 33, 8549–8558.
Guntupalli, J.S., Wheeler, K.G., Gobbini, M.I., 2017. Disentangling the representation of identity from head view along the human face processing pathway. Cerebr. Cortex 27, 46–53.
Hailstone, J., Ridgway, G., Bartlett, J., Goll, J., Buckley, A., Crutch, S., Warren, J., 2011. Voice processing in dementia: a neuropsychological and neuroanatomical analysis. Brain 134, 2535–2547.
Hanley, J.R., Damjanovic, L., 2009. It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. Memory 17, 830–839.
Hanley, J.R., Smith, S.T., Hadfield, J., 1998. I recognise you but I can't place you: an investigation of familiar-only experiences during tests of voice and face recognition. Q J Exp Psychol. Sect. A 51, 179–195.
Hasan, B.A.S., Valdes-Sosa, M., Gross, J., Belin, P., 2016. "Hearing faces and seeing voices": amodal coding of person identity in the human brain. Sci. Rep. 6, 37494.
Hölig, C., Föcker, A., Best, A., Röder, B., Büchel, C., 2017. Activation in the angular gyrus and in the pSTS is modulated by face primes during voice recognition. Hum. Brain Mapp. 38, 2553–2565.
Jenkins, R., White, D., Van Montfort, X., Burton, A.M., 2011. Variability in photos of the same face. Cognition 121, 313–323.
Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage. 62, 82-790.
Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R., Campanella, S., 2011. Cross-modal interactions between human faces and voices involved in person recognition. Cortex 47, 367–376.
Julian, J.B., Fedorenko, E., Webster, J., Kanwisher, N., 2012. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60, 2357–2364.
Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.
Kriegeskorte, N., Formisano, E., Sorger, B., Goebel, R., 2007. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc. Natl. Acad. Sci. U. S. A. 104, 20600–20605.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cognit. Sci. 17, 401–412.
Kriegeskorte, N., Mur, M., Bandettini, P.A., 2008a. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28.
Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.
Laurienti, P.J., Perrault, T.J., Stanford, T.R., Wallace, M.T., Stein, B.E., 2005. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. Exp. Brain Res. 166, 289–297.
Lavan, N., 2017. Commentary: "Hearing faces and seeing voices": amodal coding of person identity in the human brain. Front. Neurosci. 11, 303.
Lavan, N., Burston, L.F.K., Garrido, L., 2018b. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. Br. J. Psychol. (published online ahead of print 16 September).
Lavan, N., Burton, A.M., Scott, S.K., McGettigan, C., 2018a. Flexible voices: identity perception from variable vocal signals. Psychon. Bull. Rev. 26 (1), 90–102 (published online ahead of print 25 June).
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88, 365–411.
McCarthy, G., Puce, A., Gore, J.C., Allison, T., 1997. Face-specific processing in the human fusiform gyrus. J. Cogn. Neurosci. 9, 605–610.
Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047–1054.
Nestor, A., Plaut, D.C., Behrmann, M., 2011. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc. Natl. Acad. Sci. U. S. A. 108, 9998–10003.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 e1003553.
Parkinson, C., Kleinbaum, A., Wheatley, T., 2017. Spontaneous neural encoding of social network position. Nat Hum Behav 1, 0072.
Parkinson, C., Liu, S., Wheatley, T., 2014. A common cortical metric for spatial, temporal, and social distance. J. Neurosci. 34, 1979–1987.
Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat vis 10, 437–442.
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.
Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., Kanwisher, N., 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56, 2356–2363.
Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107.
Quiroga, R.Q., Kraskov, A., Koch, C., Fried, I., 2009. Explicit encoding of multimodal percepts by single neurons in the human brain. Curr. Biol. 19, 1308–1313.
Rennig, J., Beauchamp, M.S., 2018. Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus. Neuroimage 183, 25–36.
Rossion, B., Hanseeuw, B., Dricot, L., 2012. Defining face perception areas in the human brain: a large-scale factorial FRMI face localizer analysis. Brain Cogn. 79, 138–157.
Schweinberger, S.R., Herholz, A., Sommer, W., 1997. Recognizing famous voices. J. Speech Lang. Hear. Res. 40, 453–463.
Shah, N.J., Marshall, J.C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H.J., Fink, G.R., 2001. The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. Brain 124, 804–815.
Verosky, S.C., Todorov, A., Turk-Browne, N.B., 2013. Representations of individuals in ventral temporal cortex defined by faces and biographies. Neuropsychologia 51, 2100–2108.
Visconti Di Oleggio Castello, M., Halchenko, Y.O., Guntupalli, J.S., Gors, J.D., Gobbini, M.I., 2017. The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. Sci. Rep. 7, 1–14.
Van Lancker, D., Krieman, J., Emmorey, K., 1985. Familiar voice recognition: patterns and parameters. Part I: recognition of backward voices. J. Phonet. 13, 19–38.
von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L., 2005. Interaction of face and voice areas during speaker recognition. J. Cogn. Neurosci. 17, 367–376.
von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biol. 4, e326.
von Kriegstein, K., Kleinschmidt, A., Giraud, A.L., 2006. Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cerebr. Cortex 16, 1314–1322.
von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.L., Kell, C.A., Grüter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. Proc. Natl. Acad. Sci. U. S. A. 105, 6747–6752.
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137, 188–200.
Watson, R., Latinus, M., Charest, I., Crabbe, F., Belin, P., 2014a. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50, 125–136.
Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., Belin, P., 2014b. Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. J. Neurosci. 34, 6813–6821.
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.
Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cerebr. Cortex 13, 1034–1043.
Yovel, G., O'Toole, A.J., 2016. Recognizing people in motion. Trends Cognit. Sci. 20, 383–395.