BRAIN AND PERCEPTUAL REPRESENTATIONS OF FACES, VOICES, AND PERSON IDENTITY

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Maria Stephanie Tsantani

Department of Psychology, Brunel University London

September 2018

I, Maria Stephanie Tsantani, declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been clearly stated in the thesis.

I declare here that a version of Chapter 3 has been published on the preprint server BiorXiv (doi: https://doi.org/10.1101/338475).

Abstract

Specialised brain regions have been identified that selectively process faces (Kanwisher, McDermott, & Chun, 1997) and voices (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000). However, little is known regarding how information from the face and voice is integrated to represent person identity. According to two distinct models, person identity representations could exist in multimodal brain regions that process both faces and voices, or in face-selective and voice-selective regions through functional connections (Campanella & Belin, 2007). This thesis tested these predictions using functional magnetic resonance imaging and representational similarity analysis to directly compare representations of familiar faces and voices in different brain regions. A representation of person identity was found in the multimodal right posterior superior temporal sulcus, providing support for the notion that face and voice information is integrated in multimodal regions.

This thesis also showed evidence of representations of face identity in face-selective regions and voice identity in voice-selective regions that could both ‘tell apart’ different identities and ‘tell together’ different, naturalistically varying tokens of each person’s face and voice (Burton, 2013). To investigate the information processed in these regions, brain representations of faces and voices were compared with multiple models of face and voice information. Face-selective regions and voice-selective regions were found to process information regarding the perceived and objective visual/auditory similarity between faces and voices, respectively. These findings provide novel insights into the computations of these regions.

Lastly, this thesis investigated the relationship between information that is perceived from the face and the voice, and how this relationship compares between familiar and unfamiliar people. Information on social traits and perceived similarity was consistent across faces and voices, and more so for familiar compared with unfamiliar people. This finding suggests that having prior semantic knowledge about a person leads to similar judgements of their face and voice. Moreover, it suggests that some concordant information may be available even in the faces and voices of unfamiliar people.

Acknowledgements

First and foremost I would like to thank my supervisor, Lúcia Garrido, for her continuous support throughout my PhD. I thoroughly appreciate all of her encouragement and her eagerness to help with everything from designing experiments, collecting data, and explaining statistical concepts, to career advice and job interview preparation, and to providing extensive feedback on this thesis. Her enthusiasm for research, her ideas, and her incredible knowledge have been a source of inspiration for me, and I am extremely grateful to have had such an amazing supervisor.

I would also like to thank my second supervisor, Adrian Williams, and my research development advisor, Survjit Cheeta, for providing me with valuable advice throughout my PhD. I am extremely thankful to our collaborators on the fMRI study presented in this thesis, Niko Kriegeskorte and Caroyln McGettigan, for their valuable input in the experimental design of the study, for developing the methods to analyse the data, and for providing feedback on the manuscript. I am also grateful to the Leverhulme Trust for funding my PhD and making it possible.

I am very thankful to Nadine Lavan, Andrew Phillips-Hird, and Emily Mitson for helping me collect data and for being great office mates. I am especially grateful to Nadine for her help with the fMRI testing and for her advice and support during stressful times. I would also like to thank Natasha Baxter, Ibtisam Abdi, and Saira Mahmood Khan for their help in finding and processing stimuli.

Last but not least, I would like to thank my family, my partner, and my friends for being there for me though all the ups and downs of the PhD. I am extremely grateful to my parents and my grandmother for believing in me enough to fund the studies that got me to this PhD. Finally, a special shout out to Bernard for being a constant source of support and encouragement, for always assuring me that “ah sure, it will be grand”, and for always managing to put a smile on my face.

Contents

List of Tables **8**

List of Figures **9**

1 Introduction **11**

1.1 Representations of face and voice identity in the brain **12**

1.1.1 fMRI-adaptation (fMR-A) studies 13

1.1.2 Multivoxel pattern analysis (MVPA) studies 16

1.1.3 Summary 31

1.2 The integration of face and voice information in representations of person identity **32**

1.2.1 Cognitive models of face and voice recognition 33

1.2.2 Multimodal Processing (MP) model 36

1.2.3 Coupling of Face and Voice Processing (CFVP) model 47

1.2.4 Summary 55

1.3 The informational content of face and voice representations **56**

1.3.1 Physical properties 56

1.3.2 Gender 63

1.3.3 Social information 67

1.3.4 Multimodal brain regions 72

1.3.5 Summary 73

1.4 The relationship between perceived information from faces and voices 74

1.5 Organisation of thesis **78**

2 Methodology: Representational similarity analysis (RSA) **81**

2.1 Introduction to RSA **81**

2.2 Representational dissimilarity matrices (RDMs) and comparing representational geometries **83**

2.3 Motivation for the use of RSA in the current thesis **86**

2.4 Applications of RSA in the current thesis **89**

2.4.1 Comparing RDMs between faces and voices (Chapter 3) 89

2.4.2 Investigating crossmodal discriminability (Chapter 3) 90

2.4.3 Comparing face and voice RDMs to model RDMs (Chapter 4) 91

2.4.4 Comparing face/voice RDMs between different brain regions (Chapter 4) 92
2.4.5 Comparing ratings of social traits with ratings of pairwise similarity (Chapter 5) 93

3 Crossmodal representations of person identity in the brain 94

3.1 Methods 96

3.1.1 Overview of study 96
3.1.2 Participants 97
3.1.3 Stimuli 98
3.1.4 MRI data acquisition and pre-processing 100
3.1.5 Functional localisers 101
3.1.6 ROI definition 103
3.1.7 Main experimental runs: Experimental design 105
3.1.8 Main experimental runs: Statistical analysis 106

3.2 Results 111

3.2.1 Familiarity ratings 111
3.2.2 ROI definition 112
3.2.4 Analysis A: RSA comparing representational geometries 116
3.2.5 Analysis B: RSA investigating identity discriminability 120
3.2.6 Exploratory whole-brain searchlight analyses 124

3.3 Discussion 125

4 The informational content of face and voice representations 133

4.1 Methods 136

4.1.1 Stimuli for the behavioural session 136
4.1.2 Procedure for the behavioural session 136
4.1.3 Representational geometries: computing RDMs 140
4.1.4 Data analysis 146

4.2 Results 149

4.2.1 RSA comparing brain and candidate RDMs 149
4.2.2 RSA comparing brain RDMs across different ROIs 155

4.3 Discussion 158

5 The relationship between perceived information from faces and voices 170

5.1 Methods 173

5.1.1 Participants 173

5.1.2 Stimuli 174

5.1.3 Procedure 174

5.2 Results 175

5.2.1 Analyses of differences in the stimuli across both experiments 175

5.2.2 Comparison between ratings of faces and voices on social judgements 177

5.2.3 Comparison between ratings of faces and voices on pairwise perceptual similarity 180

5.2.4 Reliability of face and voice ratings 182

5.2.5 Relationship between ratings of different social judgements within modality 183

5.2.6 Comparison between ratings of faces and voices on pairwise perceptual similarity and their social evaluation 187

5.3 Discussion 189

6 General Discussion 196

6.1 Summary of main findings 196

6.2 The integration of face and voice information in the brain 197

6.3 The informational content of the person identity representation in rpSTS 203

6.4 Stimulus-invariant representations of face and voice identity 205

6.5 The informational content of face and voice identity representations .. 209

6.6 The relationship between perceived information in the face and voice 214

6.7 General limitations and future directions 218

6.8 Conclusions 221

References 223

Appendix A – Ethical approval for the studies presented in Chapters 3 & 4 .. 249

Appendix B – Ethical approval for the study presented in Chapter 5 250

List of Tables

3.1 Familiarity ratings of the face and voice of each identity…………………………....111

3.2 Average correlations between brain activity patterns for faces and for voices across scanning sessions…………………………………………………………...119

3.3 One-sample *t*-test results for mean LDC values in crossmodal RDMs……......120

3.4 One-sample *t*-test results for mean LDC values in face and voice RDMs……..121

5.1 Average correlations between face and voice ratings on trustworthiness, dominance, attractiveness, and positive-negative valence, and one-sample Wilcoxon sign-rank test results for familiar people and unfamiliar people………………………………………………………………….............179

5.2 Wilcoxon rank-sum test results for the comparison of the face-voice correlations between familiar people and unfamiliar people for each task……………………………………………………………………………….179

5.3 Correlations of face and voice ratings between the two trials in which each identity was featured for familiar people and unfamiliar people…………………………………………………………………………....183

5.4 Inter-rater agreement of face and voice ratings for familiar people and unfamiliar people…………………………………………………………………………....183

5.5 Wilcoxon sign-rank test results of analysis comparing the correlations between different judgements within-modality against zero, for familiar people and matched unfamiliar people…………………………………………………………………………....185

5.6 Average correlations between ratings of face or voice identity pairs on perceptual similarity and the Euclidean distances between their ratings on trustworthiness, dominance, attractiveness, and valence, and one-sample Wilcoxon sign-rank test results for familiar people and matched unfamiliar people……………………………………………………………………………188

List of Figures

1.1 The two models of face and voice integration…………………………………...36

2.1 Hypothetical RDM……………………………………………………………………..85

3.1 Example trial sequence and stimuli for fMRI experiment…………………….…105

3.2 Face-selective, voice-selective, and multimodal ROIs…………………………..113

3.3 Regional mean responses to faces and voices in ROIs…………………………114

3.4 Results of RSA comparing representational geometries…………………….…..116

3.5 Representational distance matrix (RDM) comparisons across scanning sessions 1 and 2 in the rpSTS……………………………………………………………...…117

3.6 Results of RSA investigating identity discriminability……………………….....120

4.1 Example trial sequence and stimuli for Trustworthiness, Dominance, Attractiveness, and Valence rating tasks……………………………………….138

4.2 Example trial sequence and stimuli for pairwise auditory and visual similarity ratings tasks……………………………………………………………………..…138

4.3 RDMs showing the Euclidean distances between face images of the 12 identities based on the OpenFace program (A) and the Gabor-Jet model (B)………..…143

4.4 RDMs showing the Euclidean distances between voice recordings of the 12 identities based on measures of f0 (A) and AVTL (B)…………………………...145

4.5 Example RDMs for faces (A) and for voices (B) for one participant…………….148

4.6 Correlations between brain RDMs for faces (A) and voices (B) and candidate model RDMs……………………………………………………………………….151

4.7 Correlations between candidate RDMs for faces and for voices………………..154

4.8 Correlations between face and voice RDMs in face-selective, multimodal, voice-selective, and anatomically defined ROIs, averaged across all participants…..156

4.9 Similarity of face RDM and voice RDMs across different ROIs…………………158

5.1 Familiarity ratings of the faces and voices of familiar people and matched unfamiliar people……………………………………………………………….....176

5.2 Representational similarity matrices showing average ratings of faces on visual similarity and ratings of voices on auditory similarity for familiar people and unfamiliar people…………………………………………………………………177

5.3 Correlations between face and voice ratings on trustworthiness, dominance, attractiveness, and positive-negative valence for familiar people and unfamiliar people……………………………………………………………………………178

5.4 Relationships between the average ratings of the face and the voice of each of the 12 identities for familiar people and unfamiliar people ……………………......180

5.5 Correlations between ratings of faces on visual similarity and ratings of voices on auditory similarity for familiar people and unfamiliar people……………….…....181

5.6 Relationships between the average visual similarity ratings of face pairs and the average auditory similarity ratings of voice pairs of each of the 66 identity pairs for familiar people and unfamiliar people…………………..…………………...181

5.7 Average correlations between different judgements within each modality for familiar people and unfamiliar people…………………………………………….…..184

5.8 Pairwise correlations between ratings of trustworthiness, dominance, attractiveness, and positive-negative valence from faces and voices, in familiar people and unfamiliar people, sorted from highest to lowest…………………....186

5.9 Correlations between ratings of face or voice identity pairs on perceptual similarity and the Euclidean distances between their ratings on trustworthiness, dominance, attractiveness, and valence, for familiar people and unfamiliar people……………………………………………………………………………188

Chapter 1

Introduction

It is possible to recognise a familiar person either from looking at their face or listening to their voice. This ability is strikingly robust to distortions in the visual appearance of the face and the sound of the voice, such as the blurring of face photographs (Hole, George, Eaves, & Rasek, 2002), and the reversing of voice recordings (Lavan, Scott, & McGettigan, 2016). Importantly, irrespective of whether a person is being recognised from a photograph of their face, or from the sound of their voice through the phone, these experiences grant access to the same knowledge about the specific person: who they are, how we know them, our relationship, etc.

In addition to their role in person identification, faces and voices play a crucial role in communication and social interactions. They convey a wealth of information about a person, such as how old they are, what gender they are, and where they are from (Yovel & Belin, 2013). As well as physical characteristics, faces and voices are also used to make inferences about socially relevant qualities of a person, such as trustworthiness, attractiveness, and dominance (McAleer, Todorov, & Belin, 2014; Oosterhof & Todorov, 2008).

The first main aim of this thesis is to determine how the brain integrates information from the faces and voices of familiar people to represent person identity. A second main aim is to determine where in the brain the different types of information conveyed by faces and voices are processed. The third and final aim is to determine how information extracted from a person’s face relates to the information extracted from their voice. To address these questions, this thesis will use representational similarity analysis (RSA) (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, Mur, & Bandettini, 2008) to compare brain activity between faces and voices, and to compare brain representations of faces and voices with models of different types of face and voice information. The novel application of this method will inform the understanding of the mechanisms through which face and voice information is integrated in the brain, and of the informational content of face and voice representations.

1.1 Representations of face and voice identity in the brain

Over the past twenty years, brain regions that selectively respond to, and process information from either faces (Kanwisher et al., 1997) or voices (Belin et al., 2000) have been identified with the use of functional magnetic resonance imaging (fMRI). Face-selective and voice-selective regions are typically defined using functional localisers that contrast brain activation to faces or voices with activation in response to control visual (such as objects, scenes, scrambled faces or objects, or other body parts) or auditory stimuli (such as natural or artificial sounds from the environment, animal sounds, or musical instruments), respectively. In this thesis the terms 'face-selective' and 'voice-selective' will be used to refer exclusively to regions defined using functional localisers. For regions defined based on contrasting responses to faces or voices with baseline activity, as opposed to a control condition, the terms 'face-responsive' and 'voice-responsive' will be used instead.

The face-selective regions that are most consistently identified across different individuals and different studies are the fusiform face area (FFA) in the fusiform gyrus (Kanwisher et al., 1997), the occipital face area (OFA) in the inferior occipital gyrus (Gauthier et al., 2000; Haxby et al., 1999), and the posterior part of the superior temporal sulcus (pSTS) (Hoffman & Haxby, 2000; Kanwisher et al., 1997). Additional regions, which are less consistently identified, include the anterior inferior temporal cortex (aIT) or the broader anterior temporal lobe (ATL) (Rajimehr, Young, & Tootell, 2009), the inferior frontal gyrus (Axelrod & Yovel, 2013; Chan & Downing, 2011; Fox, Iaria, & Barton, 2009), the amygdala (Fox et al., 2009), and regions of the mid and anterior STS (mSTS and aSTS) (Fox et al., 2009; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011). While face-selective regions have been detected in both hemispheres, responses tend to be more consistent in the right hemisphere (Duchaine & Yovel, 2015).

The main voice-selective brain regions are the “temporal voice areas” (TVAs), which are located in the bilateral STS and superior temporal gyrus (STG) (Belin et al., 2000; Pernet et al., 2015; Von Kriegstein & Giraud, 2004). In a large-scale study combining data from 218 participants, Pernet et al. (2015) showed three distinct patches in the bilateral posterior, middle, and anterior TVAs. The same study found weak, albeit significant, voice-selective responses in the bilateral inferior prefrontal cortex, amygdala, and thalamus, although these responses were not consistently discernible in the brains of individual participants.

The use of fMRI has been crucial in identifying regions that selectively respond to faces and voices. However, the identification of brain regions that process face and voice identity requires the investigation of brain activity at the level of individual faces or voices, rather than at the level of stimulus categories. Standard univariate fMRI approaches, which typically compare activation between different experimental conditions after averaging activation across multiple stimuli, do not show differential responses to individually presented faces (Kriegeskorte, Formisano, Sorger, & Goebel, 2007) or voices (Formisano, De Martino, Bonte, & Goebel, 2008). In order to overcome these issues, representations of face and voice identity have been investigated using fMRI-adaptation (fMR-A) (Grill-Spector & Malach, 2001) and multivoxel pattern analysis (MVPA) (Haxby et al., 2001; Norman, Polyn, Detre, & Haxby, 2006).

1.1.1 fMRI-adaptation (fMR-A) studies

fMR-A (Grill-Spector & Malach, 2001) has been used to investigate representations of face and voice identity in face-responsive and voice-responsive regions, respectively. fMR-A is based on the observation that BOLD (blood-oxygen-level dependent) activity in a brain region that processes a certain type of stimulus decreases with repeated presentations of that stimulus (Grill-Spector & Malach, 2001). It is assumed that decreased BOLD activity reflects reduced firing of the underlying neurons (Grill-Spector & Malach, 2001). The nature of the representations in a given brain region can be tested by modifying properties of the presented stimulus (e.g. the viewpoint of a face) and comparing the magnitude of the brain response to this modification with the magnitude of the response to repeated presentations of an identical stimulus (Grill-Spector & Malach, 2001). Increases in activation in response to the modification of a stimulus property signify release from adaptation, and indicate that representations in that region do not generalise to the particular stimulus modification (e.g. the same face presented from different viewpoints is not categorised as the same face) (Grill-Spector & Malach, 2001). In contrast, if activation remains suppressed after modifying a particular stimulus property, it is assumed that representations in that region are invariant to changes in that property (Grill-Spector & Malach, 2001). Grill-Spector & Malach (2001) proposed that adaptation indicates the presence of groups of neurons that are homogenous in terms of their representational properties, whereas release from adaptation indicates that neurons within the same group represent different stimulus properties.

Adaptation studies investigating face and voice identity typically compare activation in response to the repeated presentation of face images/voice recordings featuring the same identity with activation in response to the repeated presentation of face images/voice recordings featuring different identities. The expectation is that brain regions that process face/voice identity will show a reduced response to repeated presentations of same-identity faces/voices compared with different-identity faces/voices. Adaptation to the repeated presentations of the same face images, compared with different face images, has been found in face-selective regions, namely the FFA (Andrews & Ewbank, 2004; Ewbank & Andrews, 2008; Gauthier et al., 2000; Mur, Ruff, Bodurka, Bandettini, & Kriegeskorte, 2010; Pourtois, Schwartz, Seigher, Lazeyras, Vuilleumer, 2005; Rotshtein, Henson, Treves, Driver, & Dolan, 2005; Verosky, Todorov, & Turk-Browne, 2013; Winston, Henson, & Dolan, 2004), the OFA (Gauthier et al., 2000; Ewbank & Andrews, 2008; Mur et al., 2010), and the pSTS (Winston et al., 2004). However, adaptation to the repreated presentation of the same face image may be due to adaptation to low-level properties of an image, rather than face identity per se. Thus, studies have also investigated adaptation in response to different images of the face of the same identity, and have shown reduced responses in the FFA for same-identity faces with different emotional expressions (Winston et al., 2004), for physically-different morphed faces that were perceived as having the same identity (Rotshtein et al., 2005), for same-identity faces manipulated to have different degrees of distinctiveness (Loffler, Yourganov, Wilkinson, & Wilson, 2005), for same-identity faces presented from different viewpoints (Ewbank & Andrews, 2008; Mur et al., 2010; Verosky et al., 2013; Xu & Biederman, 2010), and for different images of the face of the same identity (Davies-Thompson, Newling, & Andrews, 2013). Adaptation in the OFA has been found for same-identity images presented from different viewpoints (Mur et al., 2010; Xu & Biederman, 2010) and with different emotional expressions (Xu & Biederman, 2010), and for different images of the face of the same identity (Davies-Thompson et al., 2013). Moroever, adaptation to different images of the face of the same identity has been found in the ATL (Yang, Susilo, & Duchaine, 2016).

For voices, a handful of studies have investigated adaptation to different voice recordings of the same identity, compared with voice recordings from different identities. These studies have shown adaptation for same-identity voices speaking different syllables in the aSTS (Belin & Zatorre, 2003) and for physically-different morphed voices that are perceived as the same identity in the voice-selective bilateral mSTS, pSTS and anterior temporal pole (Andics et al., 2010) and in the right inferior frontal cortex and in the left cingulate gyrus (Latinus, Crabbe, & Belin, 2011). However, it should be noted that these studies did not restrict their analysis to regions showing voice-selective responses.

In sum, fMR-A studies have revealed brain regions that represent face identities with invariance to different images of the same face, and voice identities with invariance to different recordings of the same voice. However, a study by Mur et al. (2010) has brought into question the extent to which these invariant representations capture identity processing. Specifically, Mur et al. (2010) found evidence of adaptation to same-identity faces that varied in viewpoint or in degree of illumination in the early visual cortex. Given the known properties of the early visual cortex, which involve the processing of low-level stimulus properties (Engel, Glover, & Wandell, 1997), it is unlikely that this region would contain viewpoint-invariant face representations, let alone face identity representations. Mur et al. (2010) speculated that the adaptation effects in the early visual cortex might have been due to the carryover of activation from a connected brain region. Specifically, a higher-order brain region that shows adaptation to different images of the same face due to true representations of identity may activate connected brain regions resulting in adaptation effects in these regions (Mur et al., 2010). As a consequence, regions that are not typically involved in identity processing may show false evidence of view-invariant face identity representations, and therefore fMR-A may not be an ideal method for identifying representations of face identity (Mur et al., 2010).

Adaptation effects have also been shown to be confounded with attention when the adapting stimuli are naturally less attention-provoking than the test stimuli, and therefore elicit less activation (Huk, Ress, & Heeger, 2001). Thus, it is possible that conditions showing different face/voice identities may engage attention to a different extent than conditions showing same-identity faces/voices, and that adaptation to same-identity faces/voices does not reflect identity processing per se (Mur et al., 2010). Conversely, adaptation has also been shown for stimuli that are explicitly attended to, compared to stimuli that are not attended to, suggesting that in some cases attention to the adaptor reduces activation (Eger, 2004; Henson & Mouchlianitis, 2007). A further limitation concerning the interpretation of findings from fMR-A studies is that, due to the presence of multiple neurons in each voxel, it is possible that some of these neurons show reduced responses due to adaptation while others show enhanced responses due to other types of stimulus processing (Krekelberg, Boynton, & van Wezel, 2006). In this case, adaptation effects may be cancelled out, and as a consequence the representational properties of the brain regions in question may be miss-interpreted (Krekelberg et al., 2006). Therefore, some regions that contain invariant representations of face or voice identity may not be identifiable using fMR-A. In conclusion, the interpretations of findings of fMR-A studies on face and voice identity representations are confounded by several limitations and should be regarded with caution.

1.1.2 Multivoxel pattern analysis (MVPA) studies

MVPA (Haxby et al., 2001; Norman et al., 2006) presents an alternative method to fMR-A for investigating representations of face and voice identity, and there is evidence that it is more sensitive to detecting subtle variations in stimulus properties compared to fMR-A (Sapountzis, Schluppeck, Bowtell, & Peirce, 2010). In contrast to univariate fMRI, which spatially averages across voxels that show a significant change in activation, MVPA is a multivariate method that is based on the spatial patterns of activation across all voxels that are contained within a brain region of interest (Norman et al., 2006). MVPA is based on the notion that individual voxels within a brain region that represents a certain stimulus category (e.g. objects) will activate to different extents in response to stimuli sub-categories (e.g. chairs and shoes), forming a unique pattern of activation across the brain region in response to each sub-category (Haxby et al., 2001). Thus, MVPA investigates the discriminability of multiple sub-categories of stimuli based on their elicited response patterns. Moreover, MVPA can be used to examine the discriminability of stimulus properties, such as face identity (e.g. Nestor, Plaut, & Behrmann, 2011), or mental states, such attended and non-attended stimuli (Haynes & Rees, 2005).

The following review will focus on studies that have used MVPA to test whether individual faces or voices elicit distinct multivoxel activity patterns in face-responsive and voice-responsive brain regions, respectively. These studies typically investigate the discriminability of different face identities or voice identities based on their response patterns in brain regions of interest. The majority of studies use linear pattern classifiers or discriminants that are trained to discriminate between different identities based on a subset of the data, and then tested on their discrimination accuracy on the remaining data. If a brain region discriminates between different face identities, each identity should elicit unique response patterns that are replicable across different presentations of that identity, and that distinguish that particular identity from all other identities. Some studies have also used similarity based MVPA, such as representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), which commonly involves comparing the similarity between the response patterns to individual face identities between same-identity faces, and between different-identity faces (Verosky et al., 2013). In this approach, a brain region that contains identity representations is expected to show higher correlations between same-identity faces than between different-identity faces.

Face identity discrimination

Two studies investigated the discrimination of individual faces in face-selective regions by comparing response patterns to single images of the face of different identities (Goesaert & Op de Beeck, 2013; Kriegeskorte et al., 2007). In one of the first studies to use MVPA to investigate representations of individual faces, Kriegeskorte et al. (2007) presented participants with two grey-scale photographs of a male and a female face, which had been matched in terms of size, viewpoint, and lighting. After localising the FFA and aIT in each participant's brain, the authors created a discriminant between the two faces in each region based on one subset of the data, and then applied this discriminant to the remaining data. They found that the two faces could only be discriminated in the right aIT. The interpretation of this finding is confounded by the fact that the two faces were of different genders, and thus the representation in aIT could be reflecting the differences related to face gender. Goesaert & Op de Beeck (2013) addressed this issue by presenting faces of the same gender. In their study, the authors presented eight male artificially generated faces, and tested the classification of pairs of these faces in the FFA, OFA, and aIT. They found that all three regions were able to discriminate between the response patterns to different faces. The results of these two studies suggest that the FFA, OFA, and aIT can distinguish between different faces based on their individual response patterns elicited in these regions.

While the findings of Kriegeskorte et al. (2007) and Goesaert & Op de Beeck (2013) provide evidence that face-selective regions can distinguish between different face images, these studies were not able to show whether these regions engage in face identity or face image discrimination. Moreover, both studies presented unfamiliar faces. Behavioural studies have shown that while people can discriminate between different face images regardless of whether the faces are familiar or unfamiliar, the recognition of face identity is contingent of the ability to categorise different images of the face of the same person as belonging to the same identity (Burton, 2013; Jenkins, White, Van Montfort, & Burton, 2011). This ability to 'tell together' different images of the same face has been shown to be relatively effortless for familiar faces but highly challenging for unfamiliar faces (Jenkins et al., 2011). In a pioneering study, Jenkins et al. (2011) presented mutliple naturalistically varying images of the faces of two identities to a group of participants who were familiar with the identities, and to a group of participants who were unfamiliar with the identities, and asked them to sort the images into piles based on identity. Jenkins et al. (2011) found that while participants who were familiar with the identities were able to correctly sort the images into two piles for the two identities, participants who were unfamiliar with the identities sorted the images into between 3 and 16 piles. This is a striking example of the very limited ability to 'tell together' different images of the face of an unfamiliar person. Therefore, a distinction must be made between face *image* discrimination and face *identity* discrimination, whereby both image and identity discrimination require the ability to 'tell people apart', but identity discrimination additionally requires the ability to 'tell people together' (Anzellotti & Caramazza, 2014; Jenkins et al., 2011; Burton, 2013). The studies by Kriegeskorte et al. (2007) and Goesaert & Op de Beeck (2013) only presented one unfamiliar face image per identity, and were thus not able to test the ability of face-selective regions to 'tell people together'. Thus, these two studies provide evidence of face image discrimination, rather than identity discrimination per se.

In this thesis, I will argue that to test whether brain regions that discriminate between different faces contain representations of identity, as opposed to image-based representations, it is necessary to meet two criteria. The first criterion is that it is necessary to test whether representations in these brain regions generalise across multiple images of the face of each identity. The second criterion is that participants need to also have an adequate level of familiarity with the faces that will allow them to perceive two different images of the face of the same person as belonging to that person. A number of MVPA studies have addressed the first criterion by investigating the generalisation of face identity representations across different images of the face depicting different viewpoints (Anzellotti, Fairhall, & Caramazza, 2014; Collins, Koski, & Olson, 2016; Guntupalli, Wheeler, & Gobbini, 2017; Natu et al., 2010; Verosky et al., 2013; Visconti Di Oleggio Castello, Halchenko, Guntupalli, Gors, & Gobbini, 2017), different emotional expressions (Nestor et al., 2011), different parts of the face (Anzellotti & Caramazza, 2016), or different photographs taken on separate occasions (Axelrod & Yovel, 2015). These studies differ in regard to three main points. First*,* they differ in regard to the extent to which they fulfil the second criterion. Specifically, with the exception of Axelrod & Yovel (2015) and Visconti Di Oleggio Castello et al. (2017), who presented familiar faces, they experimentally familiarised participants with the presented faces to different extents, ranging from a brief presentation of the faces accompanied by names (Natu et al., 2010), to training participants to associate faces with biographical information (Collins et al., 2016; Verosky et al., 2013). It has been shown that the brain responds differently to familiar faces and newly learned faces (Leveroni et al., 2000), and therefore representations of face identity may differ between different levels of familiarity. Second, these studies differ in regard to the degree of generalisation of the observed face identity representations to novel images of that face. Studies that have used different images of a face identity for the training and the testing of their classifier provide stronger evidence for the generalisation of face identity representations to novel images (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Guntupalli et al., 2017), compared with studies that train and test their classifier on the same images. Third, an important difference between these studies is the degree of natural variability between the different stimuli. This varies from highly-controlled artificially-generated stimuli with low variability (e.g. Anzellotti et al., 2014), to natural photographs of faces with higher variability (Guntupalli et al., 2017).

The issue of the natural variability of face stimuli used in studies investigating face identity representations in the brain is related to the distinction between the processing of familiar and unfamiliar faces, whereby ‘telling people apart’ is possible regardless of familiarity, but ‘telling people together’ is much easier for familiar faces (Burton, 2013; Jenkins et al., 2011). This is particularly the case for different images of the face of the same person that show a naturalistic level of within-person variability, in that they are taken using different cameras on separate days, and are unconstrained in terms of lighting, pose, and expression (Burton, 2013; Jenkins et al., 2011). Studies have shown that familiar faces are markedly easier to identify than unfamiliar faces across changes in viewpoint (Bruce, 1982; Hill & Bruce, 1996), facial expression (Bruce, 1982), and lighting (Hill & Bruce, 1996), and from poor quality images or video footage (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Henderson, Bruce, & Burton, 2001). Moreover, familiar faces can be recognised despite image distortions, such as stretching and blurring (Hole et al., 2002). Therefore, the recognition of face identity involves abstracting across different, naturalistically variable images of the same familiar person (Burton, 2013; Jenkins et al., 2011). Consequently, representations of face identity in the brain should also be able to ‘tell together’ multiple, naturalistically variable images of a person’s face, while at the same time ‘telling them apart’ from other people (Anzellotti & Caramazza, 2014).

In one of the first studies to investigate the generalisation of face identity representations across different viewpoints of the face, Natu et al. (2010) presented participants with four artificially-generated male face identities, and included four images of the face of each identity, each of which showed the face from a different viewpoint. The authors only briefly familiarised participants with each identity by presenting each face image for five seconds alongside a given name for the respective identity, prior to the scanning session. They localised a visually-responsive region in the ventral temporal cortex based on activation to both faces and control stimuli in a functional localiser. They then tested whether a classifier, which was trained to distinguish pairs of identities based on all four viewpoints of each identity, could discriminate between identities in an independent subset of the data. They showed that out of the six face identity pairs, four could be accurately discriminated in the ventral temporal cortex. However, when the authors repeated this analysis using a face-selective mask of the ventral temporal cortex, they failed to find significant discrimination of any identity pairs. As discussed previously, adequate levels of familiarity with a face identity are necessary to perceptually recognise that different images of a person’s face are the same person. Given the very short familiarisation session in this study, it is likely that participants were not familiar enough with the identities to form image-invariant representations of their faces.

(Nestor et al., 2011) tested the generalisation of face identity representations across different facial expressions, after training participants to categorise unfamiliar face photographs based on identity until they reached a near-perfect level of accuracy (>98%). They included four male face identities, each of which was represented by four different colour photographs of their face displaying different facial expressions (neutral, happiness, sadness, disgust). They used spatiotemporal information-based brain mapping with a whole-brain searchlight analysis as well as a region of interest (ROI) analysis in the FFA. A classifier was trained to discriminate between the spatiotemporal patterns in response to pairs of face identities, after averaging across the different images for each identity, and was tested using a leave-one-run-out approach. They found significant discrimination in the bilateral anterior fusiform gyrus, the right anterior middle temporal gyrus, the left posterior fusiform gyrus, and in the bilateral FFA. By ensuring that participants had learned the identity of each stimulus prior to the scanning session, Nestor et al. (2011) likely increased their chance of detecting representations of face identity. However, the acquired familiarity with the face identities was entirely visual, and did not include semantic knowledge that is typically associated with familiar people. Moroever, because the classifier was trained and tested on the same face images, the authors did not demonstrate that the observed representations generalise to novel images of the same identities. Finally, although the stimuli were not artificially-generated, the photographs of each identity were taken with the same camera in the same photo session, and therefore this study did not show face identity representations that are robust to natural within-person variability.

Two studies tested the generalisation of face identity representations to different viewpoints after training participants to associate unfamiliar face photographs with biographical information (Collins et al., 2016; Verosky et al., 2013). Verosky et al. (2013) trained participants to associate 16 unfamiliar female face identities, each of whom was represented by three colour photographs of her face taken from different viewpoints, with different names. Crucially, half of these identities were also associated with short biographical descriptions that displayed either a positive or a negative tone. Participants were able to name the stimuli with a high level of accuracy (>95%) before the scanning session. RSA was used to compare correlations between multivoxel response patterns to the same identities with correlations between response patterns to different identities across different experimental runs (after averaging across the three images for each identity) in the face-selective fusiform gyrus (including the FFA) and ATL. They found that same-identity correlations were higher than different-identity correlations in the bilateral fusiform gyrus and in the right ATL. In addition, correlations between identities that had been associated with a similar amount of semantic information (name only, or name plus biographical description) were higher than the correlations between identities associated with different amounts of semantic information only in the bilateral fusiform gyrus, suggesting that face identity representations in this region are sensitive to the level of semantic information associated with an identity.

Also investigating the association of unfamiliar faces with biographical information, Collins et al. (2016) trained participants to associate eight grey-scale, front-facing photographs of unfamiliar male faces with names and four different occupations, which were in turn associated with two different locations. Training took place over several sessions, and successful identity learning was tested using recognition matching and free recall tasks. In the scanner, the authors presented participants with four novel tokens for each of the eight identities, each showing a different viewpoint of the face, and tested for classification of identity in the OFA, FFA, and ATL, after averaging across the four different viewpoints for each face. They found that only the right ATL was able to discriminate between the different identities.

To summarise, Collins et al. (2016) and Verosky et al. (2013) showed a consistent involvement of the ATL in the discrimination of face identities that had been associated with biographical information, despite using different multivariate analysis methods. However, although training participants to associate faces with biographical information may strengthen identity representations and presents a significant advancement, this level of 'learned' familiarity falls short of the level of familiarity with a face of a previously known person, whom one has been exposed to on multiple occasions over a period of time and instantly recognises even from highly variable pictures (Jenkins et al., 2011). Furthermore, Collins et al. (2016) and Verosky et al. (2013) did not demonstrate that the observed representations generalise to novel images of the same identities, and that they are robust to natural within-person variability.

Two studies have investigated face identity representations using images of familiar faces (Axelrod & Yovel, 2015; Visconti Di Oleggio Castello et al., 2017). Visconti Di Oleggio Castello et al. (2017) presented faces of four people that were personally familiar to each participant, with each person represented by three colour photographs of their face shown from a different viewpoint and taken in the same photo session. They used a searchlight approach and a pattern classifier to test the discrimination of the four identities (using regressors that included all three images of each person) across the whole brain. This revealed several regions showing significant discrimination across the brain, including the fusiform gyrus, the inferior frontal gyrus, the temporo-parietal junction, and the middle temporal gyrus/STS. Although some of these regions overlap with known face-selective regions, this study does not provide evidence of face identity representations in face-selective regions per se.

Axelrod & Yovel (2015) also presented photographs of familiar identities, and tested identity discrimination in functionally defined face-selective regions. Specifically, they presented the faces of two famous people who were familiar to their participants. Each identity was represented by eight grey-scale images of their face that were largely front facing, had neutral facial expressions, were taken on separate occasions, and were obtained though Internet searches. The images of the two identities were equated in terms of luminance and colour, and a white background was added to all images. Classification with a leave-one-session-out procedure was performed after averaging the response to all images of each identity. Significant identity classification was found only in the right FFA, and not in other face-selective regions, including the ATL and pSTS. Importantly, Axelrod & Yovel (2015) replicated this finding in a second experiment, in which they used images of two different famous people.

The studies by Visconti Di Oleggio Castello et al. (2017), and Axelrod & Yovel (2015) present significant advancements in that they used stimuli that were familiar to the participants. Moroever, Axelrod & Yovel (2015) used completely different images of the face of each identity that were taken on separate occasions, most likely using different cameras, introducing a degree of within-person variability. However, because the images were selected to have neutral facial expressions, were converted to grey-scale, and equated in terms of luminance and colour, they cannot be considered naturalistically varying. Lastly, neither of the two studies tested whether face identity representations generalise to novel tokens of the same face.

All the studies that have been reviewed so far have trained and tested their chosen pattern discrimination method using different presentations of the same face images.

Findings from these studies show brain regions that can generalise well enough across different images of the face of the same identity that they respond in a similar way to different presentations of the *same* images. However, they do not provide evidence that these brain regions respond in a similar way to the presentation of *different* images of that identity. Therefore, this method cannot exclude the possibility that the observed face identity representations are specific to the particular images of the face that were presented. Behavioural studies have shown that while training participants to recognise faces from multiple viewpoints leads to successful identification of the trained images, it does not facilitate the recognition of novel tokens of the same face (Liu, Bhuiyan, Ward, & Sui, 2009). Therefore, the observation that face identity representations generalise across different presentations of the same images of the face of a particular identity does not necessarily mean that the representations generalise to novel images of the face of that identity. Anzellotti & Caramazza (2014) discussed this same argument and concluded that, for face identity representations to be considered image-invariant a classifier must be trained and tested on different tokens of a person's face. So far, only a handful of studies have used this approach to investigate representations of face identity (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Guntupalli et al., 2017).

Anzellotti et al. (2014) investigated the generalisation of representations of face identity to different tokens of the face of the same identity in face-selective regions, a V1 (primary visual cortex) control region, and in the hippocampus. The hippocampus was included as a ROI due to the previous discovery of neurons in this region that respond preferentially to specific people with invariance to different images of the same person (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). In their experiment, Anzellotti et al. (2014) presented five male identities, each of whom was represented by five artificially-generated images of their face shown from different viewpoints. Participants were only briefly familiarised with the stimuli before entering the scanner. One of these the identities served as a target face during an identification task in the scanner and was not included in the classification analysis. The researchers performed two analyses using different methods of classification. First, they trained and tested their classifier on different presentations of the same images, and showed above-chance classification in ATL, FFA, OFA, and in V1. Given the known properties of V1, which involve the processing of low-level image features, it seems unlikely that it would contain face identity representations. These results thus demonstrate that training and testing a classifier using the same images probably does not tap into representations of face identity per se. In their second analysis, Anzelotti et al. (2014) used a more stringent classification approach that involved training their classifier on data from four images of each identity and testing it on the fifth (untrained) image. They found above-chance classification in ATL, FFA, and OFA, but not in V1, in contrast to their previous analysis. Furthermore, they showed significant classification in the bilateral hippocampus. Based on these findings, Anzellotti et al. (2014) proposed that this method of classification presents a more stringent approach to identifying representations of face identity that generalise across different images. While this method provides stronger evidence for the presence of face identity representations, as opposed to image representations, the very brief familiarisation with the stimuli brings into question the nature of these representations. It is possible that perceiving images of an artificially-generated face presented from different viewpoints as belonging to the same identity is easier compared with different pictures of real faces, taken on different days and with different cameras, and thus does not require rigorous familiarisation with the presented identities. However, a brain region that can ‘tell together’ such images may rely on the high visual similarity between the images of the same person, rather than on the recognition that these images belong to the same identity. Thus, Anzelotti et al. (2014) do not show that face identity representations generalise across naturalistically varying images of the face of each identity.

In a related study by the same group using similar methods, Anzellotti & Caramazza (2016) investigated representations of face identity that generalise across different halves of the face, as opposed to different viewpoints. Specifically, they presented participants with three artificially-generated male identities, each of whom was represented by one front-facing full-face image of their face, and four images showing different halves of their face (top, bottom, left, right). In contrast to the previous study, participants were trained to perceptually discriminate between the three identities prior to scanning based on the full-face images. As with the previous study, one of these identities served as a target face and was not included in the classification analysis. They used recursive feature elimination mapping, a method of identifying the voxels across the brain that contribute most to classification, to identify ROIs that could discriminate between the four identities based on the full-face images. Within each ROI, classification was performed by training the classifier on three different halves of the faces and testing on the fourth half. They found above change classification only in the right ATL, and not in a control V1 region. The results of Anzellotti et al. (2014) and Anzellotti & Caramazza (2016), taken together, show that the ATL forms representations of face identity that generalise across different viewpoints and different halves of the face of the same identity. They also suggest that representations in the FFA, OFA, and hippocampus only generalise across different viewpoints. However, the involvement of these regions in Anzellotti et al. (2014), but not in Anzellotti & Caramazza (2016) may be due to the different levels of familiarisation with the stimuli used in the two studies. However, although participants in Anzellotti & Caramazza (2016) were familiarised with the faces, familiarity was purely visual in nature and was not associated with any semantic information. Moreover, similar to Anzellotti et al. (2014), this study used artificially-generated stimuli with low within-person variability, and therefore does not show that identity representations are robust to natural within-person variability.

Guntupalli et al. (2017) investigated representations of face identity that generalise across different viewpoints using a similar classification method to Anzellotti & Caramazza (2016) and Anzellotti et al. (2014) (i.e. they trained and tested their classifiers on different viewpoints of the face), but used natural photographs of faces, as opposed to artificially-generated faces. Specifically, they presented participants with the faces of four unfamiliar identities (two male, two female), each of whom was represented by five different photographs of their face showing different viewpoints, taken in the same photo session. Prior to scanning participants were trained to recognise the four identities from the different images. They used a combination of several methods, including whole-brain searchlight analysis, ROI analysis in face-selective ROIs, and RSA. For the RSA, the authors created an ‘identity’ model that predicted that response patterns to face images of the same identity would be more similar to each other than to response patterns to images of different identities. They then identified representations of face identity using the searchlight and ROI analyses, and compared these representations with the predictions of the identity model. The searchlight analysis revealed a cluster in the inferior frontal cortex, whereas the ROI analysis showed above-chance classification in the FFA and ATL, in line with the findings of Anzellotti et al. (2014). In addition, RSA showed that the face representations in the inferior frontal cortex and the FFA were correlated with the identity model, although the representations in the ATL were not. The inferior frontal cortex was not originally included as a face-selective ROI, but post-hoc analyses showed that it demonstrated face-selectivity. The findings of this study highlight the need to test a wider range of face-selective regions, as opposed to focusing on the most commonly studied (FFA, OFA, ATL, and pSTS). Furthermore, this study shows that RSA can be a useful tool of testing to what extent the observed face identity representations capture similarities in the response patterns to different images of the same identity. However, Guntupalli et al. (2017) only visually familiarised participants with the presented faces, and used face photographs taken with the same camera in the same photo session. Therefore, this study does not show evidence of representations that generalise across naturalistically variable images of the same person.

To summarise, studies investigating representations of face identity that can ‘tell people together’ as well as ‘tell people apart’ have indicated the involvement of several face-selective regions. Although no region was consistently implicated across all studies, the most consistently identified regions across studies using different MVPA methods and stimuli were the FFA (Anzellotti et al., 2014; Axelrod & Yovel, 2015; Guntupalli et al., 2017; Nestor et al., 2011; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017) and ATL (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Collins et al., 2016; Guntupalli et al., 2017; Verosky et al., 2013). Specifically, these two regions were found to distinguish between different identities in studies using familiar (Axelrod & Yovel, 2015; Visconti Di Oleggio Castello et al., 2017) or unfamiliar faces (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Collins et al., 2016; Guntupalli et al., 2017; Nestor et al., 2011; Verosky et al., 2013), in studies that trained and tested their classifiers on different presentations of the same stimuli (Axelrod & Yovel, 2015; Collins et al., 2016; Nestor et al., 2011; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017) or on presentations of different stimuli (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Guntupalli et al., 2017), and in studies using artificially-generated stimuli (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014) or face photographs (Axelrod & Yovel, 2015; Collins et al., 2016; Guntupalli et al., 2017; Nestor et al., 2011; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017). A consistent and important limitation of these studies is the lack of within-person variability across the different images of the face of each identity. The most variable stimuli used were images of faces that were obtained from the Internet and taken on separate occasions (Axelrod & Yovel, 2015), but even these images were constrained to neutral facial expressions, and they were converted to grey-scale and equated in terms of luminance and colour, thus minimising within-person variability. Therefore, none of the studies provided evidence of face identity representations that are robust to natural within-person variability in images of the same person. It is an open question whether the face identity representations in the brain regions identified in these studies would generalise to novel, naturalistically varying images of the face of the same identity. Furthermore, due to the low within-person variability of the faces, the possibility that the observed representations are based on the visual similarity between the different images of a person's face, as opposed to the recognition of these images as being the same person, cannot be ruled out.

Voice identity discrimination

Only two studies have investigated representations of voice identity, i.e. speaker identity, using MVPA (Bonte, Hausfeld, Scharke, Valente, & Formisano, 2014; Formisano et al., 2008). Formisano et al. (2008) presented participants with three unfamiliar speakers (two male, one female), each of whom was represented by nine separate recordings of their voice, in which they were vocalising three different vowel sounds (/a/, /i/, and /u/). Participants were familiarised with the stimuli prior to scanning and were able to identify the speakers from the different recordings. A univariate analysis of responses to voices showed activation in a broad region of the STS/STG. A classifier was then trained to discriminate between different speakers based on one subset of the data and tested on the remaining subset. This analysis revealed above-chance classification of speaker identity in the STS/STG and primarily in the right Heschl's gyrus/sulcus and along the right STS. To test whether speaker identity representations would generalise to different voice recordings, the authors conducted a further analysis in which different voice recordings (i.e. different vowels) were used to train and test the classifier, and found accurate classification within the regions identified in the previous analysis. These findings present evidence for representations of speaker identity that generalise across different tokens of a speaker's voice in voice-responsive regions of the STS/STG. This study addresses an important limitation of many studies investigating face identity representations, which did not test whether identity representations generalised to novel tokens of the same person's face (Axelrod & Yovel, 2015; Collins et al., 2016; Nestor et al., 2011; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017).

A second study by the same group investigated the influence of task demands on representations of voice identity (Bonte et al., 2014). They presented participants with three unfamiliar speakers (adult male, child male, child female), each of whom was represented by six separate recordings of their voice, in which they were vocalising three different vowel sounds (/a/, /i/, and /u/). As with the previous study, participants were familiarised with the stimuli prior to scanning. During scanning participants performed a delayed-match-to-sample task based on either voice identity or vowel type. A functional localiser was used to define a sound-responsive ROI and a voice-selective ROI in the STS/STG, and a classifier was trained and tested on subsets of data from within these regions. Accurate classification in both the sound-responsive and voice-selective ROIs was found only for trials in which the task involved identity processing, i.e. matching a voice to an image of an identity, as opposed to speech processing, i.e. matching voice to an image of a vowel type in written form. This finding shows that voice identity representations are also present in voice-selective regions, but suggests that the context within which a voice is being processed influences the representations. However, in contrast to Formisano et al. (2008), Bonte et al. (2014) did not use different voice recordings for the training and testing of their classifier, and therefore do not demonstrate that representations in voice-selective regions generalise to novel tokens of a person's voice.

Similar to many studies investigating face identity representations, both Formisano et al. (2008) and Bonte et al. (2014) presented unfamiliar stimuli. Much less is known regarding the distinction between familiar and unfamiliar voice processing. A recent study used the identity sorting paradigm that Jenkins et al. (2011) used with faces (described previously) to investigate the influence of familiarity on the ability to 'tell together' different recordings of the voice of the same person (Lavan, Burston, & Garrido, 2018). Specifically, the authors presented multiple voice recordings for two different identities to a group of participants who were familiar with the identities, and to a group of participants who were unfamiliar with the identities. Participants were required to 'sort' the voice recordings, presented as icons on a computer screen, into piles based on identity. Similar to Jenkins et al. (2011), the authors showed that familiar participants were better at categorising the voices based on identity than the unfamiliar participants, with familiar participants creating between three and four identity piles, and unfamiliar participants creating between four and nine identity piles. This study suggests that, similar to findings for faces (Jenkins et al., 2011), it is easier to 'tell people together' from their voices when those people are familiar, compared to when they are unfamiliar. Therefore, it is likely that the use of unfamiliar voice identities in the studies of Formisano et al. (2008) and Bonte et al. (2014) limited their ability to identify voice identity representations that 'tell together' different tokens of the voice of the same person. Lastly, both Formisano et al. (2008) and Bonte et al. (2014) presented recordings of different vowel sounds, which cannot be considered representative of the natural variability that is present in different exposures to a person's voice in everyday life (Lavan, Burton, Scott, & McGettigan, 2018). Therefore, it is not known whether identity representations within the STS/STG would generalise to more naturalistic voice stimuli with longer durations, such as words and sentences.

1.1.3 Summary

A multitude of studies have used fMR-A and MVPA to investigate representations of face and voice identity in face-responsive and voice-responsive regions. MVPA studies have brought substantial advantages for these attempts, allowing the investigation of the ability of different brain regions to distinguish between different face or voice identities based on the multivoxel activity patterns elicited by each individual identity. However, the observed representations of face and voice identity have been largely based on face and voice stimuli with substantially lower within-person variability than the faces and voices that the brain typically processes in everyday life. Therefore, these studies fail to provide evidence of face and voice identity representations that are robust to the natural variability that is present in different encounters with the face and voice of the same person.

The majority of studies have investigated face and voice representations separately, and therefore little is known on how information from the faces and voices of familiar people is integrated to form representations of person identity. The next section addresses this issue by reviewing current evidence supporting different mechanisms of face and voice integration in the brain.

1.2 The integration of face and voice information in representations of person identity

Despite significant advances in identifying representations of either face or voice identity in the brain, there is still a limited understanding of how the brain combines and integrates information from these two modalities to represent person identity. Throughout this thesis, the terms ‘multimodal’ and ‘unimodal’ will be used to refer to brain regions that process information from multiple sensory modalities, and brain regions that process information primarily from one sensory modality, respectively. The terms ‘multisensory’ and ‘unisensory’ will be used to refer to single experimental conditions or neurons that engage multiple senses, or that engage primarily one sense, respectively. The term ‘crossmodal’ will be used to refer to brain representations that combine information from multiple modalities. It should be noted that the terms ‘multimodal’, ‘multisensory’, and ‘crossmodal’, as well as the terms ‘unimodal’ and ‘unisensory’, are frequently used interchangeably in the literature (Calvert, 2001). The distinctions made here are therefore largely arbitrary and serve the purpose of clarification.

This section will first describe predictions from cognitive models of face and voice recognition regarding the integration of face and voice information to form representations of person identity. These predictions form the basis of two separate theoretical models on how faces and voice information is integrated in the brain. The evidence supporting each of these two models will then be reviewed.

1.2.1 Cognitive models of face and voice recognition

Cognitive models of face and voice recognition have included proposals about how face and voice information is integrated in representations of person identity. In their face recognition model, Bruce & Young (1986) proposed that information about familiar people is stored in a "person identity node" (PIN), which is activated when presented with information relating to this person regardless of the input modality. Specifically, Bruce & Young (1986) proposed that the PIN is a multimodal component of associative memory that can be activated by the face, voice, name, and even objects that are associated with a particular person. Activation of the PIN signals the recognition of person identity. In this model, face recognition processes are distinct from the identity recognition processes of the PINs and take place in specialised "face recognition units" (FRUs), which are only activated when presented with a face. Like PINs, there is one FRU for each known person. FRUs contain stored information about a familiar person's facial appearance, and exchange information with the corresponding PINs. According to Bruce & Young (1986), while activation of the FRU results in the recognition of a face as familiar, face identification is contingent on the activation of the PIN. Burton, Bruce, & Johnston (1990) further proposed that more than one person-specific PIN can be attached to the same "piece" of semantic information (e.g. the same first name or occupation). The face recognition model proposed by Bruce & Young (1986) has been extended to include "voice recognition units "(VRUs), which are proposed to be analogous to the FRUs and are involved in familiar voice recognition (Belin, Fecteau, Bédard, & Bedard, 2004; Campanella & Belin, 2007; H. D. Ellis, Jones, & Mosdell, 1997; Schweinberger, Herholz, & Stief, 1997; Stevenage, Hugill, & Lewis, 2012; Yovel & O'Toole, 2016). According to these extended models, PINs exchange information with both FRUs and VRUs.

Studies that proposed models that include both face and voice recognition differ in regard to their predictions of when and how face and voice information is integrated (Gainotti, 2014). H. D. Ellis et al. (1997) proposed that information from FRUs and VRUs is integrated at the level of the PIN. To corroborate this prediction, the authors presented evidence of crossmodal priming, i.e. the priming of the face of a familiar person by their voice and vice-versa, at very short time intervals (0.5 sec) between the presentation of the prime and target. Specifically, they presented primes and targets that were either the face and voice of the same person, or the face and voice of two different people. H. D. Ellis et al. (1997) showed that response times in a familiarity task were faster for targets that were primed by a stimulus of the same identity, compared with targets that that were primed by a stimulus of a different identity. The authors suggested that when a PIN for a particular person is activated through one modality it is activated faster when it is subsequently presented with the other modality within a short period of time. Moreover, they found no evidence of crossmodal priming at longer time intervals (10 min) between the presentation of the prime and target, and this was attributed to the PINs being "re-set" by the exposure to other stimuli during the prime-target time interval. Thus, H. D. Ellis et al. (1997) concluded that the activation of the PIN is required for face and voice integration, and that face and voice recognition take place independently until the PIN stage. Evidence of crossmodal priming for familiar faces and voices was also shown in a more recent study, which revealed that priming effects, i.e. reduction in response times, are stronger for voices primed by faces, than for faces primed by voices (Stevenage et al., 2012). Similar to Ellis et al. (1997), Stevenage et al. (2012) suggested that face and voice recognition are facilitated by the activation of the PIN by the other modality. Furthermore, the authors speculated that the face recognition pathway may be more strongly connected to the PIN than the voice recognition pathway, and therefore recognition is facilitated more by the other modality for voices compared with faces. However, the time interval between the presentation of the primes and the targets in this study varied across participants, and therefore their results cannot be interpreted based on this factor.

Other studies have indicated that FRUs and VRUs exchange crossmodal information with each other independently of the PINs (O'Mahony & Newell, 2012;

Schweinberger, Herholz, & Stief, 1997). Early evidence for this prediction comes from a study that showed crossmodal face-voice priming (voice-face priming wasn't tested) with a 10-minute time interval between the presentation of the prime and the target (Schweinberger, Herholz, & Stief, 1997). Specifically, the authors showed faster responses in a voice familiarity task for familiar voices that had been primed by faces compared with voices that had not been primed. This design is similar to that of H. D. Ellis et al. (1997), and contradicts their finding of no crossmodal priming at longer prime-target time intervals. Furthermore, Schweinberger et al. (1997) tested the priming of voices by names, and failed to find evidence of name-voice priming. Based on these two findings, Schweinberger, Herholz, & Stief (1997) suggested that face-voice crossmodal priming effects are unlikely to involve the PINs, which, given their proposed multimodal properties, should also support name-voice priming. Instead, they speculated that crossmodal priming effects may involve an earlier processing stage, in which faces directly activate stored representations of the voices of familiar people in VRUs. In line with this proposal, a more recent study compared response times to face-voice pairs, or face-name pairs, between identity-congruent and identity-incongruent pairs, and showed a positive effect of congruency on responses to face-voice pairs, but not face-name pairs (O'Mahony & Newell, 2012). O'Mahony & Newell (2012) proposed that facial and vocal information is integrated prior to the PIN stage, whereas face information and name information is only integrated at the level of the PINs. The authors speculated that perceptual face and voice information is integrated in memory because exposure to a person's face and voice is usually concurrent, and this exposure usually takes place before acquiring semantic knowledge about a person, such as their name. In sum, these two studies show evidence that FRUs and VRUs directly exchange information prior to the PIN stage.

The two different positions regarding the stage at which face and voice identity information is exchanged and integrated are the basis of two theoretical models of face and voice integration in the brain (Blank, Anwander, & von Kriegstein, 2011; Campanella & Belin, 2007; Yovel & O'Toole, 2016). The first model, which will be referred to as the *Multimodal Processing (MP)* model, proposes that information from face and voice identity processing systems in integrated in multimodal nodes associated with semantic person identity information (e.g. Shah et al., 2001). The second model, which will be referred to as the *Coupling of Face and Voice Processing (CFVP) model*, proposes that face and voice identity processing systems exchange crossmodal information at earlier, perceptual processing stages (e.g. von Kriegstein, Kleinschmidt, Sterzer, Giraud, et al., 2005). Within the framework of these two models, researchers have attempted to link computations from cognitive models of face and voice recognition to the processing of face, voice, and person identity in the brain. Specifically, the MP model predicts that face and voice identity information is integrated in multimodal brain regions that represent person identity, and roughly correspond to the PINs (Figure 1.1). The CFVP model predicts that face and voice information is combined at an earlier processing stage by means of direct functional and structural connections between face- and voice-responsive regions, which roughly correspond to the FRUs/VRUs (Figure 1.1). These models are not mutually exclusive, but they make different predictions about how information about faces and voices is integrated in the brain. The following section describes the evidence supporting the MP and the CFVP models separately.

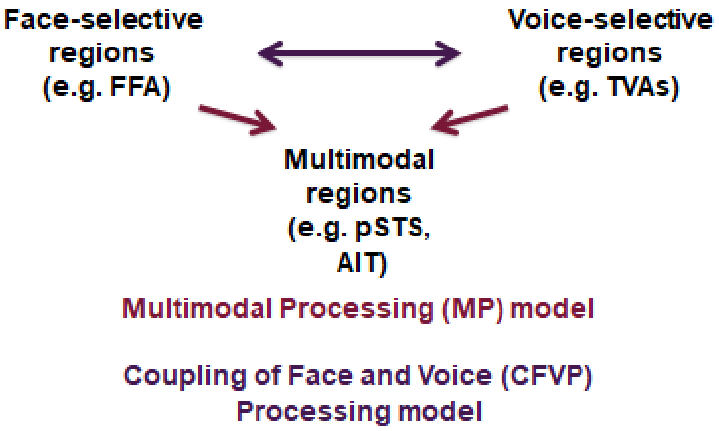

Face-selective regions (e.g. FFA)
Voice-selective regions (e.g. TVAs)
Multimodal regions (e.g. pSTS, AIT)
Multimodal Processing (MP) model
Coupling of Face and Voice (CFVP) Processing model

Figure 1.1: **The two models of face and voice integration**. Predictions of the MP model and CFVP model regarding the integration of face and voice information in the brain.

1.2.2 Multimodal Processing (MP) model

<u>Lesion studies</u>

Some of the earliest evidence supporting the existence of multimodal brain regions that process information from both faces and voices comes from studies of patients with brain damage. There have been several reports of patients with lesions to the right ATL that show impairments in the recognition of both familiar faces and familiar voices. A. W. Ellis, Young, & Critchley (1989) presented a single case study of a woman (KS) who was treated for epilepsy with a right ATL lobectomy, which included the hippocampus and amygdala. After the lobectomy, KS showed a selective memory deficit for familiar people, characterised by extreme difficulty in recognising famous faces, voices, and names, and in recalling information relating to familiar people. This study suggested that the right ATL region is involved in the processing of identity-related information from multiple modalities, including faces and voices. The importance of the right ATL in face and voice recognition was also highlighted in a study that reviewed multiple case studies of patients who had lesions that included this region, and who had been tested on both face and voice recognition abilities (Gainotti, 2011). Out of 15 patients, 12 were impaired in both face and voice recognition. However, the lesions of some of these patients were not confined to the ATL and included regions such as the STG, STS, fusiform gyrus, frontal lobe, insula, hippocampus, and amygdala. Therefore, the ATL cannot be causally implicated in face and voice recognition impairments, but it seems likely that it plays some role in both face and voice processing.

Other studies have shown evidence of co-occurring face and voice recognition deficits in patients with lesions in broader areas of cortex. Hanley, Pearson, & Young (1990) presented a case study of a woman (ELD) who had right cerebral damage, mainly in the frontotemporal region, caused by a rupture of a middle cerebral artery aneurism. After her illness, ELD showed a deficit for the learning of new visual forms, such as unfamiliar faces and objects, as well as a deficit in unfamiliar voice learning (familiar voice recognition was not tested). She also showed impaired recognition of famous faces, but not names, for people that became famous after her illness. Although ELD’s deficit was not specific to people, or selective to the processing of person identity, her difficulties with the processing of both faces and voices indicate that multimodal processing may take place in right frontotemporal regions. However, due to the extent of the brain damage, it is not possible to determine whether a single region is responsible for both face and voice processing, or whether multiple regions within the damaged area independently process faces and voices.

Neuner & Schweinberger (2000) tested a group of patients with brain damage on various tests involving face, voice, name, and object processing. They found that four patients had impairments in both face and voice recognition, three of which also showed impaired name recognition. However, none of these patients showed a selective deficit in person (i.e. face, voice, and name) recognition. Two patients had damage only to the right hemisphere, and two to both hemispheres. This study provides evidence of the co-occurrence of deficits in face- and voice-recognition in patients with brain damage. However, due to the lack of information on the precise location of the brain damage, this study does not implicate specific multimodal brain regions.

A recent study used voxel-based lesion symptom mapping to identify relationships between lesion locations and behavioural measures of voice processing abilities in a large sample of 58 patients with unilateral lesions (Roswandowitz, Kappes, Obrig, & von Kriegstein, 2018). Although the authors were primarily interested in impairments in the processing of voice identity, they also included behavioural measures of face recognition ability. Specifically, they tested voice recognition from newly-learned and familiar voices, and face recognition from newly-learned faces. Their results showed that lesions to the inferior parietal lobe were associated with both voice and face recognition deficits. Moreover, inferior parietal lobe lesions were implicated during the recognition of voices that had been paired with faces during a learning phase. Taken together, these findings suggest the inferior parietal lobe as a candidate multimodal region for the integration of face and voice identity information.

To summarise, lesion studies have associated impairments in both face and voice identity processing with damage to brain regions such as the ATL and the inferior parietal lobe. These brain regions are likely to be multimodal, and could contain representations of person identity. The main limitation of lesion studies, particularly earlier studies, is that lesions often involve large areas of cortex that may contain functionally distinct sub-regions. As a result, it is not possible to determine whether damage to a single, multimodal brain region caused the observed deficits in face and voice recognition, or whether multiple regions that selectively process either face or voices were equally damaged. The next subsection presents evidence from neuroimaging studies, which complement the findings from lesions studies by identifying regions that are functionally involved in the integration of face and voice identity information in healthy participants.

Neuroimaging studies

Studies using fMRI have presented evidence of possible multimodal brain regions that integrate face and voice information into representations of person identity. In one of the first fMRI studies to investigate face and voice identity processing in the same experiment, Shah et al. (2001) compared brain activity in response to familiar faces and voices with brain activity in response to unfamiliar faces and voices. More specifically, the authors presented participants with photographs of the faces and short (1 to 1.1sec) recordings of the voices of personally familiar people (friends and relatives) and of unfamiliar people. Although participants performed unrelated tasks, they were instructed to try to identify each stimulus. They found that the bilateral posterior cingulate, and specifically the retrosplenial cortex, was more strongly activated in response to familiar compared with unfamiliar faces and voices. Shah et al. (2001) concluded that the retrosplenial cortex is multimodal region that processes both face and voice information. However, the authors acknowledge that it is not possible to determine whether this region is involved in the processing of stimulus familiarity or person identity. It has since been shown that the retrosplenial cortex also responds more to familiar objects and places, compared to unfamiliar objects and places (Sugiura, Shah, Zilles, & Fink, 2005). Therefore, it is more likely that the retrosplenial cortex is a multimodal region that is sensitive to familiarity, rather than to person-specific identity information.

Joassin, Pesenti, et al. (2011) compared brain activity in response to faces and voices presented separately with brain activity in response to faces and voices presented simultaneously. They trained participants to associate unfamiliar face photographs, voice recordings, and names with four different identities. During the experiment, participants were presented with faces, voices, and face-voice pairs, and had to decide which of two names corresponded to each stimulus. The authors used the supra-additive criterion to identify multimodal regions, which means that for a region to be considered multimodal, activation in response to face-voice pairs had to be larger than the sum of activation in response to the faces and voices presented in isolation (Calvert, Campbell, & Brammer, 2000). This analysis revealed greater activation to the face-voice condition in the bilateral fusiform gyrus and superior temporal gyrus, the right hippocampus, and the left calcarine sulcus and angular gyrus. Out of these regions, only the hippocampus and angular gyrus were activated mainly during the face-voice condition, whereas the other regions were also activated during the unimodal face and voice conditions. A subsequent functional connectivity analysis showed that the hippocampus had strong connections to the face-responsive right FFA and voice-responsive STS. Joassin, Pesenti, et al. (2011) proposed that the hippocampus is involved in the association of face and voice information in memory, and in the recall of these associations. They also speculated that the left angular gyrus is involved in the division of attention between different modalities in response to multimodal stimuli. However, due to the use of unfamiliar face and voice stimuli in this study and the inducement of experimentally-learned associations between them, it is not possible to determine whether the aforementioned brain regions responded to the face-voice stimuli because they recognised that the stimuli belonged to the same identity, or because participants had learnt to associate them with each other. Learnt associations can be formed between completely arbitrary stimuli. Therefore, it is not clear whether the hippocampus and angular gyrus contain representations of person identity, or whether these regions are involved in associative memory more generally.

Hölig, Föcker, Best, Röder, & Büchel (2017) used a congruency manipulation paradigm to investigate the integration of face and voice information. They trained participants to associate silent videos of speaking faces, recordings of voices, and written names for twelve unfamiliar identities. In their experiment, the authors presented participants with pairs of a face and a voice, presented successively, that either belonged to the same identity or to different identities. Contrasting brain responses to identity-incongruent face-voice pairs with brain responses to identity-congruent face-voice pairs, Hölig et al., (2017) found increased activation in the right pSTS and angular gyrus. The authors proposed that viewing the faces activated stored representations of the associated voices in the right pSTS and angular gyrus, and when the subsequently presented voices did not match the activated representations brain activity increased due to the violation of expectations. However, similar to Joassin, Pesenti, et al. (2011), Hölig et al., (2017) presented faces and voices that were associated through learning, and therefore their findings could also be interpreted as the processing of learnt associations between faces and voices, as opposed to the processing of identity. Therefore, the possibility that the pSTS and angular gyrus are involved in associative memory, rather than identity processing, cannot be ruled out in this study.

The studies by Shah et al. (2001), Joassin, Pesenti, et al. (2011), and Hölig et al., (2017) all used univariate fMRI methods to investigate the integration of face and voice identity information in the brain. However, as was discussed previously, standard univariate fMRI is not optimal for detecting identity representations, because responses to individual identities cannot be differentiated from each other (Formisano et al., 2008; Kriegeskorte et al., 2007). Shah et al. (2001) compared activation in response to familiar and unfamiliar faces and voices, and revealed brain regions that are sensitive to familiarity independently from modality. However, familiarity processing does not necessarily imply identity processing, let alone the integration of face and voice information. Joassin, Pesenti, et al. (2011) used the supra-additive criterion, which was originally developed to investigate audiovisual integration (Calvert et al., 2000). This criterion states that for a brain region to be considered multimodal, it must show greater brain activity in response to audiovisual stimuli that to the sum of the activation to the auditory and visual stimuli presented in isolation (Calvert et al., 2000). This approach is based on the notion that voxels that respond more during multisensory input that during individual unisensory inputs combined are likely to contain integrative multisensory neurons, as opposed to intermixed populations of unisensory neurons, which would also be activated by the unimodal conditions (Calvert, 2001). However this approach may not reveal voxels that contain both multisensory neurons and unisensory neurons, particularly if the multisensory neurons are outnumbered (Laurienti, Perrault, Stanford, Wallace, & Stein, 2005). Moreover, it has been pointed out that this approach would not reveal multisensory neurons that display sub-additive, rather than super-additive responses to multisensory conditions (Laurienti et al., 2005). The use of the supra-additive criterion has further been criticised because of the possibility that saturation of the BOLD signal in response to one or both of the unisensory conditions could conceal signal related to the multisensory condition, leading to false negatives (Goebel & Van Atteveldt, 2009). Therefore, findings based on the supra-additive criterion should be interpreted with caution. An alternative to the supra-additive criterion is offered by congruency manipulations, demonstrated in the study by Hölig et al., (2017). Congruency manipulations are based on the assumption that, if a brain region distinguishes between congruent and incongruent pairs of stimuli, it is able to integrate information from the two stimuli (Goebel & Van Atteveldt, 2009). However, it is possible that certain brain regions respond to incongruence regardless of the modalities the stimuli presented (Laurienti et al., 2005). Therefore, higher responses to incongruent stimuli may not be due to crossmodal integration, but a reaction to incongruence per se.

Two recent studies used MVPA to investigate representations of person identity from faces and voices using crossmodal classification (Anzellotti & Caramazza, 2017; Hasan, Valdes-Sosa, Gross, & Belin, 2016). As discussed previously, the advantage of using MVPA over univariate fMRI methods is that it tests the ability of brain regions to distinguish between different conditions based on the pattern of response that they elicit across multiple voxels. To identify regions that integrate face and voice information, Anzellotti & Caramazza (2017) and Hasan et al. (2016) tested whether the multivoxel activity patterns in response to individual identities in one modality could be distinguished by a classifier that had been trained to distinguish between the same identities in the other modality. The idea behind this approach is that a brain region that represents person identity independently of modality should respond in a similar way to the face and voice of the same person.

Hasan et al. (2016) presented participants with four personally-familiar identities (two male, two female), each of whom was represented by one short (400ms) audio-visual video of their face voicing the word “had”, one muted version of the same video, and one voice recording extracted from the video. Participants performed a 4-

way identification task. To identify regions that process person identity, Hasan et al. (2016) used a whole-brain searchlight and trained a classifier to discriminate between the multivoxel response patterns to the four identities in one modality, and tested the accuracy of the classifier on data from the other modality (data from the audio-visual condition was not used in this analysis). This analysis revealed clusters in a number of brain regions that could discriminate between faces based on information from voices, and vice versa: the bilateral STS and middle temporal gyrus, the right inferior temporal gyrus, and the left inferior frontal gyrus. Hasan et al. (2016) proposed that these regions contain crossmodal representations of person identity. However, the presentation of a single token of each person's face and voice that were derived from the same audio-visual clip, which was also presented during the experiment, raises the possibility that the observed crossmodal decoding was due to learned associations between specific face and voice stimuli (Lavan, 2017). Specifically, these learned associations may have resulted in similar responses to the face and voice of the same identity not because of the fact that they belong to the same person, but because they were presented simultaneously in the audio-visual condition. Moreover, Hasan et al. (2016) did not independently test whether the brain regions revealed by the crossmodal classification were multimodal, i.e. responsive to both faces and voices. Therefore, although the authors speculated that these regions were multimodal, it is also possible that the observed crossmodal representations were the result of a coupling mechanism between face-responsive and voice-responsive regions, such as the one proposed by the CFVP model (von Kriegstein et al., 2005).

Anzellotti and Caramazza (2017) also conducted an MVPA study that tested crossmodal classification between familiar faces and voices. In their study, the authors presented participants with three famous men, each of whom was represented by two grey-scale front-facing images of their face featuring a neutral facial expression and two recordings of their voice speaking different words. The face images were equated in luminance and contrast and cropped to an oval. Participants were asked to respond to stimuli of a target identity, and only data from the two remaining (distractor) identities were used for classification. The authors used a face localiser and a voice localiser to identify face- and voice-selective regions, respectively. For each of these regions, they then trained a classifier to distinguish between the two identities in one modality, and tested it on the other modality. This analysis revealed significant crossmodal classification in the right pSTS, which showed both face-selective and voice-selective responses in a face and voice localiser, respectively. This finding suggests that this region is multimodal, and provides support to the prediction of the MP model that face and voice information is integrated in multimodal brain regions. In contrast to Hasan et al. (2016), who based their classification on data from one face and voice token for each identity, Anzellotti & Caramazza (2017) used two different tokens of the face and voice of each identity, thus providing stronger evidence of person identity processing (as opposed effects of learned associations between specific face and voice stimuli). However, they did not test whether unimodal representations of face and voice identity in the pSTS generalise across the two different face and voice tokens within the same study, possibly due to the insufficient number of tokens for this type of analysis.

Although Anzelotti and Caramazza (2017) could not determine whether the multimodal pSTS could also discriminate between identities just from the faces or just from the voices, the authors re-analysed the data from a previous study in which they tested the classification of face identities across different viewpoints (Anzellotti et al., 2014), but in which the pSTS had not been originally defined. The authors aimed to test whether the pSTS also showed face identity representations that generalised across different images of the same person's face. This new analysis showed viewpoint-invariant representations of face identity in the pSTS. Therefore, although it was not possible to determine that the crossmodal pSTS region identified by Anzelotti and Caramazza (2017) represented just face identities or just voice identities invariantly, these preliminary results suggest that this may be the case. It should be noted, however, that the stimuli in Anzellotti et al. (2014) were artificially-generated, and therefore showed low within-person variability between different images of the face of the same person. Therefore, findings based on this study do not show that representations of face identity in the pSTS are robust to natural within-person variability.

The crossmodal classification method used in Hasan et al. (2016) and Anzellotti & Caramazza (2017) presents a novel way of identifying brain regions that represent person identity. However, these studies present similar limitations to the previously reviewed studies investigating face and voice identity representations. Specifically, they show no evidence (Hasan et al., 2016) or limited evidence (Anzellotti & Caramazza, 2017) of generalisation of face and voice identity representations to different tokens of the face and voice of the same person, within the regions in which they observe person identity representations. Moreover, although Anzellotti & Caramazza (2017) presented two photographs of the face, and two recordings of the voice of each person, face photographs were processed to show low natural variability, and the voice recordings comprised a single spoken word. Therefore, it is not known whether the observed person identity representation in the pSTS would tolerate more naturalistic within-person variability across different tokens of the same person's face and voice.

Neurophysiological studies

Studies that conducted single-unit recordings in patients with epilepsy have shown evidence that structures in the medial temporal lobe, including the hippocampus, parahippocampal gyrus, amygdala, and entorhinal cortex, engage in multimodal person representation (Quiroga, Kraskov, Koch, & Fried, 2009; Quiroga et al., 2005). Specifically, Quiroga and colleagues (2005; 2009) showed that single neurons located in the medial temporal lobe respond to the face, spoken name, and written name of the same persons (Quiroga et al., 2009). These findings support neuroimaging findings showing the hippocampus as a candidate multimodal region (Joassin et al., 2011). However, responses to voices were not assessed in these studies, and therefore it is not known whether the medial temporal lobe is involved specifically in the integration of information from faces and voices.

Conclusions

Evidence in favour of the MP model comes from lesion, neuroimaging, and neurophysiological studies that have revealed a number of different brain regions that may be involved in multimodal face and voice identity processing. However, there is a lack of agreement between studies using different methods regarding the regions involved. fMRI studies have proposed a number of brain regions that may integrate face and voice information to form representations of person identity, such as the retrosplenial cortex (Shah et al., 2001), hippocampus (Joassin et al., 2011), angular gyrus (Hölig et al., 2017; Joassin et al., 2011), pSTS (Anzellotti & Caramazza, 2017; Hölig et al., 2017), and STS, middle temporal gyrus, inferior temporal gyrus, and inferior frontal gyrus (Hasan et al., 2016). Notably, these regions do not include the ATL, which has consistently been implicated in the lesion studies with patients with face and voice identity recognition impairments that were reviewed previously. A meta-analysis of neuroimaging studies in healthy participants that compared responses between familiar and unfamiliar faces, voices, or names, showed a consistent involvement of the ATL in the processing of both personally-familiar and famous-familiar stimuli (Blank, Wieland, & von Kriegstein, 2014). However, the same study failed to find consistent involvement of the ATL, or any other brain region, in studies investigating face, voice, and name recognition. Therefore, it is possible that the ATL is primarily involved in the processing of person familiarity, rather than person identification per se. The majority of patients with right ATL lesions who show deficits in face and voice recognition also show impairments in face familiarity tasks, and it may be that intact familiarity processing is a necessary pre-requisite for person identity recognition (Gainotti, 2011). However, an important point to consider is that localising the ATL using standard fMRI sequences is challenging due to the low signal-to-noise ratio in this region, particularly in the ventral part (Axelrod & Yovel, 2013). Therefore, it is also possible that the absence of the ATL in some of the fMRI studies investigating person-recognition is due to methodological issues.

Finally, lesion, neuroimaging, and neurophysiological studies each have their own limitations. For lesion studies, the often extensive damage does not allow inferences to be made regarding a specific, functionally-independent brain region. For neuroimaging studies that presented unfamiliar faces and voices, which participants learned to associate through training, it is possible that the regions implicated in multimodal processing are involved in the processing of stimulus associations rather than person identity. MVPA studies using crossmodal classification present a significant advancement, but have been limited in regard to their ability to show that person identity representations generalise across different tokens of the face and voice of each identity, and are robust to within-person variability in the face and voice. Lastly, neurophysiological studies present strong evidence of multisensory neurons, but did not test whether these neurons also respond to voices. The next section presents evidence supporting the CFVP model, which proposes that face- and voice-responsive regions exchange crossmodal information independently from multimodal regions.

1.2.3 Coupling of Face and Voice Processing (CFVP) model

Neuroimaging studies

The main evidence in support of the CFVP model comes from fMRI studies that showed crossmodal interactions between face-responsive and voice-responsive brain regions during voice recognition. In one of the first studies to investigate the crossmodal effects of voice identity processing in the brain, von Kriegstein et al., (2005) showed that the face-selective fusiform cortex is activated during the recognition of familiar voices. In their study, the authors presented participants with 47 voice recordings of sentences spoken by 14 familiar and 14 unfamiliar speakers. Participants performed either a speaker identity recognition task, in which they had to identify a target speaker irrespective of speech content, or a speech recognition task, in which they had to identify a target sentence irrespective of speaker. The same voice stimuli were presented in both tasks. Their results revealed activation in the bilateral fusiform cortex only during the identity recognition task, and only for familiar voices. A face-localiser task showed that this fusiform cortex activation either overlapped with, or was in close proximity to, the FFA in each participant’s brain. Moreover, a functional connectivity analysis of the fusiform cortex region showed an interaction with the voice-responsive bilateral middle/anterior STS during familiar voice recognition. The activation of a face-selective region during voice recognition, and the observed functional connectivity between this region and the voice-responsive STS, suggest that presence of crossmodal interactions between face- and voice-responsive regions at a sensory level. Moreover, the absence of functional connections between the voice-responsive fusiform region and any other regions (apart from the STS) during familiar voice recognition suggests that face-voice interactions may take place without top-down crossmodal information being relayed to the fusiform cortex by higher-level multimodal regions. Instead, von Kriegstein, et al (2005) proposed that the auditory modality directly recruits the visual modality during familiar voice recognition through a crossmodal coupling process that does not require the involvement of a separate multimodal region or PIN. As a potential explanation of the engagement of a face-selective region during voice recognition, the authors proposed that a familiar voice induces an “implicit imagery” of the person’s face. These findings provide support to cognitive models of face and voice identity processing that propose direct interactions between FRU and VRU prior to the PIN stage (Belin et al., 2004; Campanella & Belin, 2007).

In a further study by the same research group, von Kriegstein, Kleinschmidt, & Giraud (2006) tested whether impaired face recognition influences the previously observed responses to familiar voices in the FFA. In this study, the authors used the same experimental paradigm to test a patient (‘SO’) with developmental prosopagnosia, which is a selective impairment in the recognition of familiar faces. The experiment was identical to their previous study (von Kriegstein et al., 2005), with the exception of the stimuli, for which 47 voice recordings from seven familiar and seven unfamiliar speakers were used (as opposed to 14 in the previous study). An analysis comparing FFA activation during familiar voice recognition between SO and the (healthy) participants from the previous study, which served as controls, showed no differences in activation. In addition, no differences were found in the functional connectivity between the FFA and the voice-responsive STS. Given SO’s impaired ability to process face identity, von Kriegstein et al. (2006) interpreted the finding of intact coupling between face and voice regions as evidence that the coupling mechanism does not directly involve person identity processing, and may take place at an earlier, sensory level. However, SO also reported impairments in face imagery, and therefore the face-voice coupling is unlikely to result from the face imagery induced by a familiar voice, as had been suggested previously (von Kriegstein et al., 2005). Therefore, the reason for the involvement of the FFA during voice recognition is not clear.

The two studies described so far show activation of the FFA during the recognition of personally-familiar voices. To test whether this effect is due to having access to a visual representation of a person’s facial appearance, or whether it is dependent on the knowledge of semantic information that is associated with personal familiarity, a further study by the same group investigated the coupling between face- and voice-selective regions during voice recognition before and after learning to associate unfamiliar voices with their respective faces or with names (von Kriegstein & Giraud, 2006). During the learning session, one group of participants learnt to associate the voices of five different speakers with faces, and a second group learned to associate the same voices with names. Before and after the learning session, participants were presented with 111 two-word sentences spoken by the five speakers, and performed a voice identity recognition task. After the learning of face-voice and name-voice associations, activation during the voice recognition task was compared between the group that learnt face-voice associations and the group that learnt name-voice associations. This analysis showed greater activation in the FFA after face-voice learning compared with name-voice learning. No significant activation was found for the opposite contrast (name-voice learning>face-voice learning). A functional connectivity analysis additionally showed increased interaction between the FFA and the voice-responsive STS after the learning of face-voice associations, compared with before the learning of these associations. A comparison of improvements in behavioural voice recognition accuracy post-learning phase between the face-voice association group and the name-voice association group suggested greater improvements for the face-voice association group. However, this group also had lower voice recognition rates than the name-voice group pre-learning phase, and therefore had more room for improvement. The results of von Kriegstein & Giraud (2006) demonstrate that activation of the FFA during voice recognition is dependent on knowledge of a person’s facial appearance, but does not require a person to be personally-familiar, i.e. associated with semantic person identity information. Therefore, the crossmodal coupling between the FFA and voice-responsive regions likely involves the exchange of sensory information (as opposed to semantic information, which is typically associated with a multimodal PIN in cognitive face and voice recognition models). Furthermore, these results suggest that crossmodal coupling during voice recognition is specific to faces, and does not extend to other visual identity-related information, such as visually-presented names. Von Kriegstein and Giraud (2006) speculated that voices engage multimodal representations that are formed due to the ability of faces and voices to convey redundant identity information, in contrast to voices and names, which are arbitrarily associated. Their results are in line with behavioural studies showing shorter response times to face-voice pairs compared with name-voice pairs in familiarity tasks, and which suggest direct connections between FRUs and VRUs prior to the PIN stage (O'Mahony & Newell, 2012; Schweinberger, Herholz, & Stief, 1997). While this study demonstrated that FFA activation during voice recognition is not contingent on being personally-familiar with a person, the purpose of the FFA involvement is still unclear.

To test whether impaired face recognition influences responses in the FFA during the recognition of unfamiliar voices, after the learned association with the face, von Kriegstein et al. (2008) compared a group of patients with prosopagnosia with a healthy control group. In this study, participants learned to associate three speakers with their corresponding face, and another three speakers with an occupation (represented by a visual symbol). Participants were then presented with 20 sentences and were asked to perform a speaker and a speech recognition task, which were similar to the ones used in their original study (von Kriegstein et al., 2005). The authors then contrasted activation in response to voices that had been associated with faces with activation in response to voices that had been associated with occupations. For both prosopagnosics and controls, they found increased activity in the FFA during the voice recognition task only, in line with their previous studies. Behaviourally, while speech recognition was better in both groups for voices that been associated with faces, speaker recognition was only improved for controls. Furthermore, a significant correlation between activation in the FFA and speaker recognition was only found for controls. This study suggests that, for people with intact face recognition abilities, learning a voice together with a face enhances voice recognition, and the extent of this enhancement in different people is related to the level of activation of their FFA during voice recognition. In contrast, learning a voice together with a face did not improve voice recognition in prosopagnosics, despite similar activation of the FFA during voice recognition in this group. Therefore, this study suggests that activation of the FFA during voice recognition may reflect the ability of a stored representation of a person's face to enhance voice recognition, but only for people with intact face recognition processing. It is not clear why the FFA is activated during voice recognition in prosopagnosics, given that it was not associated with enhanced voice recognition in this group.

The studies described so far present evidence of functional connections between the FFA and voice-responsive brain regions during voice recognition. To test whether these functional connections correspond to structural connections in the brain, Blank et al. (2011) used diffusion magnetic resonance imaging to investigate white matter connections between the FFA and voice-responsive regions. To localise regions that respond to voice identity processing compared to speech processing, participants were scanned using fMRI while performing a speaker and a speech recognition task (von Kriegstein et al., 2005). Probabilistic fiber tractography was computed between the right FFA, which was defined both using a face-localiser and based on responses to the voice identity task, and the right mid, anterior, and posterior STS. The authors reported structural connections between the FFAs, defined using both methods, and the voice-responsive regions of the STS, with stronger connections to the aSTS and mSTS compared with the pSTS. These connections were present in at least 50% of the participants, with the exception of the connection between the face-selective FFA and the pSTS which was only present in 5 participants, and contained at least 10 pathways between each pair of regions. The connections between the STS regions and the voice-responsive FFA were found to be stronger compared with the face-selective FFA, suggesting that the region of the fusiform gyrus that responds to voices may be functionally distinct from the face-selective FFA. Taken together, these findings further support the prediction of the CFVP model that face- and voice-responsive regions are able to directly exchange information with each other independently from multimodal brain regions. However, the functional significance of these connections remains unclear.

A recent study by a different group provided further evidence of structural connections between the FFA and voice-selective regions in the right hemisphere (Benetti et al., 2018). In this study, probabilistic tractography was computed between the FFA, defined using a face localiser, and the TVA, defined using a voice localiser. Voxels in the TVA that overlapped with the face-selective pSTS, which was defined using the localiser, were excluded from the TVA masks based on the assumption that the pSTS engages in multimodal processing. Benetti et al. (2018) argued that excluding the pSTS would allow focus on direct structural connections involving the TVA. Therefore, the TVA mask in this study included only the mSTS. Structural connections were found between the FFA and TVA, and these connections were present in 86% of participants and contained at least 10 pathways, similar to Blank et al. (2011). This finding shows that the connections between the FFA and the STS are not specific to regions of the STS that respond selectively during voice recognition, as demonstrated in Blank et al. (2011), and also apply to voice-selective regions of the STS, i.e. regions that respond selectively to voices over non-vocal sounds.

A study using magnetoencephalography (MEG) investigated the timing of the activation of the FFA during voice recognition (Schall, Kiebel, Maess, & Von Kriegstein, 2013). They used a similar experimental paradigm to von Kriegstein et al. (2008), in which participants learnt to associate three speakers with their corresponding faces, and three speakers with different occupations. Responses during a subsequent voice identity recognition task were recorded using MEG, and the FFA was localised using a face-localiser. Comparisons were then made between voices that had been associated with faces, and voices that had been associated with occupations. These comparisons revealed significant greater activation in the FFA at approximately 100ms after the onset of the voice, for voices that had been associated with faces. Moreover, these voices led to faster M200 responses, which reflect a previously observed peak in activity at 200ms after voice onset (e.g. Charest et al., 2009), and higher recognition accuracy. An analysis comparing average M200 latencies with recognition accuracy across participants showed a positive correlation, in that participants whose recognition accuracy benefitted more from seeing the faces, as opposed to the occupations, also showed a faster M200 response. This study provides evidence that the interaction between the FFA and voice-responsive regions takes place at an early processing stage. Furthermore, it provides additional evidence that learning to associate a voice with a face facilitates voice processing, and that this facilitation is reflected in brain activity.

To summarise, studies using fMRI, functional connectivity, and structural connectivity analyses have shown that face- and voice-selective regions directly exchange information during voice identity recognition. This finding applies to the recognition of the voices of personally-familiar people, and to the recognition of voices that have been briefly associated with faces prior to the experiment. For people with intact face recognition abilities, it seems that learning to associate a voice with a corresponding face improves subsequent recognition ability of the voice, and this ability is associated with the degree of activation of the FFA during voice recognition. These effects on voice recognition were found to be specific to face-voice associations, and did not apply to name-voice or occupation-voice associations. This suggests that the connections between the FFA and the voice-sensitive regions are unique to faces and voices, and may be due to their ability to relay redundant information. However, the observed activation in the FFA of prosopagnosics during voice recognition, who are impaired in familiar face processing and face imagery, confuses the interpretation of the purpose of these connections. Crucially, it is not known whether voice-selective regions are activated during face recognition in analogous way to the FFA activation during voice recognition.

Electrophysiological studies

Joassin, Maurage, Bruyer, Crommelinck, & Campanella (2004) conducted an event-related potential study exploring the electrophysiological basis of face and voice interactions during a task involving identity processing. They trained participants to associate faces, spoken names, and written names for 12 unfamiliar identities. In an electroencephalogram (EEG) experiment, the authors presented participants with the faces and voices in pairs (as an audio-visual stimulus) and separately. Each stimulus was preceded by a written name, and participants had to decide whether the name corresponded to the identity of the stimulus. In non-matching voice trials or face-voice trials, the written name preceding the stimulus would be the same as the name being spoken by the voice, but the voice itself would not belong to the identity that had been associated with that particular name. To examine electrical activity that is unique to the multimodal condition, the authors used the supra-additive criterion and subtracted the unimodal conditions from the multimodal condition (Calvert et al., 2000). They observed three main waves of electrical activity. The first wave occured at around 100ms, was localised in the bilateral superior colliculus and fusiform gyrus, and was interpreted as potential evidence of an influence of auditory information on visual processing. The second wave occurred at around 175ms, was localised mainly in the bilateral superior colliculus, STG, and inferior frontal gyrus, and was interpreted as reflecting a possible influence of visual information on auditory processing. Finally, the third wave occurred at around 279ms and was localised mainly in a network formed of the superior colliculus, superior frontal gyrus, inferior frontal gyrus, and fusiform gyrus. Joassin et al. (2004) speculated that this third wave may reflect an interaction between unimodal regions, multimodal regions, and regions that process semantic information. While this study provides some evidence of early integration of face and voice information, the poor spatial resolution of EEG means that any interpretations of the brain regions involved and of the mechanisms of integration are largely speculative.

In a more recent EEG study, Föcker, Hölig, Best, & Röder (2011) investigated electrical activity in response to face-voice combinations that were either congruent or incongruent in terms of identity. Specifically, the authors presented participants with voices that were preceded by a face that either belonged to the same identity, or belonged to a different identity, and asked them to judge the age of the voice. At 100-140ms after the onset of the voice, they observed more negative event related potentials for voices that were preceded by a different-identity face, compared with voices that were preceded by a same-identity face. Although this study did not localise the sources of electrical activity, the authors speculated that the early modulation of activity by identity-incongruent face-voice pairings presents evidence for the integration of face and voice information at early processing stages. However, as discussed previously, brain activation to incongruent stimuli may be a response to incongruence regardless of the modalities being presented (Laurienti et al., 2005).

Conclusions

Evidence for the CFVP model comes mainly from fMRI studies showing activation of the FFA during voice recognition, and functional and structural connections between the FFA and voice-responsive regions in the STS. However, in the majority of studies FFA activation is contingent on an explicit voice recognition task, and there is a lack of evidence for the involvement of voice-responsive regions during face recognition. Therefore, it is unclear whether the proposed exchange of information between the FFA and voice-responsive regions contributes to the recognition of person identity, or whether, instead, the primary role of this exchange is to facilitate voice identity processing by providing access to face representations. Electrophysiological studies have shown evidence of early interactions between face and voice processing, but due poor spatial resolution, cannot provide precise information regarding the brain regions that are involved in these interactions.

1.2.4 Summary

To summarise, current evidence suggests that face and voice information is integrated in the brain during identity recognition through two separate mechanisms. One mechanism involves early exchange of face and voice information between face- and voice-selective regions (CFVP model), whereas the other mechanism involves the integration of face and voice information in multimodal brain regions (MP model). The ‘coupling’ mechanism, proposed by the CFVP model and the ‘multimodal’ mechanism, proposed by the MP model, are not mutually exclusive (Campanella & Belin, 2007; Gainotti, 2014). Specifically, it is possible that FRUs and VRUs communicate with each other prior to the PIN stage, in line with the CFVP model, and subsequently pass on information to be integrated and stored in the PINs, in line with the MP model. So far, the majority of studies have focused either on the ‘multimodal’ or on the ‘coupling’ mechanism. Thus, the relative contribution of each mechanism to the integration of information from faces and voices remains unclear.

Currently, there is limited knowledge on the type of face and voice information represented and exchanged between face- and voice-selective regions, and the type of information represented in multimodal brain regions during identity recognition. The next section will review fMRI studies that examine the neural correlates of different types of information in faces and voices, including visual and auditory information, gender, and social information.

1.3 The informational content of face and voice representations

Faces and voices serve as sources of multiple types of information (Yovel & Belin, 2013). As visual and auditory stimuli, they convey physical information related to their corresponding modalities: faces convey visual properties, such as shape and texture (Troje & Bulthoff, 1996), and voices convey auditory properties, such as pitch and loudness (Titze, 1989). Furthermore, faces and voices serve as sources of person identification, and convey information about physical characteristics of a person that distinguish that person from others, such as gender and age (Yovel & Belin, 2013). Finally, due to their importance in communication and social interactions, faces and voices convey socially-relevant information about a person, such as how trustworthy, dominant and attractive they are (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013). The 'special' nature of faces and voices is reflected in the brain (Belin, 2017), which contains specialised neural substrates for the processing of both faces and voices, as discussed previously. However, despite significant advances in defining face-selective and voice-selective brain regions, little is known regarding the type of face and voice information that is processed in the different regions. Moreover, while studies have shown that face-responsive regions can distinguish between different face identities (e.g. Anzellotti et al., 2014; Guntupalli et al., 2017), and that voice-responsive regions can distinguish between different voice identities (e.g. Bonte et al., 2014; Formisano et al., 2008), there is limited knowledge of the type of face or voice information used by different regions to distinguish between identities. This section will describe studies that have attempted to identify the type of information that is processed in face-selective regions, in voice-selective regions, and in regions that overlap with the known locations of face-selective and voice-selective regions. These studies are grouped into three broad categories, based on the type of face and voice information that was their primary focus: physical properties, gender, and social information.

1.3.1 Physical properties

Faces and voices convey information regarding their visual and auditory properties, respectively. This subsection describes evidence linking the FFA, OFA, and STS with the processing of visual properties in faces, and evidence linking the TVAs, or regions known to overlap with the TVAs, with the processing of auditory properties in voices.

Faces

It has been proposed that individual faces are represented as locations in a multidimensional 'face space' that is centred on a 'prototype' face, which represents the average of all faces encountered during a lifetime (Valentine & Bruce, 1986). According to this norm-based coding model (Rhodes, Brennan, & Carey, 1987), the more distant an individual face is from the prototype face, the more distinctive it is perceived to be (Valentine, 1991). Loffler et al. (2005) examined activity in the FFA in response to artificially-generated faces that had been manipulated to have different geometric distances from their mean face within a face space. Specifically, these faces varied based on head shape, hair line, and the size and location of facial features. The authors showed that distinctive faces, i.e. faces that were further from the mean in the face space, elicited higher activation in the FFA compared with less distinctive faces. However, they also showed that FFA activation in response to presentations of faces with different degrees of distinctiveness was reduced for faces with the same identity, compared to faces with different identities, suggesting image-invariant adaptation to identity in the FFA. This finding suggests that, while the FFA is sensitive to the degree of physical distinctiveness in faces, it is also sensitive to identity information and responds more similarly when different faces depict the same person.

The right FFA has been shown to be sensitive to physical differences between stimuli regardless of whether these differences result from changes in identity or changes in viewpoint (Xu, Yue, Lescroart, Biederman, & Kim, 2009). In an fMRI adaptation experiment, Xu et al. (2009) presented participants with pairs of artificially-generated faces, presented in sequence, which varied in terms of identity and/or viewpoint or were identical. The face pairs were manipulated so that changes in identity and changes in viewpoint were equivalent in terms of physical magnitude, as determined by the Gabor-Jet model (Biederman & Kalocsai, 1997). This model was designed to simulate response properties of cells in area V1, and correlates with psychophysical measures of facial similarity (Yue, Biederman, Mangini, Malsburg, &

Amir, 2012). Xu et al. (2009) showed that activation in the FFA was higher for face pairs that differed in viewpoint or identity (showing release from adaptation), compared to face pairs that were identical. Moreover, there was no difference in the degree of release from adaptation between changes in identity and changes in viewpoint. These findings suggest that the FFA represents the physical similarity between faces. However, in contrast to Loffler et al. (2005), this study did not show evidence that FFA activity is modulated by face identity.

In a subsequent study by the same group, Xu & Biederman (2010) used a similar experimental paradigm to investigate effects of changes in emotional expression and viewpoint in the bilateral FFA, OFA and STS. Similar to Xu et al. (2009), face pairs were manipulated so that changes in emotional expression and viewpoint were equivalent in terms of physical magnitude. The stimuli were pairs of artificially-generated faces that varied in terms of emotional expression and/or viewpoint, or that where identical. In the FFA, the authors showed release from adaptation for changes in expression, but not for changes in viewpoint, suggesting that the FFA contains viewpoint-invariant representations of identity, in contrast to their previous study. This may be due to the use of a bilateral FFA ROI, as opposed to the right FFA. No adaptation effects were found in the OFA or the STS for expression or viewpoint, suggesting that these regions are not sensitive to physical differences between faces.

In a MVPA study, Weibert, Flack, Young, & Andrews (2018) found that individual face representations in the FFA, OFA, and STS could be predicted by low-level image properties. Across two experiments, the authors presented grey-scale photographs of faces that varied in viewpoint and in either identity or emotional expression. Image statistics were obtained for each face image using the GIST image descriptor (Oliva & Torralba, 2001). The GIST descriptor was developed for images of scenes, and describes the spatial structure of images (Oliva & Torralba, 2001). In this study, the GIST image descriptors were compared between different experimental conditions (there were multiple images in each condition), which were defined based on combinations of face identities, viewpoints, and emotional expressions, to determine their similarity in terms of low-level image properties. For example, correlations were computed between GIST descriptors for pairs of conditions featuring faces that had the same identity but were presented from different viewpoints, with each viewpoint condition comprising several different images of the face of that identity presented from the same viewpoint. Weibert et al. (2018) showed that the similarity between the brain representations of different face conditions was predicted by the similarity between these conditions on their low-level image properties in the FFA, OFA, and STS. This study suggests that face-selective regions primarily use low-level image properties to distinguish between different face characteristics, such as viewpoint, identity, and expression. The finding of sensitivity to low-level visual properties in the OFA and STS is in contrast to Xu & Biederman (2010), who found no effect of physical differences between faces in these regions. Aside from the methodological differences of the two studies, this discrepancy may be due to Weibert et al. (2018) using face photographs, which are more naturalistically variable, and therefore distinguishable from each other, compared with artificially-generated faces, such as those used by Xu & Biederman (2010).

Another MVPA study compared different computational models of the visual system with representations of individual faces in the FFA (Carlin & Kriegeskorte, 2017). A sigmoidal-ramp tuning model and a Gaussian exemplar model were generated based on coordinates of faces in a PCA face space. This face space was computed based on four artificially-generated face identities with three different levels of distinctiveness. The sigmoidal-ramp tuning model was thought to simulate the processing of extreme visual features in area V4, whereas the Gaussian exemplar model was designed to simulate norm-based coding of individual faces (Valentine, 1991). Furthermore, a Gabor-filter model was computed based on grey-scale pixel intensities in the images, thought to simulate processing in area V1. Carlin & Kriegeskorte (2017) showed that all three models explained variance in the FFA activation in response to individual faces, but the sigmoidal-ramp tuning model and the Gabor-filter model performed better than the Gaussian exemplar model. This study therefore demonstrates that the FFA uses physical image properties to distinguish between individual faces.

Voices

A norm-based coding mechanism, equivalent to the one proposed for faces, has also been proposed for voices, whereby individual voices are represented within a 'voice space' based on their distance from a prototype voice (Latinus & Belin, 2011; Lavner, Rosenhouse, & Gath, 2001). Latinus, McAleer, Bestelmeyer, & Belin (2013) investigated whether the distance of individual voices to a prototype voice modulates activity in the TVAs. The authors presented participants with 64 male and female voices speaking the syllable "had". Separately for male and female voices, voices were represented in a three-dimensional voice space along the dimensions of fundamental frequency (f0), formant dispersion, and harmonics-to-noise ratio. Prototype male and female voices were created by morphing together all voices for each gender. The distance of each individual voice to the prototype was correlated with perceived distinctiveness. Latinus et al. (2013) found that activity in the TVAs was higher for distinctive voices, i.e. voices that with larger distances to the mean, compared with less distinctive voices. This result was replicated in a separate experiment using different participants and stimuli. These findings suggest that voice distinctiveness is represented in the TVAs in a similar way to face distinctiveness in the FFA.

Von Kriegstein, Smith, Patterson, Ives, & Griffiths (2007) investigated the neural correlates of human vocal tract length (VTL), an acoustic parameter that signals body size and is related to the filtering of sound through the vocal tract, i.e. the pharynx, mouth, and nose (Fitch, 2000). The authors manipulated a single vocalisation (the vowel 'a') from one speaker to simulate three speakers with different VTLs, and did the same for two control sounds featuring a bullfrog croak and a French horn note. Human voices and the control sounds were presented in separate blocks, and each block presented either the exact same sound or sounds featuring different VTLs. To reveal regions that are sensitive to changes in VTL, activation in response to blocks with varying VTL was contrasted with activation in response to blocks with consistent VTL. To additionally show regions that process VTL information in the human voice, results were compared between human voices and the control sounds. Von Kriegstein et al. (2007) showed that a region in the left posterior STG was significantly more sensitive to changes in VTL in the human voice compared with the control sounds. This finding suggests that the left posterior STG processes VTL information in human voices. While this study did not demonstrate voice-selectivity in this region using a conventional voice localiser, the TVAs are known to contain the left STG (Pernet et al., 2015), and therefore may also be sensitive to VTL information.

In a later study by the same group, von Kriegstein, Smith, Patterson, Kiebel, & Griffiths (2010) used a similar paradigm as von Kriegstein et al. (2007) to probe whether the previously identified region in the posterior STG that showed sensitivity to VTL is also involved in speech processing. The authors presented participants with sequences of spoken syllables that showed either constant or variable VTL, and asked them to perform either a speech recognition task or a control task (involving judgements of loudness). In the speech recognition task participants judged whether each syllable was the same as the previous one. To identify regions that process both VTL and speech, the authors compared activation in response to syllable sequences with variable VTL to activation in response to syllable sequences with fixed VTL, and additionally compared responses during the speech recognition task and the control task. The results revealed a region in the posterior STS/STG that was sensitive to VTL, in line with their previous study (von Kriegstein et al.,2007), and that also responded more during the speech recognition task that to the control task. Therefore, the posterior STS/STG appears to process both VTL and speech information. However, it is an open question whether this finding extends to the voice-selective left TVA.

Finally a MVPA study investigating representations of voice identity (described in detail in section 1.1.2) showed that regions of the bilateral STS/STG could discriminate between different speakers and between different vowel sounds (Formisano et al., 2008). Although these regions where not defined based on voice selectivity, the voice-selective TVAs are known to include the STS/STG. To determine what type of information was associated with speaker discrimination and vowel discrimination in these regions, Formisano et al. (2008) computed distances between the activity patterns in response to different speakers and different vowel sounds, and correlated the obtained distances with measured distances between the stimuli on low-level acoustic properties. Specifically, the authors extracted the voice fundamental frequency (f0), i.e. voice pitch, and the first two formant frequencies (f1, f2) from the voice stimuli. They showed that speaker identity discrimination was associated mostly with f0, whereas vowel sound discrimination was associated mostly with f1 and f2. These findings suggest that the bilateral STS/STG uses low-level acoustic information to distinguish between different voices. However, it is not known whether these findings also apply to the voice-selective portions of the STS/STG.

Conclusion

The FFA has been associated with the processing of visual information faces, including face distinctiveness (Loffler et al., 2005) and low-level visual properties described by the Gabor-Jet model (Xu & Biederman, 2010; Xu et al., 2009), the GIST image descriptor model (Weibert et al., 2018), and computational models relating to the location of faces in face space and image pixel values (Carlin & Kriegeskorte, 2017). Out of these studies, only two investigated additional face-selective regions, and they showed inconsistent findings (Weibert et al., 2018; Xu & Biederman, 2010). While Weibert et al. (2018) found that the face-selective OFA and STS used low-level information to distinguish between faces, Xu & Biederman (2010) did not show any modulation of activity in these regions related to physical differences between faces. This difference in findings may be due to Weibert et al. (2018) using naturalistically variable, as opposed to artificially-generated, face stimuli. Moreover, Weibert et al. (2018) used a MVPA approach, which may be more suited to detect sensitivity to physical stimulus properties in face-selective regions compared with adaptation designs, such as the one used by Xu & Biederman (2010). For voices, one study associated the voice-selective TVAs with the processing of voice distinctiveness (Latinus et al., 2013). Other studies have shown evidence of the processing of auditory information in voices in regions of the STS/STG, including VTL (von Kriegstein et al., 2007, 2010), vocal pitch, and voice formants (Formisano et al., 2008). While the STS/STG is known to overlap with the TVAs (Pernet et al., 2015), these studies did not use a voice localiser to explicitly define the TVAs, and therefore it cannot be assumed that the findings also apply to voice-selective regions. Therefore, there is limited information on the computations of the TVAs themselves. Finally, it should be noted that the studies presented in this subsection, which the exception of Weibert et al. (2018), used artificially-generated face images or short vocalisations as stimuli. Given that these stimuli weakly resemble the faces and voices encountered in everyday life, it is an open question whether the observed representations of physical properties in face-selective and voice-selective regions would also apply to more naturalistic tokens of faces and voices, such as colour face photographs that are unconstrained in terms of low-level image properties, and longer recordings of speech.

1.3.2 Gender

Information about a person's gender is conveyed by both their face and their voice. This subsection describes evidence suggesting that multiple face-selective regions, including the FFA, OFA, and pSTS, process information relating to face gender, and that regions of the STS/STG that are known to overlap with the TVAs process information relating to voice gender.

Faces

To investigate brain responses to different levels of gender information in faces, Freeman, Rule, Adams, & Ambady (2010) presented participants with artificially-generated grey-scale faces that had been morphed between the two genders so that they appeared highly feminine at one end of a continuum of faces, highly masculine at the other end of the continuum, and androgynous in the middle. The authors found that responses in the bilateral FFA increased as faces approached the ends of the continuum, i.e. as they became more feminine or masculine as opposed to androgynous. A similar finding of stronger responses in the FFA to very masculine and very feminine faces, compared to faces with average levels of masculinity-femininity, is reported by Mattavelli, Andrews, Asghar, Towler, & Young (2012), who also presented images of faces that had been morphed along a masculinity-femininity continuum. In addition, Mattavelli et al. (2012) showed a similar response pattern in the bilateral OFA, amygdala, and right pSTS. In the same study, Mattavelli et al. (2012) also found stronger responses to very trustworthy and very untrustworthy faces, compared with faces showing average levels of trustworthiness, in the same regions (this study is discussed in more detail in section 1.3.3). Therefore, the authors proposed that face-selective regions may have responded to the facial distinctiveness that resulted from the gender manipulation, rather than to gender itself. Given that faces with extreme levels of masculinity and femininity are likely to be perceived as more physically distinctive than faces with average levels of masculinity-femininity in everyday life, it is possible that influences of facial masculinity-femininity on brain activity cannot be dissociated from influences of face distinctiveness.

Two MVPA studies investigated the ability of face-selective regions to distinguish between individual faces based on their gender (Contreras, Banaji, & Mitchell, 2013; Kaul, Rees, & Ishai, 2011). In a study presenting a large number of grey-scale photographs of faces (320), Kaul et al. (2011) showed above-chance classification accuracy of male and female faces in multiple face-selective regions: the FFA, OFA, pSTS, inferior frontal gyrus, orbitofrontal cortex (OFC), and insula. In a study using a MVPA correlation approach, Contreras et al. (2013) presented grey-scale photographs of male and female faces of white or black race, and found that multivoxel patterns were more similar among faces with the same gender than among faces with different genders (regardless of race) in the bilateral FFA, but not in the OFA and STS (other face-selective regions were not defined). Taken together, the findings of these two studies agree that the FFA uses gender information to distinguish between different faces, but are inconsistent regarding the OFA and pSTS. This may be due to the different methodological approach used in the two studies, as well as the large number of stimuli presented by Kaul et al. (2011). In these studies the discrimination between male and female faces is unlikely to be confounded by distinctiveness. However, male and female faces differ systematically based on features of the face such as the eyebrows, nose, and chin (Bruce et al., 1993), and therefore it is likely that gender discrimination is confounded by systematic physical differences between male and female faces.

Voices

The majority of studies on the neural correlates of voice gender have focused on comparing brain responses to male and female voices. While these studies did not explicitly define voice-selective regions, some implicated regions that are known to overlap with the TVAs. For example, Sokhi, Hunter, Wilkinson, & Woodruff (2005) and Lattner, Meyer, & Friederici (2005) found stronger responses to female voices, compared with male voices, in regions of the right STS/STG. One study addressed the possibility that modulations in brain activity associated with gender are due to differences in vocal pitch between male and female voices (Weston, Hunter, Sokhi, Wilkinson, & Woodruff, 2015). Weston et al. (2015) presented participants with pitch-altered male voices (sentences) that were raised to a similar pitch level as typical female voices, and pitch-altered female voices that were lowered to a similar pitch level to typical male voices, in addition to unaltered male and female voices. They then compared the combined brain activation to original and pitch-altered female voices with the combined activation to original and pitch-altered male voices, and revealed a stronger response to female voices compared to male voices in the upper bank of the left STS. Importantly, due to the inclusion of pitch-altered voices, the male and female voice conditions had a similar average pitch level. Therefore, the left STS response is unlikely to be due to differences in pitch. Finally, to test whether this response was due to the psychological perception of gender, the authors also presented gender-ambiguous voices, and compared activation in response to voices that were categorised as male and voices that were categorised as female by participants during a task. This analysis did not reveal any significant differences in the left STS or any other region between perceived-female and perceived-male voices, suggesting that the response in the left STS cannot be explained by the psychological perception of gender. Instead, the authors speculated that the higher response to female voices in the left STS may due to acoustic measures related voice timbre, which is a measure of voice quality that is considered to be independent from voice pitch (Cleveland, 1977).

Finally, Charest, Pernet, Latinus, Crabbe, & Belin (2013) investigated representations of voice gender using an adaptation paradigm with voice stimuli (syllables) that had been morphed along male-to-female continua, in a similar way to the faces in Freeman et al. (2010) and Mattavelli et al. (2012). They showed that responses in the right aSTS increased linearly as the physical difference in gender between consecutive stimuli increased (i.e. as the stimuli were further apart along the gender continuum). This finding suggests that the right aSTS is sensitive to gender information in voices. Although this region overlapped with the right TVA, which was identified separately using a voice localizer, neither of the TVAs showed a significant adaptation effect, despite a trend in that direction. Therefore, it is possible that a sub-region of the TVAs, and not the TVAs themselves, may contain representations of voice gender. However, in a similar way to faces, it is likely that very masculine and very feminine voices are perceived as physically distinctive, and therefore influences of vocal masculinity-femininity on brain activity cannot be dissociated from influences of voice distinctiveness.

Conclusion

Studies have shown evidence that face-selective regions respond more to faces with extreme levels of masculinity or femininity (Freeman et al., 2010; Mattavelli et al., 2012). Moroever, a region in the right aSTS, which overlaps with the right TVA, responds more to voices with extreme levels of masculinity or femininity (Charest et al., 2013). However, in this review I argued that these findings may reflect a sensitivity to the facial and vocal distinctiveness in terms of physical characteristics, which results from extreme levels of masculinity and femininity in faces and voices. For faces, two MVPA studies showed that responses to male and female faces can be distinguished in mutliple face-selective regions, including the FFA (Contreras et al., 2013; Kaul et al., 2011), OFA, pSTS, inferior frontal gyrus, OFC, and insula (Kaul et al., 2011). However, given that differences in gender are confounded with systematic differences in visual appearance (Bruce et al., 1993), these studies cannot rule out the possibility that male and female faces are being distinguished in these regions based on their physical characteristics. For voices, studies have shown that regions of the STS/STG, which is known to overlap with the TVAs (Pernet et al., 2015), respond more to female voices compared with male voices (Lattner et al., 2005; Sokhi et al., 2005; Weston et al., 2015). One of these studies showed that the STS/STG response could not be explained by differences in vocal pitch between male and female voices, or by the psychological perception of gender (Lattner et al., 2005). Lattner et al. (2005) speculated that responses to gender may be explained by differences in voice quality between male and female voices. In sum, this review of studies investigating the neural correlates of face and voice gender suggests that face-selective and voice-selective regions may be sensitive to physical face/voice characteristics that differ systematically between male and female faces/voices, and which cannot be dissociated from the perception of gender in these regions.

1.3.3 Social information

Both faces and voices serve as sources of socially-relevant information about people (Yovel & Belin, 2013). Studies have shown that both faces and voices elicit impressions of social traits such as trustworthiness, attractiveness, and dominance (McAleer, Todorov, & Belin, 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zuckerman & Driver, 1989). The majority of studies investigated the social evaluation of the faces and voices of unfamiliar people, and showed that impressions of faces and voices are largely consistent across participants even after very brief exposures (McAleer et al., 2014; Oosterhof & Todorov, 2008) and are formed automatically (Ritchie, Palermo, & Rhodes, 2017). Furthermore, these judgments seem to be associated with variations in facial and vocal features, such as face shape and vocal pitch (McAleer et al., 2014; Robinson, Blais, Duncan, Forget, & Fiset, 2014; Todorov & Oosterhof, 2011). This subsection describes evidence suggesting that activity in the amygdala, FFA, OFA, and pSTS is modulated by different levels of perceived traits in faces, and that activity in regions that overlap with the TVAs is modulated by different levels of attractiveness in voices.

<u>Faces</u>

Todorov & Engell (2008) compared brain activity in response to photographs of faces with behavioural ratings of the faces on 14 social traits. The authors localised the bilateral amygdala based on anatomical masks, and face-responsive regions were defined by contrasting activation in response to the faces with baseline. Although the amygdala was not defined based on face-selectivity in this study, face-selectivity has been identified previously in this region (Fox et al., 2009). The face-responsive fusiform gyrus region was consistent with the known location of the FFA, but none of the other regions overlapped with known face-selective regions. Todorov & Engell (2008) showed that activation in the bilateral amygdala and fusiform gyrus was negatively correlated with ratings of positive traits (e.g. trustworthy, attractive), and positively correlated with ratings of negative traits (e.g. threatening, mean). Moreover, after extracting a positive-negative valence component from a principal component analysis of the trait ratings, the authors found significant correlations between this component and activity in both amygdala and the fusiform gyrus. In addition, when controlling for the variance in the ratings of each trait that was explained by the valence component, correlations between the traits and brain activity ceased to be significant. These findings suggest that the amygdala and fusiform gyrus are sensitive to a positive-negative valence dimension in faces. However, Todorov & Engell (2008) also showed that after partialling out the variance explained by the bilateral amygdala, the correlation between the fusiform gyrus and the valence dimension were no longer significant. Based on this finding, the authors proposed that activity in the fusiform gyrus is modulated by amygdala activity, and that facial valence is primarily processed in the amygdala.

Further evidence supporting a negative relationship between amygdala activation and facial valence, in which the amygdala is more activated by untrustworthy compared with trustworthy faces, has been revealed by studies that used participants' ratings of faces on social judgements as parametric modulators in the analysis of response magnitude to faces in the brain (Engell, Haxby, & Todorov, 2007; Todorov, Baron, & Oosterhof, 2008; Winston et al., 2002). In this approach, for each participant, the modelling of the brain response to each stimulus includes the modulation of their rating of the stimulus, and then random effects analysis is conducted across all participants. Moreover, two meta-analyses have shown consistent involvement of the amygdala in the processing of facial trustworthiness and attractiveness (Bzdok et al., 2011; Santos, Almeida, Oliveiros, & Castelo-Branco, 2016). Bzdok et al., 2011 conducted an activation likelihood estimation meta-analysis of 16 studies involving neural correlates of facial trustworthiness or attractiveness, whereas Santos et al., 2016 considered 20 studies involving trustworthiness only, using a meta-analysis of effect sizes in addition to activation likelihood estimation analysis. Finally, activation of the amygdala in response to untrustworthiness and unattractiveness was highlighted in a multi-level kernel density meta-analysis by Mende-Siedlecki, Said, & Todorov (2013), who examined 11 studies that contrasted faces with negative valence with faces with positive valence in terms of trustworthiness or attractiveness.

Findings of a linear relationship between valence and amygdala activation have been brought into question by evidence showing quadratic effects of facial trustworthiness and attractiveness in the amygdala, with stronger responses to either untrustworthy or trustworthy faces compared with medium-trustworthy faces (Freeman, Stolier, Ingbretsen, & Hehman, 2014; Mattavelli et al., 2012; Said, Baron, & Todorov, 2009; Said, Dotsch, & Todorov, 2011; Todorov et al., 2008; Todorov, Said, Oosterhof, & Engell, 2011), and stronger responses to attractive and unattractive faces compared with average-attractive faces (Winston, O'Doherty, Kilner, Perrett, & Dolan, 2007). These findings raise the possibility that the amygdala may be responding to the facial distinctiveness that results from the extreme presence or absence of social traits.

Evidence that the amygdala responds to face distinctiveness has been provided by two studies that demonstrated that the amygdala shows similar quadratic responses to valence and to non-social face dimensions (Mattavelli et al., 2012; Said et al., 2011). Mattavelli et al. (2012) investigated the neural correlates of facial trustworthiness and face gender by using face stimuli that were manipulated to vary either in terms of trustworthiness or in terms of gender. Specifically, the authors first collected trustworthiness ratings on a large number of male and female faces (500) obtained from the Internet, which were unconstrained in terms of age, pose, and expression. Based on these ratings, for each gender the authors created an average trustworthy and an average untrustworthy face. The final stimuli were created by morphing these prototype faces to create four face continua for four different levels of trustworthiness, consisting of faces that varied in terms of gender, and four face continua for four different levels of masculinity-femininity, consisting of faces that varied in terms of trustworthiness. Importantly, an equal amount of morphing was applied to individual faces in the valence continua and in the gender continua, so that faces in the same position in the continua were equally distinct from the prototype faces. In the experiment, faces from each continuum were presented in separate blocks. Mattavelli et al. (2012) identified face-selective ROIs in the amygdala, FFA, OFA, and right pSTS, and tested for linear or quadratic responses to trustworthiness and gender within these regions. The results of Mattavelli et al. (2012) showed that the amygdala, FFA, OFA, and right pSTS showed stronger quadratic responses, compared with linear responses, for both trustworthiness and gender. Specifically, activation in these regions was higher in response to highly trustworthy, untrustworthy, masculine, and feminine faces, compared with faces showing medium levels of trustworthiness and masculinity-femininity. No differences were found when comparing responses to trustworthiness and responses to gender. Given that faces in both the trustworthiness and gender continua were equally distinct from the average face these findings strongly suggest that the amygdala is involved in the processing of facial distinctiveness, as opposed to valence or gender per se. It should be noted that the majority of previous studies did not define the amygdala based on face-selectivity, and also did not explicitly investigate responses to valence in other face-selective regions. The finding of sensitivity to face distinctiveness in the FFA is in line with the previously described study by Loffler et al. (2005). Moroever, the findings of Mattavelli et al. (2012) suggest that the OFA and pSTS may also be sensitive to face distinctiveness in terms of physical features.

Similar findings to Mattavelli et al. (2012) regarding quadratic relationships between activity in the amygdala and FFA for both social and non-social dimensions were reported by Said et al. (2011). This study used a computational model of social traits in faces (Oosterhof & Todorov, 2008) to manipulate artificially-generated faces on valence and on an undefined dimension orthogonal to valence, which was found to be less related to trait judgements of the faces compared with valence. Similar to Mattavelli et al. (2012), the faces in the two dimensions were equated based on their distance to an average face. The results showed a quadratic response to both valence and the control (non-social) dimension in the FFA and amygdala. Taken together, the findings of Mattavelli et al. (2012) and Said et al. (2011) suggest that both the amygdala and the FFA may be sensitive to facial distinctiveness regardless of the dimension that is being manipulated.

Voices

One study investigated the effect of different levels of perceived attractiveness in voices on response magnitude in voice-responsive regions (Bestelmeyer et al., 2012). Bestelmeyer et al. (2012) presented participants with voices speaking the sound “ah”, which had been previously rated on attractiveness by a separate group of participants, while they performed an unrelated task in the scanner. The authors found a negative correlation between attractiveness ratings and responses in the right inferior frontal gyrus and bilateral STS/STG, with stronger responses to voices as perceived vocal attractiveness decreased. Bestelmeyer et al. (2012) demonstrated that the STS/STG overlapped with the TVAs, which were defined separately using a voice localiser; however, responses to voice attractiveness were not directly tested in the TVAs. Although the inferior frontal gyrus is not commonly defined a voice-selective region, voice-selective responses have been observed previously in the inferior prefrontal cortex (Pernet et al., 2015). To determine whether the sensitivity of the STS/STG and inferior frontal gyrus to vocal attractiveness could be explained by variations in acoustic features, the authors repeated their analysis while controlling for variance attributed to the distance of each voice to the average voice for each gender, and to the harmonics-to-noise ratio of each voice, which is a measure of voice quality. This further analysis showed that acoustic measures explained much of the variance for the STS/STG, but not for the inferior frontal gyrus. This suggests that the inferior frontal gyrus processes higher-level perceptual information in voices that cannot be fully explained by the variation in acoustic features. In contrast, the STS/STG, and potentially the TVAs, may primarily process lower-level acoustic information that is associated with differences in vocal attractiveness.

Conclusion

Multiple studies have implicated the amygdala in the processing of valence information in faces (Bzdok et al., 2011; Mende-Siedlecki et al., 2013; Santos et al., 2016). However, findings showing that the face-selective amygdala shows stronger responses to physically distinctive faces compared with average faces, regardless of whether this distinctiveness results from the extreme presence or absence of a social or a physical characteristic (Mattavelli et al., 2012; Said et al., 2011), suggest that the amygdala may primarily respond to physical distinctiveness in faces. Similar findings were also demonstrated in the FFA, OFA, and the pSTS (Mattavelli et al., 2012; Said et al., 2011), suggesting that these regions show a similar response profile to the amygdala. For voices, one study showed that regions of the STS/STG that overlapped with the TVAs were sensitive to the acoustic sound properties that were associated with vocal attractiveness (Bestelmeyer et al., 2012). However, the processing of social information in the TVAs has not directly been investigated. Moreover, to the best of my knowledge, no studies have investigated the ability of face-selective and voice-selective regions to discriminate between individual faces or voices based on perceived social information.

1.3.4 Multimodal brain regions

Little is known regarding the type of face or voice information that is processed in multimodal brain regions that respond to both faces and voices. Faces and voices often convey similar information about people, including physical characteristics such as masculinity-femininity (Smith, Dunn, Baguley, & Stacey, 2016a), and social characteristics such as trustworthiness, dominance, and attractiveness (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zuckerman & Driver, 1989). It is possible that this information is processed in separate regions for separate modalities, even if using similar coding principles (Yovel & Belin, 2013), and the evidence reviewed above suggests that this may be the case at least to some extent in face- and voice-responsive regions. However, it is possible that some of these characteristics are processed in the same brain regions for both modalities, such that a multimodal region would represent this information independently of input modality. Two previous studies suggest that this may be the case, in terms of representations of emotion (Peelen, Atkinson, & Vuilleumier, 2010) and positive-negative valence (Chikazoe, Lee, Kriegeskorte, & Anderson, 2014).

Peelen et al. (2010) showed that activity patterns in the medial prefrontal cortex and the left STS in response to faces, voices, and bodies expressing the same emotion were more similar to each other than to activity patterns in response to the same classes of stimuli expressing different emotions. This finding is in agreement with a behavioural study showing that emotions are conceptualised in similar way regardless of whether they are perceived through the face or the voice (Kuhn, Wydell, Lavan, McGettigan, & Garrido, 2017). Using similar methods to Peelen et al. (2010), Chikazoe et al., (2014) found that brain activity patterns in the medial and lateral OFC in response to pleasant images of scenes and tastes were more similar to each other than to activity patterns to unpleasant images and tastes (and vice versa). Notably, the authors also found modality-specific valence representations in the ventral temporal cortex and in the anterior insula. Taken together, these studies indicate that the brain contains both crossmodal and modality-specific representations, and that crossmodal representations are likely to be found in multimodal brain regions such as the pSTS and OFC.

1.3.5 Summary

The computations of face-selective and voice-selective regions have been associated with physical face/voice properties, gender, and social information. However, the current evidence suggests that the face-selective FFA, OFA, pSTS, and amygdala may primarily process visual face properties, and that the TVAs may be mainly involved in the processing of acoustic voice properties. Specifically, the above review showed that many of the findings associating face-selective and voice-selective regions with the processing of gender and social information could potentially be explained by physical face and voice properties, such as distinctiveness in terms of physical features. One limitation of the described studies is that the majority focused on investigating one type of face or voice information (with the exceptions of Mattavelli et al., 2012 and Said et al., 2011). Therefore, there is a lack of studies using the same experimental paradigm and stimuli to investigate the processing of different types of face/voice information in different brain regions. As demonstrated by Mattavelli et al. (2012) and Said et al. (2011), such studies can directly compare brain responses to different types of information.

An important limitation concerning studies investigating neural correlates of information extracted from the voice is that many studies did not independently define the TVAs using a voice localiser (with the exceptions of Charest et al., 2013 and Latinus et al., 2013). Therefore, even though many findings involved regions that are known to overlap with the TVAs, it is not certain that these findings would apply to voice-selective regions. For studies investigating the processing of physical properties in face-selective and voice-selective regions, one concern is the use of face and voice stimuli with low variability. It is not known whether findings from these studies would apply to more naturalistically variable face and voice tokens that would more resemble the faces and voice encountered in everyday life. Furthermore, the majority of studies discussed in this review used univariate fMRI methods to investigate modulations in brain activity that are associated with information extracted from faces and voices. While univariate fMRI methods can reveal brain regions that are sensitive to a given type of information in faces or voices, they are not able to show whether such regions use this information to distinguish between individual faces or voices. Therefore, less is known regarding the type of information that may be used by different face-selective and voice-selective regions to discriminate individual faces and voices. Finally, virtually nothing is known regarding the informational content of face and voice representations in multimodal brain regions that respond to both faces and voices, and it is possible that these regions represent information that can be extracted from both the face and the voice.

The next section focuses on how information that is extracted from a person's face relates to information extracted from their voice in terms of behavioural face and voice judgements, and how this relationship compares between familiar and unfamiliar people.

1.4 The relationship between perceived information from faces and voices

The previous section discussed the possibility of the existence of crossmodal representations of face and voice information that is perceived from both the face and the voice, such as physical characteristics and social traits (Yovel & Belin, 2013), in multimodal brain regions. If these representations exist at the level of individual person identities, it would require that the information perceived from a person's face should be consistent with the information perceived from their voice. For example, in order for a person's face and voice to elicit similar representations in a brain region that responds to trustworthiness, their face and voice should also elicit similar behavioural evaluations of trustworthiness. Regarding physical person-specific information, it has been shown that faces and voices convey highly consistent (r ≥.70) information regarding masculinity-femininity, health and height (Smith et al., 2016a). Specifically, Smith et al. (2016a) compared ratings of unfamiliar faces and voices on these characteristics and showed that they were highly correlated across the two modalities. However, little is known regarding the relationship between social information perceived from the face and the voice of the same person. The majority of studies investigating the social evaluation of faces and voices have focused on one of the two modalities, and it is not clear whether the face and voice of the same person convey consistent information regarding the social traits of that person.

For unfamiliar people, faces and voices cannot be associated with each other based on person-specific semantic knowledge, and a small number of studies that have compared ratings of faces to ratings of voices of unfamiliar people have shown largely inconsistent results. The majority of these studies have focused on judgements of attractiveness, and have shown weak to moderate (.20-.59) correlations between face and voice ratings which are often dependent on stimulus and/or participant gender, with some finding correlations for male participants rating female stimuli (Abend, Pflüger, Koppensteiner, Coquerelle, & Grammer, 2015; Valentova et al., 2017; Wells, Baguley, Sergeant, & Dunn, 2013), or for both genders rating opposite-sex stimuli (Lander, 2008). Other studies found correlations between face and voice attractiveness ratings when including stimuli of both genders in their analysis, but these correlations were no longer significant when splitting the analysis by gender (Rezlescu et al., 2015; Saxton, Burriss, Murray, Rowland, & Craig Roberts, 2009). Lastly, one study found no relationship between face and voice ratings on attractiveness (Oguchi & Kikuchi, 1997). Findings in regard to dominance are contradictory, whith one study finding a modate negative correlation (-.52) between face and voice ratings in male participants (Rezlescu et al., 2015), and another finding a low positive correlation (.37) for perceived threat potential, a component derived from ratings of dominance, strength, and body size for participants of both genders rating male stimuli (Han et al., 2017). Finally, in regard to trustworthiness, one study showed a moderate correlation (.47) between face and voice ratings, that was, however, no longer significant when male and female stimuli were analysed separately (Rezlescu et al., 2015).

A potential issue affecting the interpretation of findings relating the social evaluation of the face and voice in unfamiliar people is that, even within each modality, judgments may not be consistent for different instances of the face or voice of the same person. Studies have shown that the variance between the ratings of different face photographs of the same person on social judgements is equal or larger than the variance between ratings of different face identities that have been averaged across all photographs of each person's face (Sutherland, Young, & Rhodes, 2017; Todorov & Porter, 2014). Specifically, different images of the same person's face are evaluated differently, perhaps due to cues extracted from changeable aspects of the face, such as emotional expression and face viewpoint (Sutherland et al., 2017). In contrast, in regard to voices, Rezlescu et al., (2015) found moderate to high correlations between ratings of different vocalisations spoken by the same speakers on social judgements. However, Rezlescu et al. (2015) used vowel sounds as stimuli, and it is not known whether the evaluation of speech stimuli with longer durations (e.g. sentences), which better capture the natural variability in voices that we encounter in everyday life (Lavan, Burton, et al., 2018), would also be consistent across different tokens from the same speaker. Ultimately, given the variability in ratings of different images of the same face within modality, it is highly likely that variability in ratings of different face and voice tokens from the same person across modalities would be even greater, and may account in part for the inconsistencies between different studies.

It is possible that the different nature of the cues extracted from unfamiliar faces and voices results in the formation of largely independent associations between faces and voices and social traits. However, a second possibility is that, through the experience of other people in everyday life, joint associations are formed between specific facial and vocal features, such as face shape and voice pitch, and certain social traits, such as trustworthiness or dominance. Over and Cook (2018) recently proposed that trait evaluations of faces reflect mappings between locations in 'face space' and locations in 'trait space' that are learned through experience. For example, different encounters with people whose faces are a certain shape and who also display a certain trait can lead to an association between that particular face shape and trait. Based on Over and Cook's (2018) framework, it seems plausible that people could also learn associations between 'voice space' and 'trait space', and, given that exposure to faces and voices is usually concurrent during social interactions, people could additionally learn joint associations of facial features and vocal features with their corresponding traits. Thus, the evaluation of a person's face would be similar to the evaluation of their voice because of the knowledge that these particular face and voice features typically co-occur with certain traits. For example, if people perceived as dominant tend to have faces with a large jaw and low-pitched voices, both of these features will be associated with dominance.

Certain face and voice features may co-occur across different people because they reflect physical characteristics of a person such as masculinity-femininity, health, and height, that are conveyed concordantly by both the face and the voice (Smith et al., 2016a). Moreover, social judgements of faces and voices have been associated with some of these characteristics. For example, dominance has been associated with masculinity in both voices (McAleer et al., 2014) and faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013), and masculinity itself has been associated with influences of testosterone on facial appearance (Penton-Voak & Chen, 2004) and voice pitch (Dabbs & Mallinger, 1999; Evans & Davis, 2015). Therefore, it is likely that a person with high levels of testosterone will have a masculine facial and vocal appearance, and that both the face and voice of such a person would be evaluated as dominant. For attractiveness, evaluations of both faces and voices have been negatively associated with fluctuating asymmetry (Gangestad, Thornhill, & Yeo, 1994; Hughes, Harrison, & Gallup, 2002), a measure of the extent to which bilateral external and internal body structures are symmetric (Van Valen, 1962), and which is considered an indicator of health (Thornhill & Moller, 1997). In sum, it is likely that some social judgements are derived, at least in part, from physical characteristics that are conveyed by both the face and the voice (Smith et al., 2016a). This may encourage the learning of associations between facial features, vocal features, and social traits through experience.

Previous work investigating the relationship between the social evaluation of the face and the voice (Abend et al., 2015; Han et al., 2017; Lander, 2008; Oguchi & Kikuchi, 1997; Rezlescu et al., 2015; Saxton et al., 2009; Valentova et al., 2017; Wells et al., 2013) and the evaluation of physical person characteristics from faces and voices (Smith et al., 2016a) has used stimuli from unfamiliar people, and therefore virtually nothing is known regarding the relationship between information extracted from a familiar person's face and the information extracted from their voice. For familiar people, exposure to faces and voices during social interactions is often simultaneous, and these two sources of information are naturally associated through experience (Yovel & Belin, 2013). Moreover, faces and voices of familiar people both provide access to stored semantic knowledge about a person (Damjanovic & Hanley, 2007). Therefore, it is likely that similar judgements regarding social and physical person characteristics would be made from the face and the voice of a familiar person due to prior knowledge and experience of that person. Furthermore, in contrast to unfamiliar people, for whom social judgements of the face and voice are likely to be primarily influenced by physical face and voice characteristics, social judgements of familiar faces and voices are likely to be influenced mainly by prior knowledge of the person's character, and therefore judgements should be more similar between faces and voices for familiar people, compared with unfamiliar people. However, the influence of familiarity on the relationship between information perceived from the face and the voice has not yet been investigated.

Conclusion

Despite extensive knowledge of the social evaluation of faces and voices separately, very little is known regarding how the evaluation of a person's face relates to the evaluation of their voice, and how this relationship compares between familiar and unfamiliar people. For familiar people, judgements of the face and voice are likely to be similar due to prior knowledge of the person. For unfamiliar people, this relationship is complicated by the within-subject variability across different tokens of the same person's face and voice, and it is an open question whether the face and the voice convey redundant information about a person's social traits.

1.5 Organisation of thesis

This thesis attempts to contribute to the three main issues reviewed in this introduction, namely: (1) how face and voice information is integrated in the brain to form representations of person identity, (2) what is the informational content of face and voice representations and (3) how information extracted from the face relates to information extracted from the voice.

Chapter 2 describes representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), a multivariate data analysis method that will be used throughout this thesis. This method has the unique benefit of allowing comparisons of brain activity from different sensory modalities, here vision and audition. Moreover, it makes possible comparisons between brain representations and models of perceived or objective stimulus properties, to determine the informational content of these representations, and comparisons between behavioural judgements obtained using different types of tasks.

Chapter 3 aims to disentangle the relative contributions of the ‘multimodal’ integration mechanism, proposed by the MP model, and the ‘coupling’ integration mechanism, proposed by the CFVP model, to the integration of face and voice information. In an fMRI experiment, RSA was used to compare the multivoxel activity patterns elicited by the faces and voices of the same identities in independently localised face-selective, voice-selective, and multimodal brain regions. In contrast to the majority of previous studies that examined face, voice, and person identity representations, this study used naturalistically varying face and voice stimuli, and presented multiple tokens of the face and voice of each identity. Thus, this work aimed to identify representations that are robust to the natural within-person variability encountered in faces and voices in everyday life.

Chapter 4 investigates the informational content of face and voice representations in face-selective, voice-selective, and multimodal brain regions. It presents a study in which RSA was used to compare brain representations of individual faces and voices with models of both perceived and objective face and voice properties. The majority of previous work has focused on identifying modulations in the magnitude of the response of different brain regions to categorical levels of a single stimulus property (such as trustworthy or untrustworthy faces). This study will complement these findings by testing multiple stimulus properties that may be used by different brain regions to distinguish between individual stimuli based on their multivoxel response patterns.

Lastly, Chapter 5 examines the relationship between information extracted from the face and information extracted from the voice, and how this relationship compares between familiar and unfamiliar people. To overcome the issue of different tokens of the face and voice of a person eliciting different judgements, a novel paradigm is used in which face and voice ratings are based on multiple, naturalistically varying tokens of the face or voice. This study will inform the discussion on whether faces and voices convey concordant information about a person.

Chapter 2

Methodology: Representational similarity analysis (RSA)

The majority of data presented in this thesis were analysed using representational similarity analysis (RSA) (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, Mur, & Bandettini, 2008). This chapter introduces RSA by giving a general overview of this method, and explains the motivation for using RSA to address the aims of this thesis. A brief overview is then given to describe how RSA was used in each chapter.

2.1 Introduction to RSA

RSA was developed to address the problem of making comparisons between data obtained using different methods, and between data and computational models that describe different units (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, Mur, & Bandettini, 2008). For example, fMRI data is measured in voxels, which capture neural activity from multiple neurons, whereas many information-processing models make predictions about specific neurons. This greatly complicates the comparison between measured brain activity and model predictions. To overcome this problem, RSA abstracts from the individual measurement units by computing the dissimilarity between pairs of experimental conditions based on the data corresponding to each condition (Kriegeskorte, Mur, & Bandettini, 2008). It is then possible to compare whether two distinct data sources represent the dissimilarity between conditions in a similar manner.

In their seminal study, Kriegeskorte, Mur, Ruff, et al. (2008) used RSA to compare representations of images of objects in the human inferior temporal cortex, measured using fMRI, with representations of the same objects in the monkey inferior temporal cortex, measured using single cell recordings. This approach revealed a distinction between representations of animate and inanimate objects in the inferior temporal cortex that was strikingly similar across both species, suggesting that the human and monkey brain represent objects in a similar way. In the same study, Kriegeskorte, Mur, Ruff, et al. (2008) also used RSA to compare representations of objects in the human inferior temporal cortex with representations of the same objects in the human early visual cortex. This approach, which was termed “representational connectivity”, compares the representational dissimilarities of the same experimental conditions/stimuli between two different brain regions. Using this method, the authors revealed that pairs of stimuli that had similar representations in the inferior temporal cortex also tended to have similar representations in the early visual cortex, suggesting that these two brain regions represent objects in a similar way.

In addition to comparing brain representations across different species and imaging methods (Kriegeskorte, Mur, Ruff, et al., 2008), and across different brain regions (Guntupalli et al., 2016; Kriegeskorte, Mur, Ruff, et al., 2008; Pegado et al., 2018; Visconti Di Oleggio Castello et al., 2017), RSA has also been used to compare dissimilarities between brain representations of experimental conditions with perceived dissimilarities between the same conditions obtained from behavioural ratings (Charest & Kriegeskorte, 2015; Connolly et al., 2012; Hiramatsu, Goda, & Komatsu, 2011; Mur et al., 2013; Saarimaki et al., 2015; Saarimäki et al., 2018; Said, Moore, Engell, & Haxby, 2010; Sormaz, Watson, Smith, Young, & Andrews, 2016). For example, Mur et al., 2013 compared the dissimilarities between brain representations of objects in the human inferior temporal cortex, measured using fMRI, with participants’ judgements of the perceived similarity of the same objects. This comparison revealed that pairs of objects with similar brain representations in the inferior temporal cortex also tended to be perceived as similar. A further use of RSA is to compare perceived similarities between different conditions with the objective similarity between the conditions as predicted by different computational models or models of stimulus properties (Carlin & Kriegeskorte, 2017; Mur et al., 2013). In the current example, Mur et al. (2013) compared the perceived similarity of objects with the visual similarity of these objects as predicted by multiple models of low-level image properties such as luminance and colour.

Other studies have used RSA to directly compare brain representations to computational models or models of stimulus properties (Carlin, Calder, Kriegeskorte, Nili, & Rowe, 2011; Carlin & Kriegeskorte, 2017; Connolly et al., 2012; Guntupalli et al., 2017; Hiramatsu et al., 2011; Verosky et al., 2013; Weibert et al., 2018). For example, Connolly et al., 2012 showed that brain representations of images of animals in the early visual cortex were similar to a computational model of the properties of this region. RSA has also been used to compare brain representations elicited by different sensory modalities within the same brain regions (Chikazoe et al., 2014). For example, Chikazoe et al. (2014) compared representations of pleasant and unpleasant tastes and visual scenes, and showed that the OFC distinguished between pleasant and unpleasant stimuli independently from modalitiy. Lastly, RSA has been used to compare different behavioural judgements regarding the same experimental conditions (Stolier, Hehman, & Freeman, 2018). For example Stolier et al. (2018) compared representations of personality traits that were obtained based on judgements of faces and based on the measurement of stereotypes. In sum, RSA is a highly flexible method that can be used to investigate relationships between different types of data obtained using different methods and from different sources.

2.2 Representational dissimilarity matrices (RDMs) and comparing representational geometries

In multivariate fMRI, RSA is based on activity patterns across all voxels in a brain region of interest (ROI), which is elicited by an experimental condition or stimulus. An activity pattern is interpreted as a representation of that particular condition/stimulus in that particular ROI (Kriegeskorte, Mur, & Bandettini, 2008). The dissimilarities between the representations of all experimental conditions are arranged in a representational dissimilarity matrix (RDM).

RDMs (Figure 2.1) are computed using a measure of dissimilarity between data such as the correlation distance, the Euclidean distance, the Mahalanobis distance, or the linear discriminant contrast (LDC) (Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014; Walther et al., 2016). The correlation distance is computed as 1 minus the Pearson correlation between the multivoxel activity patterns in response to different conditions (Haxby et al., 2001; Kriegeskorte, Mur, Ruff, et al., 2008; Walther et al., 2016). The Euclidean distance is calculated as the square root of the sum of the squared differences between the activity patterns (Brooks & Freeman, 2018; Kriegeskorte, Mur, Ruff, et al., 2008; Walther et al., 2016). In a study comparing the reliability of multiple distance measures, i.e. the similarity of RDMs computed based on different subsets of the same data, Walther et al. (2016) showed that the correlation distance and the Euclidean distance were equally reliable. However, the authors noted that the interpretation of RDMs computed using the correlation distance is confounded by the fact that experimental conditions that elicit a high magnitude of brain activity are shown as more similar to each other compared with experimental conditions that elicit a lower magnitude of brain activity. In contrast, the computation of the Euclidean distance is not affected by the magnitude of the brain response. The Mahalanobis distance and the LDC are both based on the Euclidean distance (Walther et al., 2016). Specifically, the Mahalanobis distance is estimated by computing the Euclidean distance of activity patterns that have been normalised by their estimated multivariate noise covariance (Kriegeskorte, Goebel, & Bandettini, 2006; Walther et al., 2016), whereas the LDC is estimated by computing the Mahalanobis distance across different subsets of data though crossvalidation (Carlin & Kriegeskorte, 2017; Walther et al., 2016). Walther et al. (2016) showed that both multivariate noise normalisation and crossvalidation improve the reliability of RDMs, highlighting the benefits of using the LDC.

The number of rows and columns of a RDM is equal to the number of experimental conditions, and each cell contains a value that expresses the dissimilarity between the two conditions in the corresponding row and column (Kriegeskorte, Mur, & Bandettini, 2008). Figure 2.1 shows an illustration of a hypothetical RDM computed from the multivoxel activity patterns elicited by different faces in a hypothetical region of interest. RDMs provide a description of the representational geometry of the experimental conditions, i.e. their position within the multidimensional representational space defined by the units of measurement (Kriegeskorte & Kievit, 2013). For fMRI data, representational geometry refers to the position of each condition within the space defined by all voxels in a ROI. This position is based on the activity pattern across all voxels in response to a particular condition. A brain RDM thus provides information regarding the ability of a ROI to distinguish or group together different conditions in its representational space. RDMs can also be computed for measures of participant responses or model predictions regarding the same conditions. For example, an RDM for the same faces shown in Figure 2.1 could be computed from ratings of the faces on perceived pairwise similarity, or based on the location of features in the faces.

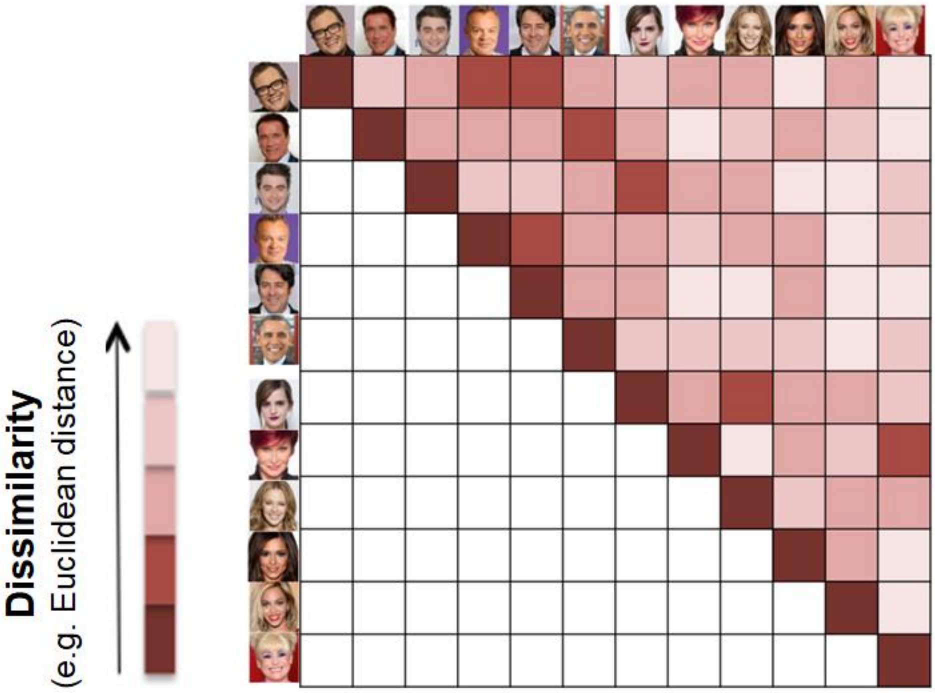


Figure 2.1: Hypothetical RDM. A hypothetical brain RDM for a hypothetical region of interest (ROI) showing the dissimilarity between the multivoxel activity patterns elicited by different faces. Each cell in the upper triangle shows the colour-coded dissimilarity between the activity patterns elicited by the faces in the corresponding row and column. Dissimilarities are represented on a dark-to-light colour scale for small-to-large dissimilarities. The diagonal features comparisons between identical faces, which are usually not of interest, and these cells are therefore assigned a value corresponding to the maximum possible similarity on the chosen scale.

Different types of RDMs (e.g. brain, behavioural, model) for the same conditions can be directly compared to assess the relationships between them (Kriegeskorte, Mur, & Bandettini, 2008). This is possible because the RDMs abstract from the original units of measurement (e.g. brain activity patterns, behavioural ratings) to represent the dissimilarities between the representations of different conditions. Comparing two RDMs reveals the extent to which the representational geometry of a set of conditions in one RDM is similar to the representational geometry of the same conditions in the other RDM. If the dissimilarities between different conditions in two RDMs are similar, it is likely that the information driving the representations of those conditions is also similar. For example, if two faces elicit similar response patterns in a given brain region, and those faces are also rated as being visually similar, it is likely that that brain region represents the perceived visual similarity between faces. Model RDMs of predicted dissimilarity relationships between different conditions in the brain can also be computed, in order to determine the extent to which those predictions explain the dissimilarity relationships between different conditions in the brain.

Comparisons between different types of RDMs are commonly made using a correlation type measure, such as Pearson's correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau a rank correlation coefficient (Nili et al., 2014). Pearson's correlation coefficient (or 1 minus the correlation) is used when two RDMs are expected to show a linear relationship (Kriegeskorte, Mur, Ruff, et al., 2008; Walther et al., 2016). In contrast, Spearman's and Kendall's rank correlation coefficients (or 1 minus the correlation) are used when a linear relationship between two RDMs cannot be assumed due to comparing different types of data, e.g. brain RDMs and model RDMs (Kriegeskorte, Mur, & Bandettini, 2008; Walther et al., 2016). Kendall's tau a coefficient has been shown to perform better than Spearman's rank coefficient when comparing data with tied ranks (such as model RDMs) and data without tied ranks (such as brain RDMs) (Nili et al., 2014).

2.3 Motivation for the use of RSA in the current thesis

The first aim of this thesis, which was to determine how information for the face and the voice is integrated to form a representation of person identity, was addressed by using RSA to directly compare representations of faces and voices in face-selective, voice-selective, and multimodal regions across the brain. Comparisons of brain activity elicited through different sensory modalities in fMRI studies are complicated by the different nature and noise levels of the resulting brain activation. Univariate fMRI studies are not able to directly compare brain activity elicited by individual faces and voices because individual stimuli cannot be differentiated based on differences in activation magnitude (Formisano et al., 2008; Kriegeskorte et al., 2007). MVPA studies have compared multivoxel activity patterns elicited by the face and voice of the same identities by testing whether the activity patterns elicited by a pair of face identities can be distinguished based on the activity patterns elicited by their corresponding voice identities, using crossmodal pattern classification (Anzellotti & Caramazza, 2017; Hasan et al., 2016). However, pattern classification does not provide information on how representations of multiple face or voice identities are related to each other, and on the extent to which the relationships between identities in one modality are similar to the relationships between the identities in the other modality. In contrast, RSA can be used to directly compare the representational geometry of a set of face identities with the representational geometry of the corresponding voice identities. This is possible because RSA abstracts from the data, i.e. the multivoxel activity patterns, to create RDMs. A comparison between representational geometries for face and voice identities can reveal the extent which the geometries are similar, and therefore the extent to which a given brain region represents person identity information that is common to both faces and voices and therefore independent from modality. To the best of my knowledge, RSA has not been used previously to compare representations of faces and voices in the brain.

In addition to the comparison of representational geometries of faces and voices, crossmodal representations of person identity were investigated by using RSA to compute the crossmodal discriminability of identities in one modality based on information in the other modality. This approach was possible using the LDC distance measure, which typically involves computing discriminants between pattern estimates for pairs of conditions in a subset of the data, and evaluating those discriminants on a different subset of the data for crossvalidation (Carlin & Kriegeskorte, 2017; Walther et al., 2016). In the current thesis, the LDC was also used to compute pattern discriminants in one modality and test them on the other modality, obtaining a measure of crossmodal discriminability. To the best of my knowledge, this is the first time that the LDC had been used to examine crossmodal discriminabily. This analysis shares some similarities with crossmodal pattern classification analyses, such as those used by Anzellotti & Caramazza (2017) and Hasan et al. (2016) to compare activity patterns in response to faces and voices. The main benefit of using the LDC, as opposed to a pattern classifier, is that the LDC is a continuous measure of discriminability that provides information about the extent to which two patterns are discriminable, whereas classifiers produce a binary output that merely indicates whether a given pattern can, or cannot, be correctly classified as belonging to one of two pre-defined categories (Walther et al., 2016). Moreover, a study comparing the reliability of the two methods across different splits of the same data sets showed higher reliability for the LDC (Walther et al., 2016).

The second aim of the present thesis, which was to determine where in the brain the different types of information conveyed by faces and voices are processed, was addressed by using RSA to compare representational geometries for faces and voices in face-selective, voice-selective, and multimodal regions with model geometries describing different types of information that can be extracted from the face and the voice. RSA has been used previously to compare brain representations of faces in face-selective regions with computational models of low-level visual properties, leading to informative insights into the computations of these regions (Carlin & Kriegeskorte, 2017; Weibert et al., 2018). However, aside from their visual properties, faces convey information related to a person's identity, such as their gender and personality. Therefore, in the present thesis RSA was used to compute multiple models of perceived social and physical information from faces and voices derived from participants' ratings, as well as models of objective face and voice properties. To the best of my knowledge, RSA has not been used before to compare brain representations of voices to models describing information extracted from the voice.

Lastly, although not directly related to the main aims of this thesis, RSA was used to compute RDMs based on ratings of individual faces or voices on social traits, in order to test whether dissimilarities between the ratings of pairs of faces/voices on social traits could be explained by their ratings on perceived visual/auditory similarity, obtained from pairwise similarity ratings tasks. Thus, ratings from two different types of behavioural tasks involving the same stimuli could be directly compared.

2.4 Applications of RSA in the current thesis

2.4.1 Comparing RDMs between faces and voices (Chapter 3)

In an fMRI study presented in Chapter 3, RSA was used to compare RDMs of the faces of familiar people with RDMs of their voices in face-selective, voice-selective, and multimodal regions across the brain. For each ROI, one RDM was computed based on the multivoxel activity patterns elicited by familiar faces (similar to Figure 2.1), and a second RDM was computed from the activity patterns elicited by the corresponding voices of the same identities. In both of these RDMs, the rows and columns corresponded to the same individuals (e.g. row and column 1 in both the face RDM and the voice RDM corresponded to identity 1), making the two RDMs directly comparable in terms of person identity. The aim of this analysis was to identify brain regions in which the geometries for faces and voices are similar, by testing whether pairs of identities that elicit similar response patterns in one modality also elicit similar response patterns in the other modality. Matching face and voice geometries for the same identities would indicate that the activity patterns elicited by the face and voice of each individual identity are similar. In other words, a brain region with matching representational geometries for faces and voices shows similar representations of the face and voice of the same identity, suggesting that it recognises and processes person identity regardless of the input modality.

The LDC distance measure was used to compute the face and voice RDMs. In contrast to other commonly used distance measures, such as 1-correlation or Euclidean distance, the LDC performs multivariate noise normalisation on the activity patterns by taking into account the noise covariance between all voxels in a ROI (Walther et al., 2016). As mentioned previously, multivariate noise normalisation has been shown to improve the reliability of brain RDMs (Walther et al., 2016). A further advantage of using the LDC is that a value of zero can be interpreted as the true absence of discriminability between two activity patterns (Walther et al., 2016). This similarity cannot be attributed to noise covariance because the crossvalidation across different partitions of the data renders the noise between the partitions independent. Thus, under the null hypothesis the LDC is symmetrically distributed around zero and unbiased. RDMs computed using the LDC show the representational geometry in terms of the discriminability between the activity patterns in response to different conditions in a ROI.

Finally, for the comparisons between face and voice RDMs, Pearson's correlation coefficient was selected because face and voice RDMs were compared within the same brain regions, and were therefore likely to show a linear relationship. A correlation between face and voice RDMs was computed for each ROI.

2.4.2 Investigating crossmodal discriminability (Chapter 3)

In the fMRI study that was presented in Chapter 3, in addition to the comparison of face and voice RDMs, a second method was used to investigate representations of person identity, which involved computing crossmodal RDMs between faces and voices in ROIs. To compute crossmodal RDMs using the LDC, the activity patterns of identity pairs in one modality were used to create a linear discriminant, which was then applied to differentiate the activity patterns of the same identity pairs in the *other* modality. The representational geometry of these crossmodal RDMs shows how discriminable the representations of different identities are in one modality based on information that discriminated their representations in the other modality. In other words, crossmodal RDMs show the degree to which pattern discriminants for each pair of identities generalise from one modality to the other.

The LDC provides a continuous measure of discriminability for each pair of conditions, whereby an LDC value of zero or lower indicates no discriminability, and higher values indicate higher discriminability (Nili et al., 2014; Walther et al., 2016; Carlin and Kriegeskorte, 2017). Due to this property, it is possible to calculate the mean LDC value across all cells in an RDM to determine the overall ability of a ROI to discriminate between the activity patterns in response to the different conditions. Mean LDC values across participants can then be subjected to random-effects inference comparing against zero. In this study, mean LDC values were calculated for crossmodal RDMs to determine the overall ability of each ROI to discriminate between different identities based on crossmodal information. Mean LDC distances for each ROI across participants were then compared against zero using one-sample t-tests. A brain region that shows crossmodal discriminability should display similar representations of the face and voice of the same identity, and therefore should be able to recognise and process person identity regardless of the input modality.

This analysis complements the previously described analysis comparing face and voice representational geometries, and overcomes some of the constraints of the latter. Specifically, the analysis comparing representational geometries is constrained by two assumptions. The first assumption is that there is sufficient variability in the representational distances between different identities within-modality, i.e. different degrees of dissimilarity between identities. If all identities are equally distinct from each other, it would not be expected to find correlations between geometries across the two modalities. The second assumption is that modality-general information dominates over any modality-specific information that may be present in the same voxels. Specifically, it is possible that the voxels comprising the pattern estimates contain both unisensory and multisensory neurons (Driver & Noesselt, 2008; Laurienti et al., 2005; 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. In contrast to the analysis comparing representational geometries, the analysis investigating person identity discriminability 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 is likely to be more sensitive to detecting modality-general person identity representations even in the presence of modality-specific information.

2.4.3 Comparing face and voice RDMs to model RDMs (Chapter 4)

In a study presented in Chapter 4, RSA was used to compare the brain RDMs for faces and voices that were computed in the study described in Chapter 3 to model RDMs that were computed from perceived and objective characteristics of the same faces and voices. The aim of this analysis was to characterise the informational content of the brain representations of faces and voices in different ROIs by comparing brain representational geometries for faces or voices with model representational geometries. Model RDMs of perceived face or voice characteristics were computed based on ratings of faces and voices on perceived trustworthiness, dominance, attractiveness, positive-negative valence, and pairwise visual/auditory similarity. Model RDMs of objective face or voice characteristics were computed for faces based on computational models describing face properties, and for voices based on different acoustic measures. A binary model for face and voice gender was also constructed. Model RDMs, with the exception of perceived similarity and gender, were computed using Euclidean distance.

Each candidate model RDM was compared to a reference brain RDM using Kendall's tau a rank correlation coefficient, which was the most suitable similarity measure because of the different number of expected tied ranks in brain RDMs and in model RDMs. Similar representational geometry between a brain RDM and a model RDM would indicate that pairs of stimuli that are distinguishable in a given brain region based on their activity patterns would also be dissimilar in terms of the property described by the model (e.g. they would differ in their perceived trustworthiness). Put simply, it would suggest that the brain region processes the type of information described by the model.

2.4.4 Comparing face/voice RDMs between different brain regions (Chapter 4)

In the study presented in Chapter 4, an exploratory analysis also used RSA to compare the "representational connectivity" (Kriegeskorte, Mur, Ruff, et al., 2008; Kriegeskorte, Mur, & Bandettini, 2008) of face or voice RDMs across different ROIs (within the same modality) to assess their similarity in terms of information content. Face or voice RDMs in every ROI were compared with the same-modality RDM in every other ROI using Spearman's rank correlation. This distance measure was chosen because, although a linear relationship is not expected between RDMs across different brain regions, these RDMs are likely to contain a similar number of tied ranks. Although this analysis does not reveal the type of information being shared across regions, it complements the previously described analysis, which attempts to characterise the informational content of brain RDMs by comparing them to different model RDMs, by revealing which ROIs process similar information from faces or voices.

2.4.5 Comparing ratings of social traits with ratings of pairwise similarity (Chapter 5)

In a study presented in Chapter 5, RSA was used to compare ratings of faces/voices on social traits, namely trustworthiness, dominance, attractiveness, and positive-negative valence, with ratings of the faces/voices on pairwise visual/auditory similarity. For this analysis, RDMs were computed for each trait, separately for faces and voices, using the Euclidean distance between ratings of all pairs of identities. For each modality, the RDMs for each trait were each compared with the ratings of pairwise similarity using Spearman correlation, given that a linear relationship cannot be assumed between different dissimilarity scales. These comparisons revealed whether pairs of face/voice identities that were given similar ratings on social traits were also perceived as being visually/acoustically similar.

Chapter 3

Crossmodal representations of person identity in the brain

This chapter addresses the first aim of this thesis, which was to determine how the brain integrates information from the faces and voices of familiar people to represent person identity. It describes a study that aimed at disentangling the relative contributions of the ‘multimodal’ integration mechanism, proposed by the MP model, and the ‘coupling’ integration mechanism, proposed by the CFVP model, to the integration of face and voice information (Blank et al., 2011; Campanella & Belin, 2007; Yovel & O’Toole, 2016). Briefly, the MP Model proposes that information from faces and voices is integrated in multimodal brain regions, and lesion studies (Ellis et al., 1989; Gainotti, 2011) and fMRI studies (Anzellotti & Caramazza, 2017; Hölig et al., 2017; Joassin et al., 2011; Shah et al., 2001) suggest the ATL, the pSTS, the angular gyrus, the retrosplenial cortex, and the hippocampus as candidate multimodal regions. In contrast, the CFVP proposes that the direct coupling between face- and voice-responsive brain regions is crucial for the integration of person identity information (von Kriegstein et al., 2005). In particular, 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., 2008, 2006, 2005; von Kriegstein & Giraud, 2006).

A recent MVPA study found support for the MP model by showing that a multimodal region in the right pSTS could discriminate between the activity patterns elicited by a pair of familiar faces based on the activity patterns elicited by their corresponding voices, and vice-versa (Anzellotti & Caramazza, 2017). However, as discussed In the Introduction chapter, this study was limited in its ability to show that these crossmodal representations of person identity generalise to different tokens of the face and the voice of each identity, and presented just two tokens for each identity. Therefore, this study could not rule out the possibility that the observed person identity representations were in some way specific to the particular face and voice stimuli that were presented in the study. Moreover, the presented faces and voices were constrained in terms of their natural variability, and it is not certain whether the observed crossmodal representation in the pSTS would be robust to more naturalistically varying face and voice stimuli.

In the present study, multivoxel fMRI activation patterns were measured in response to the faces and voices of 12 famous individuals. It was important to use highly familiar individuals because of the need to guarantee that participants were well acquainted with the faces and voices of those individuals. Thus, participants were only recruited for the full study if they demonstrated that they were familiar with the majority of the famous individuals in an online recognition task. In addition, and in contrast to previous studies, this study presented multiple, naturalistically varying face videos and voice recordings of 12 different identities. Thus, this study was able to sample the variability of visual and auditory appearance encountered in everyday life, and to better capture processes of person identification, which are distinct from image or sound recognition (Burton, 2013).

In order to directly compare the mechanisms proposed by the MP model and the CFVP model, the present study localised face-selective, voice-selective, and multimodal brain regions, and tested for crossmodal representations of person identity within each of these regions. Specifically, RSA was used to (a) compare the representational geometries of faces and voices in each region, and to (b) test whether linear discriminants computed based on data in one modality could discriminate pairs of identities in the other modality. The expectation was that, if a region shows a crossmodal person identity representation, the representational geometry of face and voice identities will match, and/or pattern discriminants will generalise across faces and voices. The MP model predicts that crossmodal person identity representations will exist in multimodal brain regions, and the CFVP predicts that these representations will also exist in face-selective and voice-selective brain regions.

3.1 Methods

3.1.1 Overview of study

Participants were recruited after completing an online Recognition Task to demonstrate that they were familiar with the famous individuals used as stimuli. The full study then consisted of two MRI scanning sessions and one behavioural session, with each session taking approximately 90 minutes. All three sessions took place on separate days. Before entering the scanner at the start of the first MRI session, participants repeated the Recognition Task in the presence of the experimenter and also completed a Familiarity Task in which they rated all face and voice stimuli on perceived familiarity.

In each MRI session participants completed three functional runs (main experimental runs) in which they viewed the faces and listened to the voices of the famous people in an event-related design. In addition, participants underwent two structural scans (one in each session) and functional localisers for face-selective, voice-selective, and multimodal regions of interest (ROIs). 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 multimodal (face-voice) localiser, and (4) a voice localiser. Finally, participants completed a behavioural testing session (the methods and results of this session will be presented in chapter 4).

To investigate the existence of crossmodal person identity representations in each of our ROIs, RSA was used 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). Analysis A focused on the representational geometry of all of the 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. Analysis B focused on the discriminability of pairs of identities, and 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). As discussed in the Method chapter, these two analyses complement each other and allow the testing of different predictions regarding the nature of crossmodal person identity representations. For Analysis A (RSA comparing representational geometries), it was predicted that brain regions with crossmodal person identity representations would show matching representational geometries for face identities and voice identities. For Analysis B (RSA investigating identity discriminability), it was predicted that brain regions with crossmodal person identity representations would be able to discriminate between pairs of identities in one modality based on their representational distance in the other modality.

3.1.2 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 (see below) as part of the screening procedure for the study and were only invited if they were able to recognise at least 75% of a set of famous people from both their face and their voice.

Thirty-one healthy 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 males) 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 (see Appendix A).

Recognition Task

Participants completed a face and voice Recognition Task to determine whether they could recognise at least 75% of the famous people (i.e. at least 9 out of 12) from both the face and the voice. Face stimuli consisted of single photographs of each of the 12 famous people that were obtained from the Internet through Google Image searches. Photographs included the top part of the body and were front-facing. Voice stimuli consisted of single sound-clips for each of the 12 famous people and were obtained from YouTube videos. Sound-clips were approximately 8-seconds long and were root-mean-square (RMS) normalized using Praat (version 5.3.80; Boersma and Weenink, 2014; www.praat.org). None of these face or voice stimuli were presented in the main experiment.

Stimuli were presented using Qualtrics (Qualtrics, Provo, UT). For each stimulus participants had to identify the person shown in the picture or the person speaking (by providing their name or other uniquely identifying biographical information). In the online task participants typed their responses below each stimulus, and in the lab task responses were made verbally.

3.1.3 Stimuli

Six silent, non-speaking video clips of moving faces, and six recordings of voices for each of the 12 famous people (six female, six male) were obtained from videos on YouTube (in total, 72 stimuli per modality). These people had been identified in previous pilot studies conducted in the lab 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. 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.

It has been shown that famous voices are harder to recognise than famous faces (Damjanovic & Hanley, 2007; Hanley & Damjanovic, 2009). Therefore, a pilot experiment was conducted to determine the minimum amount of time that participants needed to listen to a recording of the voice of the 12 famous people in order to be able to reliably identify them. Ideally, the aim for the fMRI study was to present as many stimuli as possible in the shortest amount of time possible in order to maximise design efficiency. In the experiment, participants (*N*=8) were presented with recordings of the voices of the 12 famous people that featured three, five, or seven seconds of speech. Three different voice recordings were presented for each of the three time duration conditions (nine stimuli per person). After listening to each recording participants were asked to verbally identify the person by name or other uniquely identifying biographical information. The average number of voices that were correctly identified in each time duration condition (out of a total of 36 voices) was calculated for each participant, and then averaged across participants. The results showed that participants recognised an average of 32.6 (*SD*=3.54) voices in the 3 s condition, 33.5 (*SD*=2.62) in the 5 s condition, and 32.6 (*SD*=3.07) in the 7 s condition. There were no significant differences in recognition accuracy between the three duration conditions [*F*(2,14)=2.27, *p*=.140], and therefore a duration of three seconds was chosen for the voice recordings to be presented in the fMRI experiment.

The face videos 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 videos were edited using Final Cut Pro X (Apple, Inc.) so that they were three seconds 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 recordings were edited using Audacity® 2.0.5 recording and editing software so that they contained three seconds of speech after removing long periods of silence. The recordings were then converted to mono with a sampling rate of 44100, low-pass filtered at 10KHz, and RMS normalised using Praat. Six recordings of the voice of the same person were obtained from different original videos. All of the voice recordings had a different verbal content and were non-overlapping. The recordings were selected so that the speakers' identity could not be determined based on the verbal content, conforming to the standards set by Van Lancker, Krieman, & Emmorey (1985) and Schweinberger, Herholz, & Sommer (1997).

Familiarity Task

Before entering the scanner, participants rated all stimuli that would be presented in the main experimental runs on perceived familiarity. Participants were presented with the face stimuli first, followed by the voice stimuli, in separate blocks. Stimuli were presented using the Psychophysics Toolbox (version 3; Brainard, 1997; Pelli, 1997) running in Matlab (version R2013b; MathWorks). Face stimuli were presented in the centre of the screen. Participants listened to the voice stimuli through headphones (Sennheiser HD 202). Participants rated each stimulus on scale from 1 (very unfamiliar) to 7 (very familiar). Each block took approximately 5 minutes to complete.

3.1.4 MRI data acquisition and pre-processing

Participants were scanned using a 3.0 Tesla 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 x 1.0 in-plane resolution; slice thickness, 1.0mm; 176 axial interleaved slices; PAT, Factor 2; PAT mode, GRAPPA (GeneRalized Autocalibrating Partially Parallel Acquisitions); repetition time (TR), 1900ms; echo time (TE), 3.03ms; flip angle, 11°; matrix, 256x256; field of view (FOV), 256mm].

For all functional runs T2*-weighted whole-brain functional scans were acquired using echo-planar imaging (EPI) [3.0 x 3.0 in-plane resolution; slice thickness, 3.0mm; PAT, Factor 2; PAT mode, GRAPPA (GeneRalized Autocalibrating Partially Parallel Acquisitions); 34 sequential (descending) slices; repetition time (TR), 2000ms; echo time (TE), 30ms; flip angle, 78°; matrix, 64x64; field of view (FOV), 192mm]. 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 multimodal localiser runs we obtained 227 brain volumes.

Data were pre-processed using Statistical Parametric Mapping (SPM12; Wellcome Department of Imaging Science, London, UK; http://www.fil.ion.ucl.ac.uk/spm) operating in Matlab. Pre-processing was performed separately for each scanning session. All runs within each session (main experiment or localizer 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 were in the same space.

3.1.5 Functional localisers

TVA localiser

The TVA localiser developed by Belin et al. (2000), which contains vocal and non-vocal auditory stimuli, was used to localise the TVAs. Stimuli were presented in 40 blocks of 8 seconds 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 minutes.

Face, Voice, and Multimodal localisers

New face, multimodal, and voice localiser runs were created that shared the same experimental design and presented stimuli from comparable categories (people and objects/scenes). Importantly, these localisers used videos and not static images of faces. Dynamic face stimuli have been shown to be more effective that 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 multimodal localiser the stimuli were audio-visual 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). The voice localiser 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.

Videos (640 x 360 pixels) were presented at the centre of the screen. The screen resolution was 1024 x 768 pixels, and from a distance of 85 cm, the videos subtended 20.83 x 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. Each stimulus lasted 8 seconds and each run presented 48 stimuli. 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-second 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 multimodal 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 minutes.

To identify face-selective (face localiser), voice-selective (voice localiser and TVA localiser), and people-selective (multimodal localiser) brain regions, mass univariate time-series models were computed for each participant. Regressors modelled the blood-oxygenation-level-dependent (BOLD) response following the onset of the stimuli and were convolved with a canonical hemodynamic response function (HRF). We used a high-pass filter cutoff of 128 seconds, and autoregressive AR(1) model to account for serial correlations. For the face, voice, and multimodal localisers there were three experimental regressors: (1) famous faces/voices/people, (2) non-famous faces/voices/people, and (3) objects and scenes. For the TVA localiser there were two experimental regressors: (1) voices and (2) non-voices. For all localisers six head motion parameters computed during realignment were included as covariates. Selectivity was defined with a *t*-test contrasting the responses to faces/voices/people (famous and non-famous) *versus* responses to the control stimuli.

3.1.6 ROI definition

Probabilistic maps from previous studies were used to define regional masks in which our regions of interest (ROIs) were predicted to be located. ROIs were then defined by extracting all selective voxels within those regional masks for each participant. This approach is similar to the one implemented by Julian, Fedorenko, Webster, & Kanwisher (2012) and avoids experimenter biases in ROI definition.

Probabilistic maps were thresholded to only show voxels that were present in 20% of the participants and binarised to create regional masks. We used a probabilistic map of the TVAs created by Pernet et al. (2015) and obtained from neurovault (http://neurovault.org/images/106/) to create separate masks for the right and left TVA (rTVA, lTVA). For all other regional masks, we used probabilistic maps that were obtained from a previous study conducted in the lab (unpublished data). In this previous study, 22 participants were tested using the same face and voice localisers as the current study (the multimodal localizer was not used in this previous study). We defined face-selective and voice-selective *t*-test images for each participant, thresholded each image at *p*<.05 (uncorrected), binarised the resulting image, and summed all images across participants to create face-selective and voice-selective probabilistic maps. In cases where there was some overlap between the masks for different regions we manually defined the borders of these masks using anatomical landmarks.

Regional masks of face-selective regions were created for the right fusiform face area (rFFA), the right occipital face area (rOFA), and the right posterior superior temporal sulcus (rpSTS). Regional masks of voice-selective regions were created for the right and the left superior temporal sulcus and gyrus (rSTS/STG, lSTS/STG). Regional masks of multimodal regions were created based on joint face-selective and voice-selective probabilistic maps. These masks were created for a number of regions that showed *both* face-selective and voice-selective responses in most participants: precuneus/posterior cingulate, orbitofrontal cortex (OFC), frontal pole (FP), 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 did not create a mask of the multimodal STS using this method due to the voice-selective STS region being much larger than the face-selective STS region. However, there was large overlap between the mask of the face-selective rpSTS and the masks of the rSTS/STG and rTVA, suggesting that this face-selective rpSTS region also responds to voices.

All of the regional masks (in MNI space) were registered and resliced to each participant’s native space using FSL (version 5.0.9; Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). These masks were then used to extract ROIs from the *t*-test maps obtained from the contrasts of interest from the face, voice, TVA, and multimodal localisers from the current study. All voxels that fell within the boundaries of the mask and that were significantly activated at *p*<.001 (uncorrected) were included in the subject-specific ROI. If there were fewer than 30 voxels at *p*<.001 the threshold was lowered to *p*<.01 or *p*<.05. If we could not define 30 selective voxels even at *p*<.05, the ROI for that participant was not included in the analyses. It was required that all ROIs be present in at least 20 participants (out of 30).

3.1.7 Main experimental runs: Experimental design

Design and procedure

Face and voice stimuli were presented using the Psychophysics Toolbox via a computer interface inside the scanner (Figure 3.1). Face and voice clips of all 12 identities were intermixed within each run. A fixation point was always present and participants were asked to fixate. The videos were 640 x 360 pixels and, from a viewing distance of 85cm, videos subtended 20.83 x 12.27 degrees of visual angle. The six face videos and the six voice recordings for each of the 12 identities were evenly distributed among the three runs so that each run contained two different videos of the face and two different recordings of the voice of each identity. Each individual stimulus was presented twice within each run. Therefore, in each run there were 96 experimental trails (48 face trials, 48 voice trials) in total.

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.

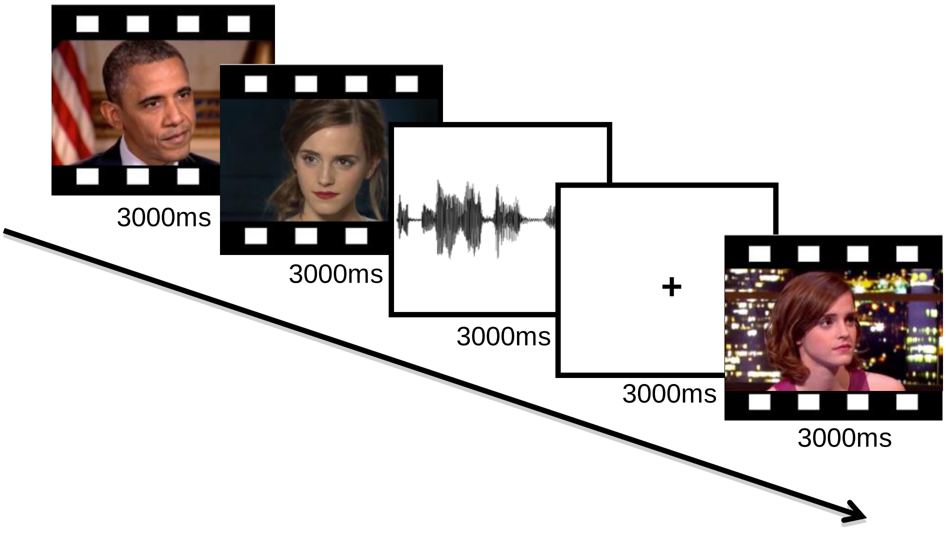


Figure 3.1: **Example trial sequence and stimuli for fMRI experiment.** Please note that there was a 1000ms inter-trial-interval that is not depicted in this figure. Therefore, the total time from the start of one trial to the start of the next trial was 4000ms.

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 three seconds with a SOA of four seconds. 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 minutes.

The presentation order of the three runs 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.

3.1.8 Main experimental runs: Statistical analysis

General linear models

Mass univariate time-series models were computed for each participant. Models were defined separately for each scanning session and each experimental run (six runs in total). 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 seconds and autoregressive AR(1) model to account for serial correlations. 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 and six head motion parameters computed during realignment were included as regressors of no interest.

As part of the crossvalidation procedure used in the RSA 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).

Mean response to faces and voices in ROIs

We conducted an analysis to characterise the responses to faces and voices 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 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*<.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*<.05.

Analysis A: RSA comparing representational geometries

For this analysis we computed representational dissimilarity matrices (RDMs) for faces and voices separately for each participant, each scanning session and each ROI. All analyses were performed using in-house Matlab code and the RSA toolbox (Nili et al., 2014). RDMs were computed using the LDC between the pattern estimates (beta estimates across all voxels within an ROI) elicited by the different face or voice identities. To calculate the LDC, crossvalidation was performed using a leave-one-run-out procedure between runs that presented different tokens of the face and voice of each identity. This procedure ensured that face and voice representations reflected face and voice *identity*, rather than specific face videos and voice recordings. In each crossvalidation fold 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). Multivariate noise normalisation was applied by computing a noise variance-covariance matrix based on the residual time courses obtained from the general linear model (GLM) 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 & 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 an LDC value showing the discriminability of the two identities. The resulting RDMs with LDC values from each crossvalidation fold were averaged to create a single RDM.

The resulting 12x12 RDMs 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. 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 (Figure 3.5).

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 (Figure 3.5). 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 scanning 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 $q < .05$. Correlations between face and voice RDMs that are significantly greater than zero would indicate that that a region contains crossmodal person identity representations.

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 LDC to compute a linear discriminant based on the activity patterns of identity pairs in one modality, and then applied the discriminant to the activity patterns of the same identity pairs in the other modality. 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. We then calculated the mean LDC value across all cells in each RDM and each ROI to determine the overall ability of that ROI to discriminate between identities.

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 RDMs that were created in the previous analysis. For this analysis the corresponding RDMs (e.g. faces session 1 and faces session 2) for each scanning session were averaged across the two sessions, and then the mean LDC across the vectorised matrix was calculated.

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. 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*<.05. Mean LDC distances that are significantly greater than zero in crossmodal RDMs would indicate that a region contains crossmodal person identity representations. Moreover, LDC distances that are significantly greater than zero in face or voice RDMs would indicate a region contains representations of face or voice identity.

Exploratory whole-brain searchlight analyses

Despite including a broad range of functionally defined ROIs, it is possible that crossmodal person identity representations may exist in brain regions not covered by these 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 the main ROI analyses. We note that we focused solely on crossmodal person identity representations in this exploratory analysis, as that was the main aim of this study.

For each participant we created 6mm 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 6mm 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, Ridgway, Webster, Smith, & Nichols, 2014) in FSL for inference on the resulting statistical maps (5000 sign-flips). Clusters were identified with threshold-free cluster enhancement, and p-values were corrected for multiple comparisons (FWE < 0.05).

3.2 Results

3.2.1 Familiarity ratings

Familiarity ratings of both faces and voices were high (Faces: *M* = 6.28, *SD* = 0.5; Voices: *M* = 6.2, *SD* = 0.49). Average familiarity of each identity's face and voice are shown in Table 3.1.

Table 3.1: Familiarity ratings of the face and voice of each identity. Ratings are averaged across participants and show the mean (*M*) rating of the face and voice of each identity across all face videos and all voice recordings of that identity, the standard deviation (*SD*) of participants' ratings of each identity, and the range of mean ratings for the six face tokens and six voice tokens for each identity. The rating scale ranged from 1 (very unfamiliar) to 7 (very familiar).

		AC	AS	BO	DR	GN	JR	BK	BW	CC	EW	KM	SO
Faces	M	6.48	5.72	6.94	6.76	6.27	6.42	6.36	5.65	6.49	6.73	5.25	6.32
	SD	0.75	1.19	0.2	0.59	0.85	0.59	0.94	1.45	0.79	0.45	1.48	0.92
	Token	6.37-	4.83-	6.90-	6.67-	5.87-	6.17-	6.07-	5.33-	6.37-	6.47-	4.6-	6.17-

	range	6.60	6.13	7	6.87	6.5	6.57	6.57	5.9	6.60	6.9	5.57	6.47
Voices	M	6.59	5.66	6.73	6.69	6.37	6.54	6.23	5.54	6.63	6.07	5.3	6.02
	SD	0.54	1.48	0.63	0.57	0.77	0.71	1.04	1.74	0.74	0.94	1.45	1.03
	Token range	6.37-6.83	5.43-5.87	6.7-6.8	6.57-6.8	6.07-6.67	6.3 6.7	6.07-6.37	5.47-5.67	6.4-6.77	4.6-6.53	5-5.53	5.5-6.43

3.2.2 ROI definition

Using functional localisers we defined face-selective ROIs (rFFA, rOFA, rpSTS), voice-selective ROIs (rSTS/STG, rTVA, lSTS/STG, lTVA), and multimodal ROIs (OFC, FP, rTP-aIT, lTP-aIT, Prec./P.Cing. [including the retrosplenial cortex]) in each participant. We were able to localise these ROIs with at least 30 voxels in all 30 participants, except for the face-selective rFFA (28 participants) and rOFA (29 participants), the Prec./P.Cing. (26 participants), and the OFC (21 participants). 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 multimodal rTP-aIT ROIs. In addition, the voice-selective ROIs in the left hemisphere (lTVA, lSTS/STG) overlap with each other and with the multimodal lTP-aIT ROI. For visualisation purposes only, probabilistic maps of all ROIs were created by normalising the single subject ROIs to MNI space and summing them. Figure 3.2 shows these maps thresholded to display all voxels that were present in at least 20% of the participants.

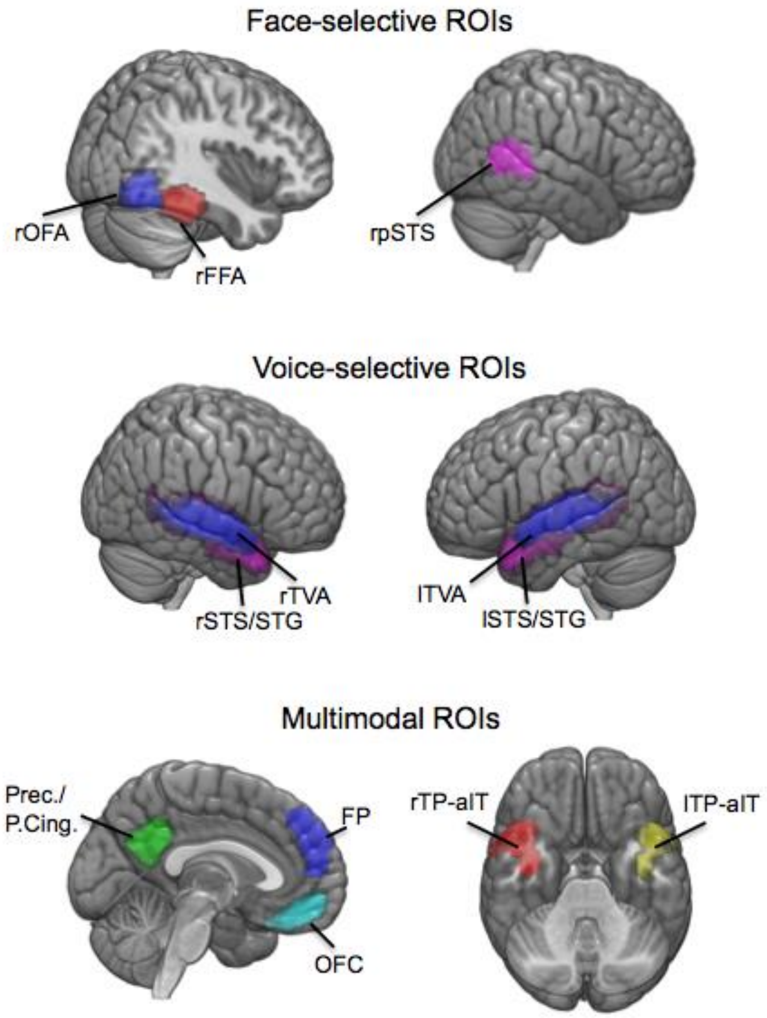


Figure 3.2: Face-selective, voice-selective, and multimodal ROIs. Location of ROIs that resulted from the face, voice, and multimodal localisers in MNI space.

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.

3.2.3 Mean response to faces and voices in ROIs

In order to confirm that each ROI showed the expected responsiveness to faces and voices, we computed the regional mean of the parameter estimates for faces and for voices across participants for each ROI and modality (Figure 3.3). As expected, mean beta values for faces were high and significantly greater than zero in all three face-selective ROIs (all one-sample *t*-tests with *p*<.0001). Mean beta values for voices were also significantly greater than zero in the rFFA (*p*<.0001) and rpSTS (*p*<.0001), but not in the rOFA. The rFFA and the rOFA showed significantly greater responses to faces compared with voices (both paired-samples *t*-tests with *p*<.0001). In contrast, the rpSTS showed significantly greater responses to voices compared with faces (*p*=.0002) despite being defined using our face localiser. This is most likely due to the large overlap between this ROI and the voice-selective rSTS/STG and rTVA ROIs. This finding demonstrates that the rpSTS also showed substantial responses to voices.

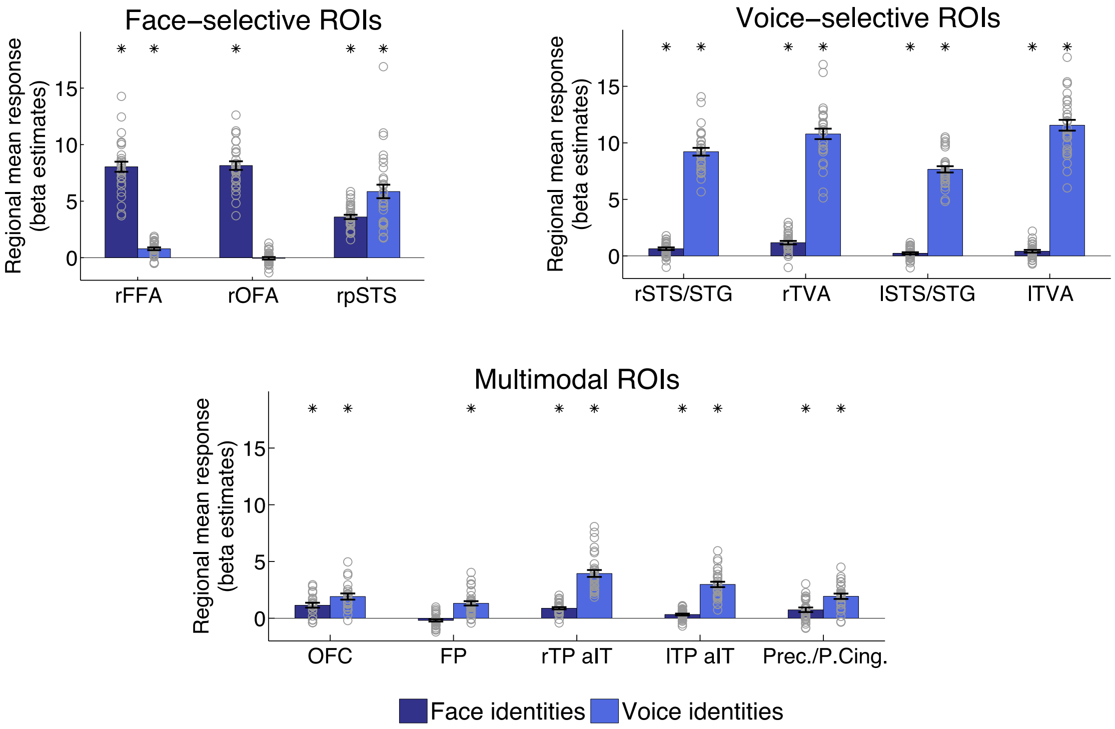


Figure 3.3: Regional mean responses to faces and voices in ROIs. Regional mean responses for all face identities and for all voice identities in face-selective, voice-selective, and multimodal ROIs (mean beta estimates across all voxels of each ROI, and across all runs). Bars show mean responses across 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*≤.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*≤.0011 (FDR corrected for all 12 ROIs).

It could be that the responses to voices in rpSTS were due to the voices being familiar, and not because of being voices *per se*. To determine whether this region responded to voices more generally or just to familiar voices, we investigated the responses in rpSTS to familiar voices, unfamiliar voices, and non-voices during the functional voice localisers. For each participant, we calculated the mean parameter estimates across all voxels of the face-selective rpSTS for each condition of the voice localiser (familiar voices, unfamiliar voices, and auditory scenes) and of the TVA localiser (vocal and non-vocal sounds). For the voice localiser, both the familiar and the unfamiliar voices had significantly higher parameter estimates than the auditory scenes (both *p*<.0001). For the TVA localiser, the rpSTS also showed significantly higher responses to voices than non-voices (*p*<.0001). These results show that the face-selective rpSTS also responds to voices in general and not only familiar voices (for similar results, see Deen et al., 2015), and therefore in the rest of this article we will refer to this rpSTS region as displaying multimodal responses.

Returning to the analysis of the parameter estimates for faces and voices during the main experimental runs, the mean beta values for voices were significantly greater than zero for all four voice selective ROIs (all *p*<.0001). Mean beta values for faces were also significantly greater than zero for all voice-selective ROIs (all *p*≤.0209), but the parameter estimates were significantly lower than for voices (all *p*<.0001).

For the multimodal ROIs mean beta values for faces and for voices were significantly greater than zero in all ROIs (all *p*≤.0009) except the frontal pole for faces. This result demonstrates that, although we still included the frontal pole ROI in the main analyses, one cannot be confident about the multimodal responses of this ROI. Also, we note that in all multimodal ROIs (OFC, FP, rTP-aIT, lTP-aIT, Prec./P.Cing.) mean beta values for voices were significantly higher than mean beta values for faces (all *p*≤.0011). This was observed consistently across all participants.

3.2.4 Analysis A: RSA comparing representational geometries

The first main analysis compared the representational geometry of the 12 famous identities across and within modalities in each ROI. Face and voice RDMs were computed separately for each session using the LDC and RDMs were compared using Pearson correlation (Figures 3.4 & 3.5). We then tested whether these correlations were significantly above zero.

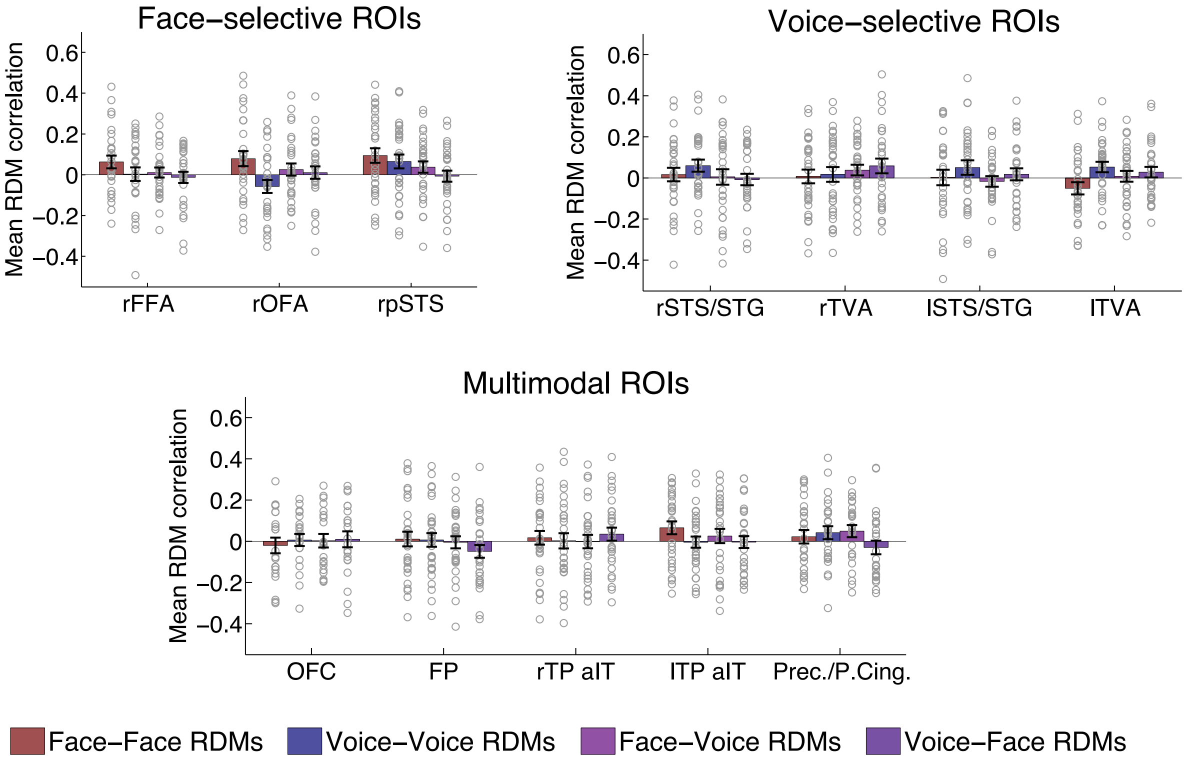


Figure 3.4: Results of RSA comparing representational geometries. Comparisons between the representational distance matrices (RDMs) from two scanning sessions using Pearson’s correlation coefficient. Bars show mean correlations across participants, error bars show standard error, and grey circles show 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 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 $p \leq .0001$ (FDR corrected for all 48 comparisons). Note that in this figure the rpSTS is classed as a face-selective ROI for consistency purposes only, but in fact it demonstrated multimodal properties.

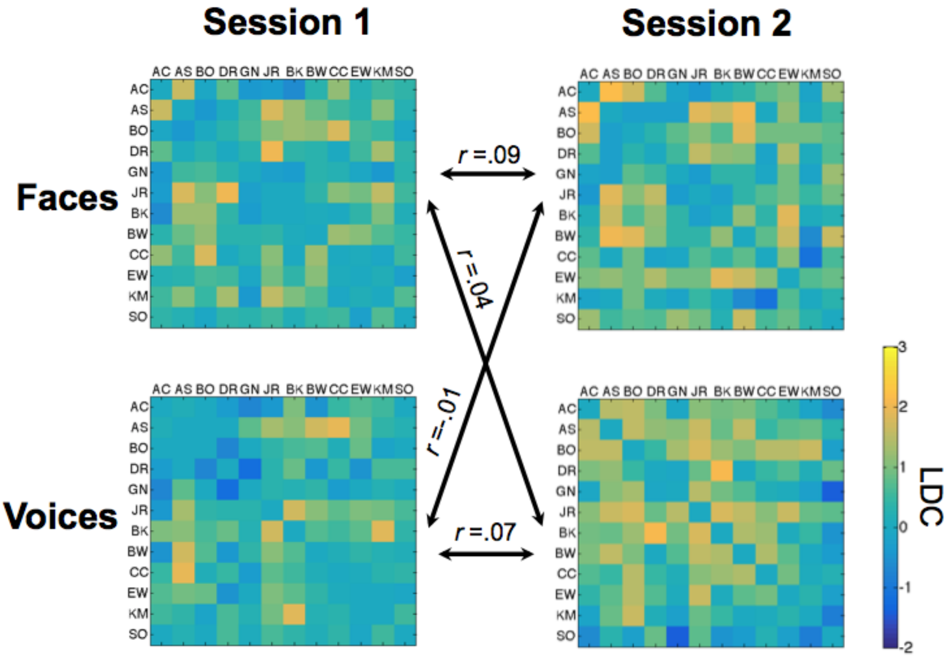


Figure 3.5: Representational distance matrix (RDM) comparisons across scanning sessions 1 and 2 in the rpSTS. Face and voice RDMs for the rpSTS were averaged across all 30 participants for illustration purposes. Each cell shows the discriminability of the brain activity patterns corresponding to a pair 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.

We expected that face and voice RDMs would be correlated in ROIs that represent person identity independently from modality. However, the results showed no significant correlations between face and voice RDMs in face-selective, voice-selective, or multimodal ROIs (Figure 3.4). It is possible that comparing RDMs across different scanning sessions taking place on separate days did not allow the detection of 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. However, we still found no significant correlations between face and voice RDMs. Therefore, using this method we found no evidence of crossmodal person identity representations in our ROIs.

We also expected that there would be correlations between RDMs within the same modality 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.

It is possible that the low correlations between same-modality RDMs across the two scanning sessions could be due to low intra-subject reliability of the brain activity patterns elicited by individual face and voice identities. To investigate this possibility, for each participant, each ROI, and each modality, I averaged the activity patterns (betas) elicited by each of the 12 identities across the three runs in each scanning session. This resulted in an average activity pattern for each identity in session 1, and an average activity pattern for each identity in session 2. I then computed the Pearson correlation between the average activity patterns for each identity across the two sessions. Finally, for each ROI and each participant I averaged these correlations across the 12 identities, and then across all participants. This resulted in two average correlations for each ROI: one showing the intra-subject reliability of activity patterns elicited by face identities in that ROI, and the other showing the intra-subject reliability of activity patterns elicited by voice identities. High correlations (*r* >.70) were found between activity patterns for faces in the face-selective rFFA and rOFA, and for voices in voice-selective bilateral STS/STG and TVAs (Table 3.2). Moderate to high correlations (*r* > .50) were found between activity patterns for faces and voices in the rpSTS, and for voices in the bilateral TP-aIT (Table 3.2). These results suggest that the low correlations between same-modality RDMs across scanning sessions were not due to the low reliability of the activity patterns elicited by individual face and voice identities. In fact, the activity patterns elicited by face identities in face-selective regions, and by voice identities in voice-selective regions, were highly reliable, despite the low reliability of the RDMs in these regions.

Table 3.2: Average correlations between brain activity patterns for faces and for voices across scanning sessions. Correlations in each modality were averaged across the 12 identities and across participants. The rpSTS is classed as a face-selective ROI for consistency purposes only, but in fact it demonstrated multimodal properties. *M* = mean, *SD* = standard deviation.

	Faces		Voices	
	M	*SD*	*M*	*SD*
Face-selective ROIs				
rFFA	.729	.164	.209	.143
rOFA	.805	.170	.164	.155
rpSTS	.668	.179	.846	.133
Voice-selective ROIs				
rSTS/STG	.381	.172	.879	.131
rTVA	.496	.186	.865	.132
lSTS/STG	.245	.130	.861	.135
lTVA	.303	.183	.858	.136
Multimodal ROIs				
OFC	.133	.108	.321	.156
FP	.179	.120	.446	.195
rTP-aIT	.223	.137	.641	.178
rTP-aIT	.128	.116	.518	.209
Prec./P. Cing.	.31	.161	.484	.164

3.2.5 Analysis B: RSA investigating identity discriminability

The second main analysis tested the generalisation of pattern discriminants from one modality to the other. More specifically, we computed crossmodal RDMs and we tested whether linear discriminants computed on pairs of faces could be used to discriminate between pairs of voices, and vice-versa. We also tested whether each ROI could discriminate between pairs of stimuli within the same modality. Mean LDC distances across all cells in crossmodal, face, and voice RDMs were compared against zero.

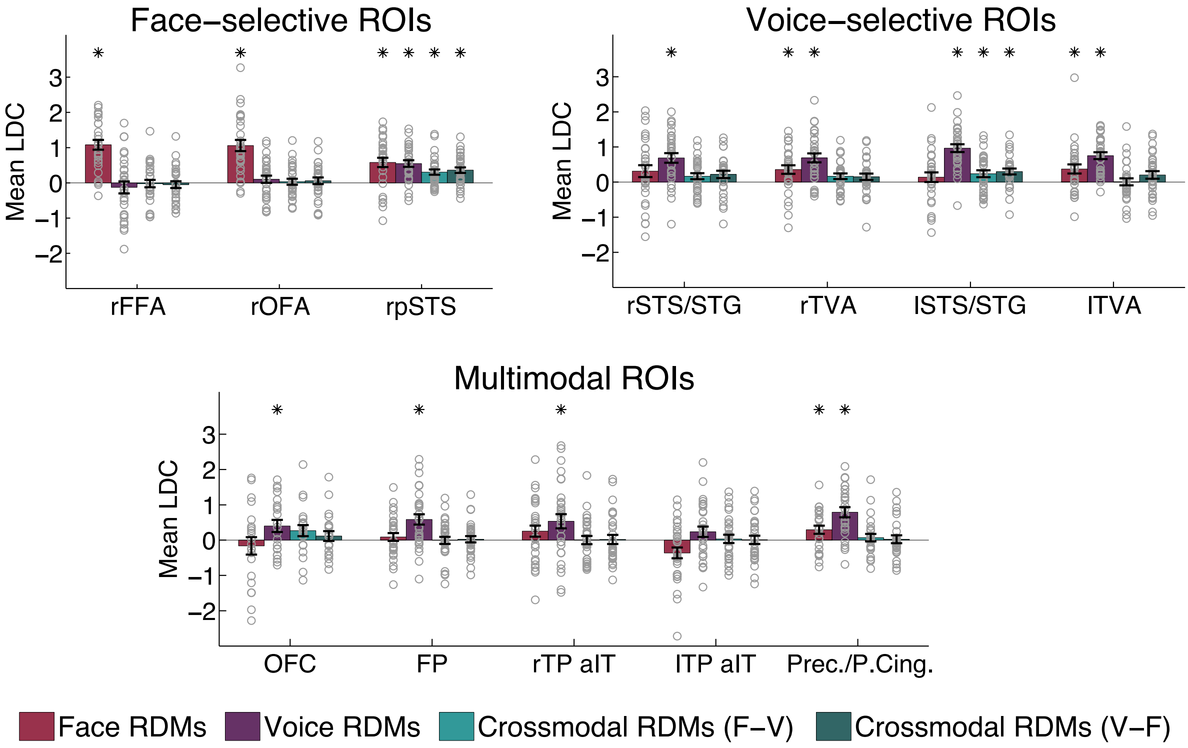


Figure 3.6: Results of RSA investigating identity discriminability. Mean LDC between identities in face RDMs, voice RDMs, and crossmodal RDMs in face-selective, voice-selective, and multimodal ROIs. There are two types of crossmodal RDMs: (a) face discriminant applied to voices (F-V), and (b) voice discriminant 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*≤.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. Note that in this figure the rpSTS is classed as a face-selective ROI for consistency purposes only, but in fact it demonstrated multimodal properties.

Table 3.3: One-sample *t*-test results for mean LDC values in crossmodal RDMs. Stars represent statistical significance at *p*≤.0150 (FDR corrected for all 48 comparisons in face, voice, and crossmodal RDMs). The rpSTS is classed as a face-selective ROI for consistency purposes only, but in fact it demonstrated multimodal properties.

		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
Multimodal 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

Table 3.4: One-sample *t*-test results for mean LDC values in face and voice RDMs. Stars represent statistical significance at *p*≤.0150 (FDR corrected for all 48

comparisons in face, voice, and crossmodal RDMs). The rpSTS is classed as a face-selective ROI for consistency purposes only, but in fact it demonstrated multimodal properties.

		Face RDMs			Voice RDMs		
	df	t	Sig. (1-tailed)	d	t	Sig. (1-tailed)	d
Face-selective ROIs							
rFFA	27	7.764	.0001*	1.47	-0.753	.7711	0.14
rOFA	28	6.707	.0001*	1.25	0.995	.1641	0.18
rpSTS	29	4.378	.0001*	0.80	5.871	.0001*	1.07
Voice-selective ROIs							
rSTS/STG	29	1.850	.0373	0.34	5.025	.0001*	0.92
rTVA	29	2.945	.0031*	0.54	5.447	.0001*	0.99
lSTS/STG	29	1.019	.1583	0.19	8.667	.0001*	1.58
lTVA	29	2.846	.0040*	0.52	7.834	.0001*	1.43
Multimodal ROIs							
OFC	20	-0.662	.7424	0.14	2.337	.0150*	0.51
FP	29	0.799	.2153	0.15	4.007	.0002*	0.73
rTP-aIT	29	1.617	.0583	0.30	2.685	.0059*	0.49
lTP-aIT	29	-2.369	.9877	0.43	1.630	.0570	0.30
Prec./P. Cing.	25	2.538	.0089*	0.50	5.524	.0001*	1.08

We expected that in brain regions with crossmodal person identity representations the mean LDC values for crossmodal RDMs would be significantly greater than zero. The results showed that mean LDC values in these RDMs were significantly greater than zero in the rpSTS, and in the voice-selective lSTS/STG (Figure 3.6; Table 3.3). 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 while the mean LDC values for crossmodal RDMs in the lSTS/STG were 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 crossmodal 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 crossmodal person identity representation and it could be driving the positive result in the lSTS/STG. However, we were not able to test this because we could not localise the lpSTS in our participants using our face localiser.

We also expected 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 multimodal Prec./P. Cing. (Figure 3.6; Table 3.4). 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 (originally defined as face-selective), and in the multimodal OFC, FP, rTP-aIT and Prec./P. Cing. (Figure 3.6; Table 3.4).

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 (Figure 3.6 & Tables 3.2, 3.3) and for each participant we compared the mean LDC values for different-gender identity pairs (calculated across 36 pairs) 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 ≤.0010), and for voice RDMs in the bilateral TVAs and STS/STG (all p≤.0005), suggesting that gender 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*≤.0239), suggesting that identity discrimination in these regions is not solely driven by differences in gender.

3.2.6 Exploratory whole-brain searchlight analyses

We conducted additional exploratory searchlight analyses across the whole brain to determine whether there were brain regions with crossmodal person identity representations that are not included in the 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 generalization 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 ≤ .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 ≤ .005 (k=592, *p*=.001, *t*=4.05, x=48, y=-49, z=11) that overlapped with the rpSTS ROI.

3.3 Discussion

This study showed that there is a crossmodal person identity representation in a multimodal region of the rpSTS, demonstrating that this region is able to discriminate familiar identities based on crossmodal information in faces and voices. More specifically, the rpSTS could discriminate pattern estimates for pairs of face identities based on linear discriminants computed from pattern estimates for pairs of voice identities, and vice-versa. A crucial and novel aspect of this study is the finding 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 using different tokens of the face and voice to obtain and test pattern discriminants, it was demonstrated that the face- and voice-elicited person identity representations in the rpSTS are stimulus-invariant and modality-invariant. Stimulus-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, there was no evidence of matching representational geometries for faces and voices, across or within modalities, and possible reasons for this will be discussed below.

A crossmodal and invariant person identity representation in the rpSTS

The finding of a crossmodal person identity representation in a multimodal region of the rpSTS supports the MP model, which proposes that face and voice information is integrated in multimodal brain regions (e.g. Hölig et al., 2017; Joassin et al., 2011). In contrast, no support was found for the prediction from the CFVP Model (e.g. von Kriegstein & Giraud, 2006; von Kriegstein et al., 2005) that there would be a crossmodal identity representation in face-selective regions of the fusiform gyrus.

The finding that the face-selective rpSTS also shows voice-selectivity is in agreement with multiple studies showing overlap between face-selective and voice-selective regions in the rpSTS (Anzellotti & Caramazza, 2017; Davies-Thompson et al., 2018; Deen, Koldewyn, Kanwisher, & Saxe, 2015; Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009; Watson, Latinus, Charest, Crabbe, & Belin, 2014; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003), suggesting that the pSTS is not only multimodal but also shows a preference for people-related stimuli regardless of modality (Watson, Latinus, Charest, et al., 2014). The pSTS has previously been associated with person identity processing by a study showing that this region responded more to voices that were primed by the face of a different identity, compared with voices that were primed by the faces of the same identity (Hölig et al., 2017). Moreover, the rpSTS has been associated with crossmodal representations of emotion from faces and voices (Watson, Latinus, Noguchi, et al., 2014). Finally, the pSTS has been implicated in audiovisual integration by studies comparing responses to audiovisual conditions with responses to auditory and visual conditions presented in isolation, for voices and speaking mouths (Calvert et al., 2000), sound-producing animals and tools (Beauchamp, Lee, Argall, & Martin, 2004), emotional faces and voices (Davies-Thompson et al., 2018; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007; Robins, Hunyadi, & Schultz, 2009), and neutral faces and voices (Watson, Latinus, Charest, et al., 2014).

The main finding of a crossmodal representation of person identity in the rpSTS is in agreement with a study showing crossmodal classification of pattern estimates for familiar faces and voices in this region (Anzellotti & Caramazza, 2017). Similar to the present study, the rpSTS region in Anzellotti & Caramazza (2017) showed both face- and voice-selectivity. In contrast to Anzellotti & Caramazza (2017), who tested the discrimination of just two identities, in the present work discrimination was tested across a larger set of 12 identities. Moreover, the present work additionally showed that the rpSTS contains representations of face and voice identity that are invariant to different tokens of the same face and voice. Although Anzellotti & Caramazza (2017) presented two tokens of the face and voice of each identity, they did not test whether the two identities could be discriminated within each modality by training and testing their classifiers using different tokens. 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) (Anzellotti & Caramazza, 2014; Burton, 2013). Finally, the present work was based on the presentation of multiple naturalistically varying face videos for each identity, in sharp contrast to Anzellotti & Caramazza (2017), who presented grey scale images of faces with low within-person variability, and thus shows that identity representations in the rpSTS are robust to substantial within-person variability. While behavioural studies have shown the importance of within-person variability for recognition (Jenkins et al., 2011; Burton, 2013; Burton et al., 2016), this is rarely taken into account in neuroimaging experiments, which typically use highly similar or artificial stimuli for the same person.

It should be noted that while there was also evidence of crossmodal person identity discrimination in the voice-selective lSTS/STG, this region could not discriminate between individual faces, and therefore this finding should be interpreted with caution. It has previously been shown that the left pSTS shows crossmodal representations of emotion from faces and voices (Peelen et al., 2010) and is involved in audiovisual integration (Beauchamp, Lee, et al., 2004; Calvert et al., 2000, 2001; Robins et al., 2009). Thus, it is possible that a multimodal pSTS region within the voice-selective lSTS/STG ROI shows person identity representations. However, the lpSTS was not included as a ROI in the present study because it was not consistently activated across participants in the face localiser.

Invariant representations of face identity and voice identity

In addition to discriminating between person identities across modalities, this study also demonstrated that the pSTS could discriminate between individual face and voice identities within each modality by generalising across different tokens of the face and voice of each identity. For faces, these findings complement previous work showing that the pSTS can discriminate between face identities by generalising across different viewpoints of the face (Anzellotti & Caramazza, 2017; Visconti Di Oleggio Castello et al., 2017). Evidence for representations of face identity in the pSTS also come from a study showing adaptation to repeated presentations of the same identity (Winston et al., 2004). For voices, previous studies have shown that regions in the STS/STG can discriminate between different voice identities by generalising across different recordings of the voice (Bonte et al., 2014; Formisano et al., 2008). Moreover, adaptation to morphed voices with the same perceived identity has been shown in the bilateral pSTS (Andics et al., 2010). In contrast with these studies, which largely used stimuli with low within-person variability, the present study shows that representations in these regions generalise across highly variable, naturalistic face videos.

The face-selective rFFA and rOFA were also 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 to novel viewpoints of the face. Moreover, studies have shown evidence of representations of face identity in the FFA that generalise across different face images (Axelrod & Yovel, 2015), different emotional expressions (Nestor et al., 2011), and different viewpoints of the face (Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017), although these studies did not test whether these representations generalised to novel tokens of the face. Adaptation studies have also shown evidence of invariant representations of face identity in the FFA for same-identity faces with different emotional expressions (Winston et al., 2004), for physically-different face morphs that were perceived as having the same identity (Rotshtein et al., 2005), and for same-identity images presented from different viewpoints (Ewbank & Andrews, 2008; Mur et al., 2010; Verosky et al., 2013). In contrast with these previous studies, the current work shows that face identity representations in the rFFA and rOFA are robust to naturalistic changes in the appearance of the face of the same person.

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 voice identity that generalise to novel utterances of different vowels in the lateral Heschl’s gyrus/sulcus and in the right STS. Bonte et al. (2014) also showed some evidence of voice identity representations in the TVAs, but did not test whether these representations generalise to novel utterances from the same voice. The present study extends previous findings by showing that generalisation to different recordings of the same voice is possible in voice-selective regions even when using short sentences with variable speech content that were recorded in different settings. Finally, the finding of voice identity representations in voice-selective regions is also in agreement with adaptation studies showing adaptation to same-identity voices in the mid and posterior STS (Andics et al., 2010) and anterior STS (Belin & Zatorre, 2003).

The present study also showed invariant discrimination of face identity and of 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.

Representational geometries

There was no evidence of matching representational geometries across faces and voices in rpSTS despite the finding of crossmodal generalisation of the pattern discriminants. It is possible that all identities were equally distinct from each other within each modality (i.e. the nature of person identity code in these regions does not result in variable representational distances between identities). In addition, the rpSTS showed both modality-specific and crossmodal representations, and it is possible that the former had stronger influence on the representational geometry. Beauchamp, Argall, Bodurka, Duyn, & Martin (2004) showed that the pSTS contains intermixed visual, auditory, and multisensory patches, and future studies could use higher-resolution neuroimaging methods to probe person identity representations in this region.

There was also no evidence of stable representational geometries across scanning sessions for face identities or voice identities in any of the ROIs. Again, it could be that identities were equally distinct across from each other within each modality, or it could be that experimental conditions would need to be improved to obtain more reliable representational geometries.

Anterior temporal lobe and searchlight results

There was no 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 (A. W. Ellis et al., 1989; Gainotti, 2011). The fact that the TP-aIT ROIs responded more to voices that to faces suggests that the multimodal region localizer used in the present study was not optimal for detecting multimodal responses in the anterior temporal lobe. Moreover, the sequences used 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.

It is possible that crossmodal representations exist outside face- and/or voice-selective regions, and the 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.

Limitations and future directions

A possible limitation of the present study is that the set of identities whose faces and voices were presented in the study were highly variable in terms of facial and vocal appearance. Specifically, the identities differed in respect to their gender, age, nationality, accent, and race. Moreover, these famous identities were selected (based on pilot studies previously conducted in the lab) because they were highly recognisable based on both their face and voice, and thus are likely to display particularly distinctive facial and vocal features. Although the different face videos and voice recordings for each identity were also highly variable within the same identity, it is possible that the between-person variability was higher that the within-person variability. Thus, it can be argued that discrimination between identities, both within modality and across modalities, was facilitated by their distinct appearances. Future studies should attempt to replicate findings of the current study using a more homogenous set of familiar identities.

A further potential limitation concerns the use of famous-familiar identities. Although participants reported being highly familiar with the famous people from both their faces and voices, it is unlikely that participants ever interacted with these people personally. Instead, familiarity will have been obtained from a third-person perspective, and most likely through the mass media. In contrast, familiarity with personal acquaintances, such as friends and relatives, is usually obtained through direct interactions. Studies contrasting brain activation in response to personally-familiar faces with activation in response to famous-familiar faces showed stronger activation in a number of regions, including the bilateral pSTS, posterior cingulate/precuneus, and fusiform gyrus (Gobbini, Leibenluft, Santiago, & Haxby, 2004; Sugiura, Mano, Sasaki, & Sadato, 2011). Therefore, it is possible that representations of individual face identities, voice identities, and person identities may differ depending on the type of familiarity with a person, and future studies should investigate potential differences between the representations of faces and voices of famous-familiar and personally-familiar identities.

A third potential limitation relates to the use of dynamic face stimuli and the interpretation of the finding of a person identity representation in the pSTS, given its association with the processing of dynamic faces (Fox et al., 2009; Pitcher et al., 2011) and audiovisual speech processing (Calvert et al., 2000; Kreifelts et al., 2007; Robins et al., 2009; Watson, Latinus, Charest, et al., 2014). It is possible that the observed crossmodal representation in this region is specific to dynamic faces, and dynamic speech-related information that is shared between the face and the voice (Yovel & O’Toole, 2016). Behavioural studies have shown that it is possible to successfully match faces and voices of unfamiliar people when presented with speech samples and silent videos of speaking faces, even when the faces are speaking a different sentence to the speech sample (Kamachi, Hill, Lander, &

Vatikiotis-Bateson, 2003; Lachs, Lorin, Pisoni, 2004; Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007; Smith, Dunn, Baguley, & Stacey, 2016b). This raises that possibility that dynamic faces and voices of the same identity share information even when the person is unfamiliar, and that this information may be cross-decodable in the brain. However, in contrast to these studies, the dynamic face stimuli used here did not feature silent speech, and thus should not necessarily engage auditory representations relating to speech (Calvert et al., 1997). Moreover, person identity representations in the rpSTS have been shown previously in a study using static face stimuli (Anzellotti & Caramazza, 2017). However, future work should probe the existence crossmodal face-voice representations in the rpSTS using unfamiliar stimuli.

Conclusion

To conclude, this study showed a crossmodal person identity representation that generalises across different, naturalistically varying face videos and voice recordings of the same person in a multimodal region of the rpSTS. This supports the MP Model for face and voice integration. This study also showed 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). The next chapter will focus on the type of information that is represented in these different regions.

Chapter 4

The informational content of face and voice representations

Chapter 3 presented an fMRI study that revealed a number of face-selective and voice-selective regions, and multimodal brain regions that selectively respond to both faces and voices. While previous work has demonstrated that some of these regions consistently respond to faces and voices, the type of information from the face or the voice that is represented in each region is unclear. Therefore, this chapter will address the second aim of this thesis, which was to determine where in the brain the different types of information conveyed by faces and voices is processed. It describes a study that compared representations of faces and voices in face-selective, voice-selective, and multimodal brain regions with models of different types of information that can be extracted from faces and voices.

Previous studies have shown evidence that face-selective and voice-selective regions, or regions known to overlap with the location of these regions, process information related to physical properties of faces and voices (Formisano et al., 2008; Weibert et al., 2018), gender (Kaul et al., 2011; Weston et al., 2015), and social information (Bestelmeyer et al., 2012; Todorov & Engell, 2008). However, a review of findings from these studies, presented in Chapter 1, suggested that many of the findings relating to representations of gender and social information in face-selective and voice-selective regions may be explained by variations in physical face and voice characteristics. For gender, differences between male and female faces or voices are related to differences in facial features (Bruce et al., 1993) and vocal features (Titze, 1989), respectively, confounding the interpretation of studies showing discrimination of male and female faces in face-selective regions (Contreras et al., 2013; Kaul et al., 2011), and stronger responses to female voices in voice-responsive regions (Lattner et al., 2005; Sokhi et al., 2005; Weston et al., 2015). Moreover, findings that highly masculine and highly feminine faces and voices increase brain activity in face-selective and voice-selective regions are likely to be confounded by the high distinctiveness of these faces and voices in terms of physical features (Charest et al., 2013; Freeman et al., 2010; Mattavelli et al., 2012). For social information, a similar confound exists for studies showing increased brain activity in face-selective regions in response to faces and voices with very positive or very negative valence (Mattavelli et al., 2012; Said et al., 2011). However, the majority of previous studies focused on one type of face and voice information, and were therefore not able to directly compare different types of information.

Few studies have investigated the type of information used by face-selective and voice-selective brain regions to distinguish between individual faces and voices. Weibert et al. (2018) showed that the similarity between the representations of different face conditions (featuring multiple face images that could vary based on identity, viewpoint, and/or emotional expression) in the FFA, OFA, and STS was predicted by the similarity between the low-level properties of the images, as measured by the GIST descriptor model. The use of low-level visual information to distinguish between different faces in the FFA was also demonstrated in a study showing that response patterns to artificially-generated faces in this region were explained by computational models of low-level visual properties, including simulations of visual processing in areas V1 and V4 (Carlin & Kriegeskorte, 2017). Moreover, studies have shown evidence of discrimination of male and female faces in multiple face-selective regions, including the FFA, OFA, and pSTS (Contreras et al., 2013; Kaul et al., 2011), suggesting that these regions are sensitive to visual information that distinguishes faces based on gender. For voices, Formisano et al. (2008) showed that regions of the STS/STG, which are known to overlap with the voice-selective TVAs (Pernet et al., 2015), use information about vocal pitch to distinguish between individual voice identities, and information about formant frequencies to distinguish between different vocalisations. However, the information used by the TVAs to distinguish between different identities has not been explicitly investigated. Moreover, to the best of my knowledge, the ability of face-selective and voice-selective regions to distinguish between individual faces and voices based on social information has not been investigated. Finally, virtually nothing is known regarding the informational content of face and voice representations in multimodal regions that respond to both faces and voices. In Chapter 1, it was argued that multimodal regions may represent information that is extracted from both faces and voices, such as information relating to social traits (McAleer et al., 2014; Oosterhof & Todorov, 2008).

The present study used RSA to compare representations of individual face and voice identities in face-selective, voice-selective, and multimodal regions (computed in Chapter 3), and in the amygdala with multiple models of physical face/voice properties, gender, and social information, with the aim to determine the informational content of representations in these regions. Although the amygdala was not included in the analysis in Chapter 3, it was included in the present study to test its proposed involvement in the processing of social information in faces (Bzdok et al., 2011; Mende-Siedlecki et al., 2013; Santos et al., 2016). Models of perceived physical properties were computed based on ratings on visual/auditory pairwise similarity tasks, and models of objective physical properties were computed based on measures of stimulus similarity obtained using the OpenFace and Gabor-Jet programs for faces, and using f0 and AVTL for voices. A model of gender predicted that response patterns would be more similar between same-gender faces/voices than between different-gender faces/voices. Models of social information were computed from ratings of faces and voices on social traits, namely trustworthiness, dominance, attractiveness, and positive-negative valence. It was predicted that face-selective and voice-selective regions would primarily process information relating to the perceived and objective visual/auditory similarity of the faces/voices, as well as gender. In addition, it was predicted that multimodal regions would primarily process social information from faces and voices. Finally, it was expected that the amygdala would process social information from faces as well as information on the visual similarity of faces.

An additional exploratory analysis was performed that compared the representations of faces and voices between different face-selective regions, voice-selective regions, multimodal brain regions, and the amygdala with the aim to identify brain regions that share information with each other. Although this analysis does not reveal what information is being shared across regions, it provides insights into which regions share information, and how the different regions are organised in terms of their representations of faces and voices. Moreover, it enables the identification of brain regions that share types of information that were not included in one of models.

4.1 Methods

The present study used the fMRI data that was reported in detail in Chapter 3. For the analysis in this chapter, we used the face RDMs and voice RDMs that showed the pairwise discriminability of the response patterns for all pairs of face identities and all pairs of voice identities in face-selective, voice-selective, and multimodal brain regions.

The same participants that took part in the fMRI study reported in Chapter 3 also completed a behavioural session. In this study, participants rated the same faces and voices (i.e. the stimuli that they were presented with in the scanner) on trustworthiness, dominance, attractiveness, positive-negative valence, and on perceived pairwise visual and auditory similarity. We then created model RDMs based on the distance between stimuli for each of these attributes, and used correlation to compare the face and voice RDMs in different brain regions to the model RDMs. In addition, face and voice brain RDMs were compared with RDMs of visual and acoustic features and gender. Given that details of the methods for the fMRI sessions are presented in detail in Chapter 3, below I focus on the behavioural session only.

4.1.1 Stimuli for the behavioural session

The same stimuli that were presented in the fMRI task (described in detail in Chapter 3) were presented in the behavioural session. However, although the duration of the original face and voice stimuli was 3000ms, only the first 1500ms of each stimulus were presented in this study. The length of the original stimuli had been determined to accommodate the fMRI study, and they were cropped for the behavioural session in order to shorten the length of the tasks.

4.1.2 Procedure for the behavioural session

Trustworthiness, Dominance, Attractiveness, and Valence rating tasks

There were 8 blocks in total (4 judgements x 2 modalities). Faces and voices were presented in separate blocks, and presented using ABBA counterbalancing. The modality of the first block was counterbalanced across participants. The presentation order of the four judgements was also counterbalanced across participants. All tasks and stimuli were presented using the Psychophysics Toolbox (version 3; Brainard, 1997; Pelli, 1997) running in Matlab (version R2013b, The MathWorks, Inc.). Face stimuli were presented in the centre of the screen. Participants listened to the voice stimuli through headphones (Sennheiser HD 202).

In each trial of each task, a face or voice identity was represented by three tokens of their face or voice, presented successively with no gap in between them (Figure 4.1). Each identity was presented in two trials; one trial presented three face/voice tokens randomly selected from the six available, and the other trial presented the remaining three tokens. This resulted in 24 trials in each face/voice block (12 identities x 2 presentations). Tokens within each trial were presented in a random order, and trial order was also randomised. Each video and audio clip was presented for 1500ms and there was a 1000ms inter-trial-interval following the response.

In each trial, participants were asked to rate how trustworthy/dominant/attractive the face/voice is, or how they feel about the face/voice (for valence), basing their judgement on three tokens of the face/voice that were presented successively (Figure 4.1). The rating scale ranged from 1 (very untrustworthy/non-dominant/unattractive/negative) to 7 (very trustworthy/dominant/attractive/positive) and participants responded using the corresponding keys on the keyboard. Trustworthiness was defined as ‘able to be relied on as honest and truthful’. Dominance was defined as ‘having power and influence over other people’. No definition was provided for valence or attractiveness. Participants were advised that there was no time limit to their responses and that they should go with their first instinct. The duration of each block was approximately 3 minutes.

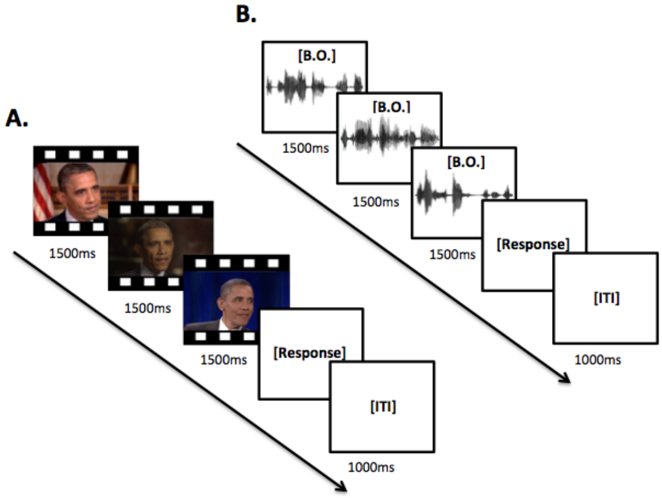


Figure 4.1: Example trial sequence and stimuli for Trustworthiness, Dominance, Attractiveness, and Valence rating tasks. A: Example trial in face block. B: Example trial in voice block.

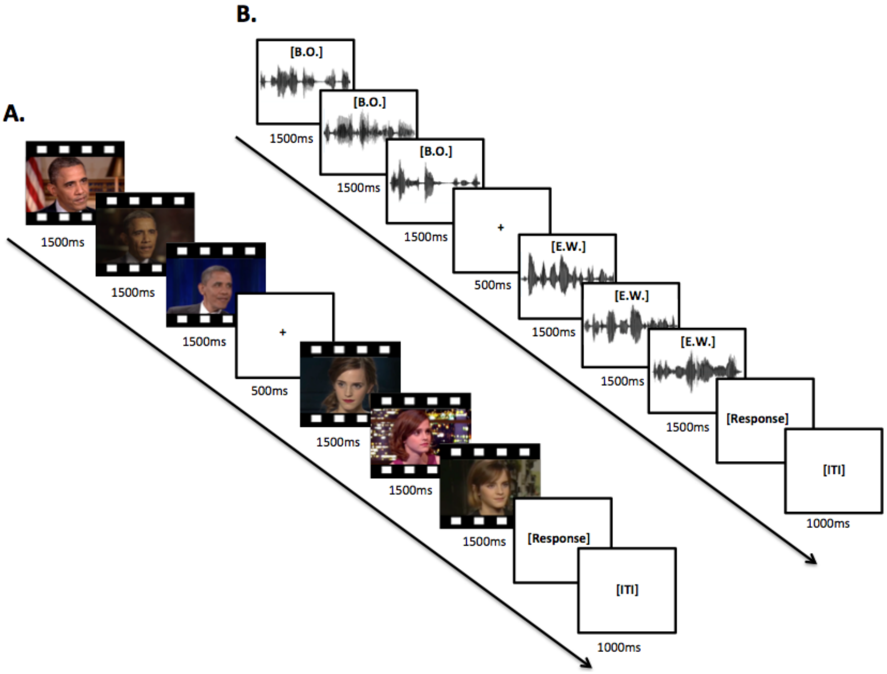


Figure 4.2: Example trial sequence and stimuli for pairwise auditory and visual similarity ratings tasks. A: Example trial in face block. B: Example trial in voice block.

Visual and Auditory Similarity rating tasks

In these tasks, participants rated the visual or auditory similarity of pairs of face identities or pairs of voice identities, respectively. For each task, each of the 12 identities was paired with the other 11 identities to create 66 identity pairs. Each identity pair contained three tokens for each identity that were randomly selected from the six available tokens for that identity. Each identity pair was presented in two trials to counterbalance the presentation order of the two identities within the trial. Because the tokens for each identity were selected randomly, the two trials did not necessarily present different tokens. There were 132 trials in each task (66 identity pairs x 2 presentations).

The presentation order of the visual similarity task and the auditory similarity task was counterbalanced across participants. In addition, the presentation order of the pairwise similarity tasks in relation to the other ratings tasks (Trustworthiness, Dominance, Attractiveness, Valence) was counterbalanced across participants so that the similarity tasks were either completed before or after the other tasks.

In the visual similarity task, participants were instructed to rate the similarity between the visual appearance of the two face identities in each pair, focusing on the facial features. In the auditory similarity task participants were instructed to rate how similar the two voice identities sounded in terms of the characteristics of their voices. In both tasks, participants were told to ignore similarities between identities that were related to non-perceptual, biographical or semantic information (e.g. identities that were both actors). Furthermore, to encourage participants to base their judgements on perceptual information, participants were advised to consider to what extent two identities could potentially be related to each other, i.e. be part of the same family, based on how they looked (visually similarity task) or sounded (auditory similarity task).

In each trial, participants were first presented with the three tokens of the face/voice of identity A. Following a 500ms fixation screen, they were presented with the three tokens of the face/voice of identity B. Stimuli were presented using Psychtoolbox in Matlab 2013b. Tokens for each identity were presented successively with no gap in between (Figure 4.2). Participants were then asked to rate how similar the two faces/voices looked/sounded on a scale from 1 (very dissimilar) to 7 (very similar) by pressing the corresponding key on the keyboard. Participants were advised that there was no time limit to their responses and that they should go with their first instinct. Each video and audio clip was presented for 1500ms and there was a 1000ms inter-trial-interval (ITI) following the response. The presentation order of the trials within each task was randomised. The duration of each task was approximately 30 minutes.

4.1.3 Representational geometries: computing RDMs

Brain RDMs

We used the RDMs for faces and voices, defined as described in Chapter 3, for all the ROIs previously defined, namely (1) face-selective regions (right fusiform face area (FFA), occipital face area (OFA), and posterior superior temporal sulcus (pSTS) — the latter was originally defined as face-selective, but shown to display multimodal properties), (2) voice-selective regions (bilateral superior temporal sulcus/gyrus (STS/STG) and temporal voice areas (TVAs), and (3) multimodal regions (orbitofrontal cortex (OFC), frontal pole (FP), precuneus/posterior cingulate (Prec/P.Cing.), and bilateral temporal pole and anterior inferior temporal cortex (TP-aIT)). In addition to these RDMs, we also computed RDMs for the right and left amygdala (Amyg). The amygdala was not consistently activated across participants in the face localiser, and therefore anatomical masks of this region were obtained from the Harvard-Oxford brain atlas in FSL. These masks were thresholded to include voxels that were present in at least 20% of participants. The masks were then transformed to each participant's native space and used as ROIs. In contrast to all other ROIs, which were created by masking the results of the functional localisers, the anatomical amygdala masks themselves were used as amygdala ROIs. These ROIs were then used to compute GLMs and RSA using the same procedures reported in Chapter 3.

For each participant, face and voice brain RDMs for each ROI were averaged across the two scanning sessions, resulting in one 12x12 face RDM and one 12x12 voice

RDM per ROI. RDMs showed the discriminability (computed using the LDC) of the response patterns for all 12 face identities and all 12 voice identities.

Candidate model RDMs for behavioural ratings

RDMs for ratings of the faces and voices on trustworthiness, dominance, attractiveness, and positive-negative valence were computed for each participant. For each modality and each judgement, the Euclidean distance between the ratings of each possible pair of identities was calculated (ratings were averaged across the two trials in which the same identity was presented), separately for faces and voices. Thus, for each judgement, we obtained a 12x12 matrix of Euclidean distances between ratings for 66 identity pairs.

The RDMs based on judgments of perceptual similarity were computed in a different manner given that the judgments themselves were already of the similarity of each identity pair. The ratings of the face and voice pairs were averaged across the two trials in which each identity was presented, and were reverse-coded to match the LDC and Euclidean distance measures, where a higher value indicated higher dissimilarity. Thus, after reverse scoring, a similarity value of one reflected maximum similarity and a value of seven reflected maximum dissimilarity. The resulting values were arranged into 12x12 RDMs for face identities and for voices identities.

Candidate model RDM for gender

A 12x12 RDM for identity gender was created by assigning a value of 0 to same gender identity pairs, and a value of 1 to different-gender identity pairs.

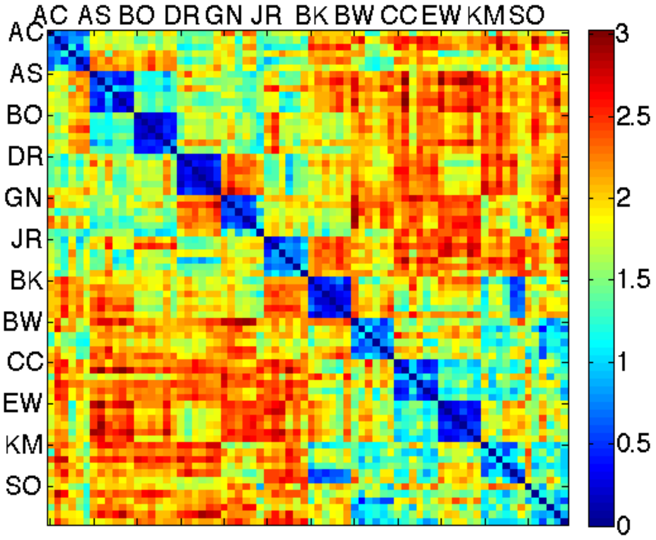
Candidate model RDMs for visual features

In addition to the perceived similarity between stimuli, we also used objective measures of the similarity between the visual appearances of the faces of the 12 identities based on models or descriptors of the visual features of the faces. Here, we used two methods to do this, one based on neural networks using OpenFace (Amos, Ludwiczuk, & Satyanarayanan, 2016) and one using the Gabor-Jet model (Biederman & Kalocsai, 1997; Margalit, Biederman, Herald, Yue, & von der Malsburg, 2016; Yue et al., 2012).

The first model of the similarity between visual features of the faces was computed using the OpenFace face recognition library (Amos et al., 2016; http://cmusatyalab.github.io/openface/). Briefly, this program detects a face, does an affine transformation so that the eyes and mouth appear in approximately the same location, and then creates a descriptor or low-dimensional representation of the face. The program then uses a deep neural network to map each face to its low dimensional representation (128 measurements), and computes the distance between the low dimensional representations of each two faces. For use as input to OpenFace, we extracted one still frame from each face video used in the experiment. Thus, we obtained six different images of the face of each identity, taken from the six different videos in which the identity was presented, resulting in 72 images in total. The low-dimensional representations for each face that were obtained from OpenFace were used to compute Euclidean distances between each pair of face images. Figure 4.3A shows the 72x72 RDM with the OpenFace distances between each pair of images. Zero means it is the same image and values between 0 and 1 likely indicate that two pictures likely belong to the same person. Larger values than 1 indicate that the two pictures belong to different people. We found that OpenFace performed quite well at grouping images of the same person (*M* = 0.652) compared to images of different people (*M* = 1.914) in our stimuli. To obtain a 12x12 RDM for the 12 identities, which would be comparable to the brain RDMs, for each identity pair we averaged the distances across all possible stimulus pairings for that pair. Thus, we obtained an RDM showing the similarity between the visual appearances of the faces of the 12 identities, averaged across all stimulus pairs featuring those identities (Figure 4.5A).

A second measure of similarity between face images was obtained using the Gabor-Jet model (Biederman & Kalocsai, 1997; Margalit et al., 2016; Yue et al., 2012). This model was designed to simulate response properties of cells in area V1. Moreover, Gabor-Jet similarity has been found to correlate with psychophysical measures of facial similarity (Yue et al., 2012). The same images that were used as input to the first model were also used in this analysis. First, OpenFace 2.0 (Baltrusaitis, Zadeh, Lim, & Morency, 2018) was used to detect faces in each image, and the images were then greyscaled. The Matlab script provided in www.geon.usc.edu/GWTgrid_simple.m was then used to create a 100 x 80 Gabor descriptor for each face. After transforming these matrices into vectors, we computed the Euclidean distance between the vectors from each pair of faces. Figure 4.3B shows the matrix with the Gabor-Jet distances between each pair of images. In contrast to the OpenFace distance model, there was a very small difference between the mean Euclidean distances for different tokens of the same people (*M*= 411) compared with different tokens of different people (*M*= 438). Finally, to obtain a 12x12 identity RDM for comparison with the brain RDMs, we followed the same procedure as with the OpenFace model (Figure 4.5A). From Figure 4.3B it is apparent that identity 'AC' shows larger dissimilarity between the different stimuli of AC compared with other identities, and Figure 4.5A shows that this identity is dissimilar to all other identities. It is likely that these effects are due to AC wearing glasses in the images, and perhaps this analysis introduced some errors in locating the eyes in each picture of AC.

A. OpenFace



B. Gabor-Jet

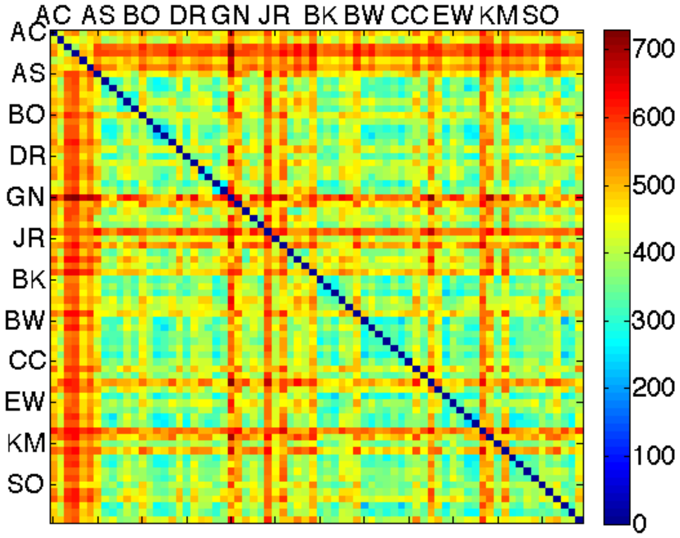


Figure 4.3: RDMs showing the Euclidean distances between face images of the 12 identities based on the OpenFace program (A) and the Gabor-Jet model (B). Matrices are symmetric around a diagonal of zeros, and each cell shows the distance between the stimuli in the corresponding row and column.

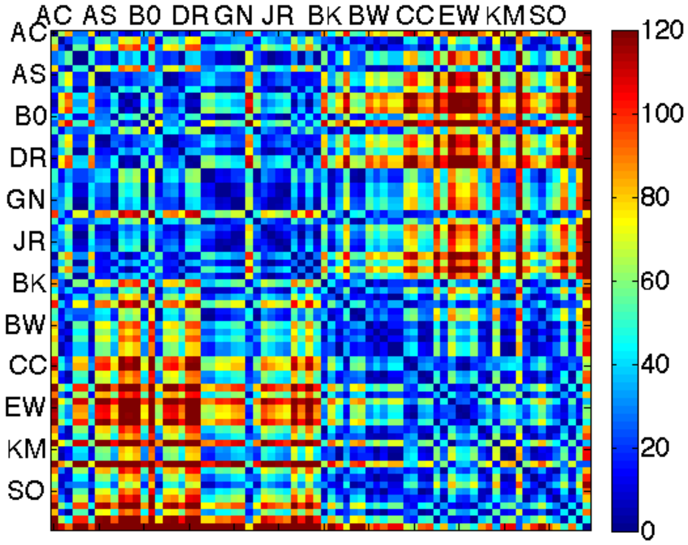
Candidate model RDMs for acoustic features

For voices, we also included measures of objective similarity between identities, this time based on acoustic properties of the stimuli. Based on their relevance for voice identity processing (Baumann & Belin, 2009; Lavner et al., 2001), RDMs were computed based on the similarity of the mean fundamental frequency (f0; perceived as voice pitch), and apparent vocal tract length (AVTL) of the voices of the 12 identities. F0 is related to the rate of vocal fold vibration in the larynx, where sound is produced, whereas AVTL is related to the filtering of sound through the vocal tract, i.e. the pharynx, mouth, and nose (Fitch, 2000).

The mean f0 of each voice stimulus was extracted using Praat (version 5.3.80; Boersma and Weenink, 2014; www.praat.org). To calculate AVTL, the first four formant frequencies (f1, f2, f3, f4) for each voice stimulus were extracted using Praat, and the AVTL for each stimulus was then computed using the formula described in Cartei, Cowles, & Reby (2012). Briefly, the average formant spacing (ΔF) for each voice, i.e. the distance between all neighboring formants, was calculated as the slope of a linear regression line obtained by plotting the values for the four formant frequencies against increments of formant spacing that are predicted by a model of the vocal tract, and which correspond to 0.5, 1.5, 2.5, and 3.5 for formants f1, f2, f3, and f4, respectively. The known approximate speed of sound in the vocal tract, which is 35000 m/s, was then divided by the resulting ΔF value, after multiplying the value by 2. The resulting AVTL value is expressed in centimetres. Figure 4.4 shows the matrices with the f0 and AVTL distances between each pair of sounds. For both f0 and AVTL, the mean Euclidean distances for different tokens of the same people (f0: *M*= 30.2, AVTL: *M*= 0.52) were only slightly lower compared with different tokens of different people (f0: *M*= 53.03, AVTL: *M*= 1.26).

To create identity RDMs, the mean f0 values and AVTL values for the six different voice stimuli for each speaker were averaged to create one mean f0 value and one AVTL value for each of the 12 identities. Finally, we calculated the Euclidean distance between the mean f0 values, and between the AVTL values, for all possible pairs of identities, resulting in a 12x12 f0 RDM and a 12x12 AVTL RDM (Figure 4.5B).

A. f0



A. AVTL

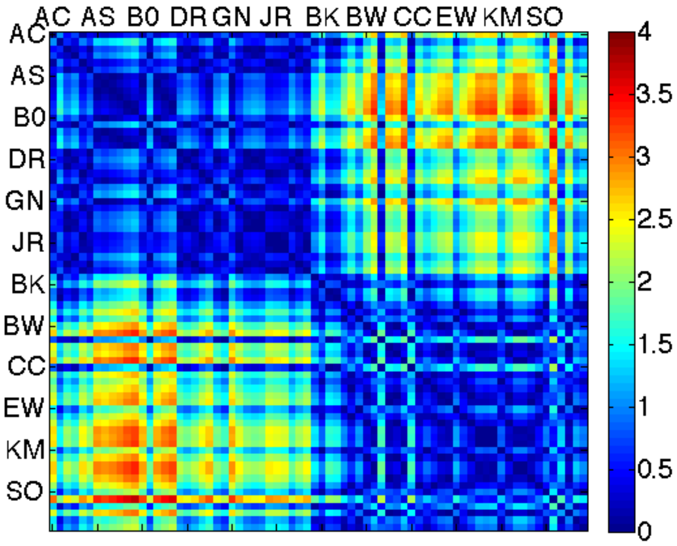


Figure 4.4: **RDMs showing the Euclidean distances between voice recordings of the 12 identities based on measures of f0 (A) and AVTL (B).** Matrices are symmetric around a diagonal of zeros, and each cell shows the distance between the stimuli in the corresponding row and column.

4.1.4 Data analysis

RSA comparing brain and candidate RDMs

For each individual participant and each ROI, we used their brain RDM for faces or for voices as a reference, and compared it with candidate RDMs computed from their behavioural ratings of the faces or voices (trustworthiness, dominance, attractiveness, positive-negative valence, perceptual similarity), with a RDM for identity gender, and for RDMs of visual (OpenFace, Gabor-Jet) and acoustic features (f0, AVTL) (Figure 4.5). Comparisons between the reference and candidate RDMs were made using Kendall’s tau a correlation, and p-values for each comparison were obtained by computing a one-sample Wilcoxon signed-rank test across the correlation values for all participants, comparing them against zero (Nili et al., 2014). P-values were corrected for multiple comparisons using FDR correction (*q*=.05) across all comparisons within each ROI (8 for faces, 8 for voices).

An estimate of the noise ceiling was calculated for each ROI and modality to indicate the maximum achievable correlation with a candidate RDM given the noise in the brain RDMs. A lower bound of this noise ceiling (Nili et al., 2014) was calculated by correlating (tau a) the brain RDM for each participant with the average brain RDM across all other participants, after sign rank transforming the RDMs, and then averaging these correlations across all participants. Please note that this estimate of noise ceiling is also an estimate of inter-subject reliability of the brain RDMs.

RSA comparing brain RDMs across different ROIs

Finally, for each individual participant, face or voice RDMs were compared between different brain regions to assess their similarity in terms of information content. Individual RDMs were compared using Spearman correlation, and a p-value for each comparison was computed using one-sample Wilcoxon signed-rank test across the correlation values for all participants, comparing them against zero. P-values were corrected for all comparisons using FDR correction with *q*=.05. Finally, correlations were averaged across participants for visualisation purposes by Fisher-transforming the correlation values, averaging them, and then reverse transforming the resulting values.

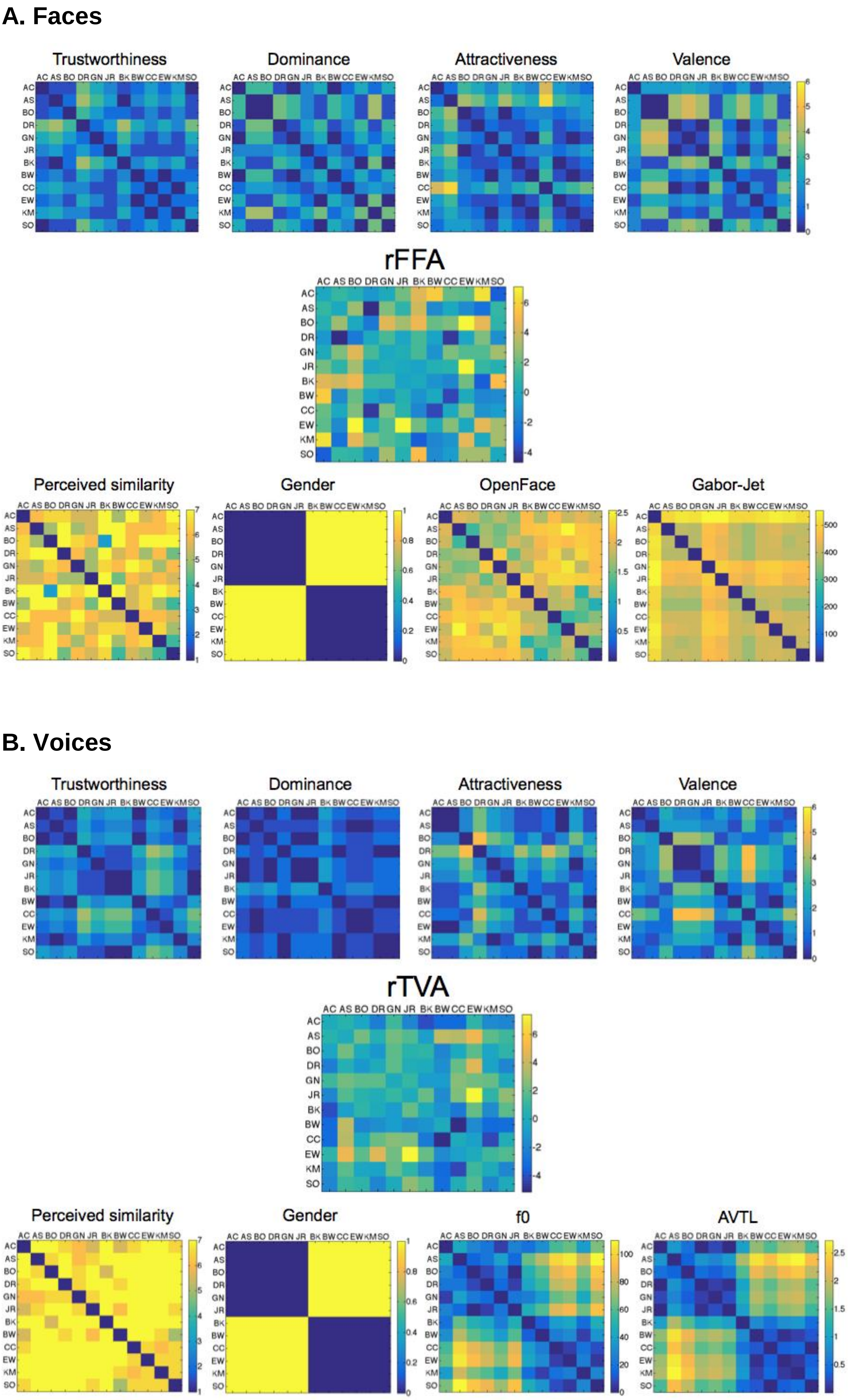

A. Faces
Trustworthiness
Dominance
Attractiveness
Valence
rFFA
Perceived similarity
Gender
OpenFace
Gabor-Jet
B. Voices
Trustworthiness
Dominance
Attractiveness
Valence
rTVA
Perceived similarity
Gender
f0
AVTL
AC AS BO DR GN JR BK BW CC EW KM SO

Figure 4.5: **Example RDMs for faces (A) and for voices (B) for one participant.** Matrices are symmetric around their diagonals. A: Example brain RDM for the rFFA (centre) and candidate RDMs (top and bottom row). B: Example brain RDM for the rTVA (centre) and candidate RDMs (top and bottom row). Note that the OpenFace and Gabor-Jet (for faces), f0 and AVTL (for voices), and gender RDMs are based on stimulus characteristics, and are therefore the same for all participants. The gender RDM is the same for both faces and voices.

4.2 Results

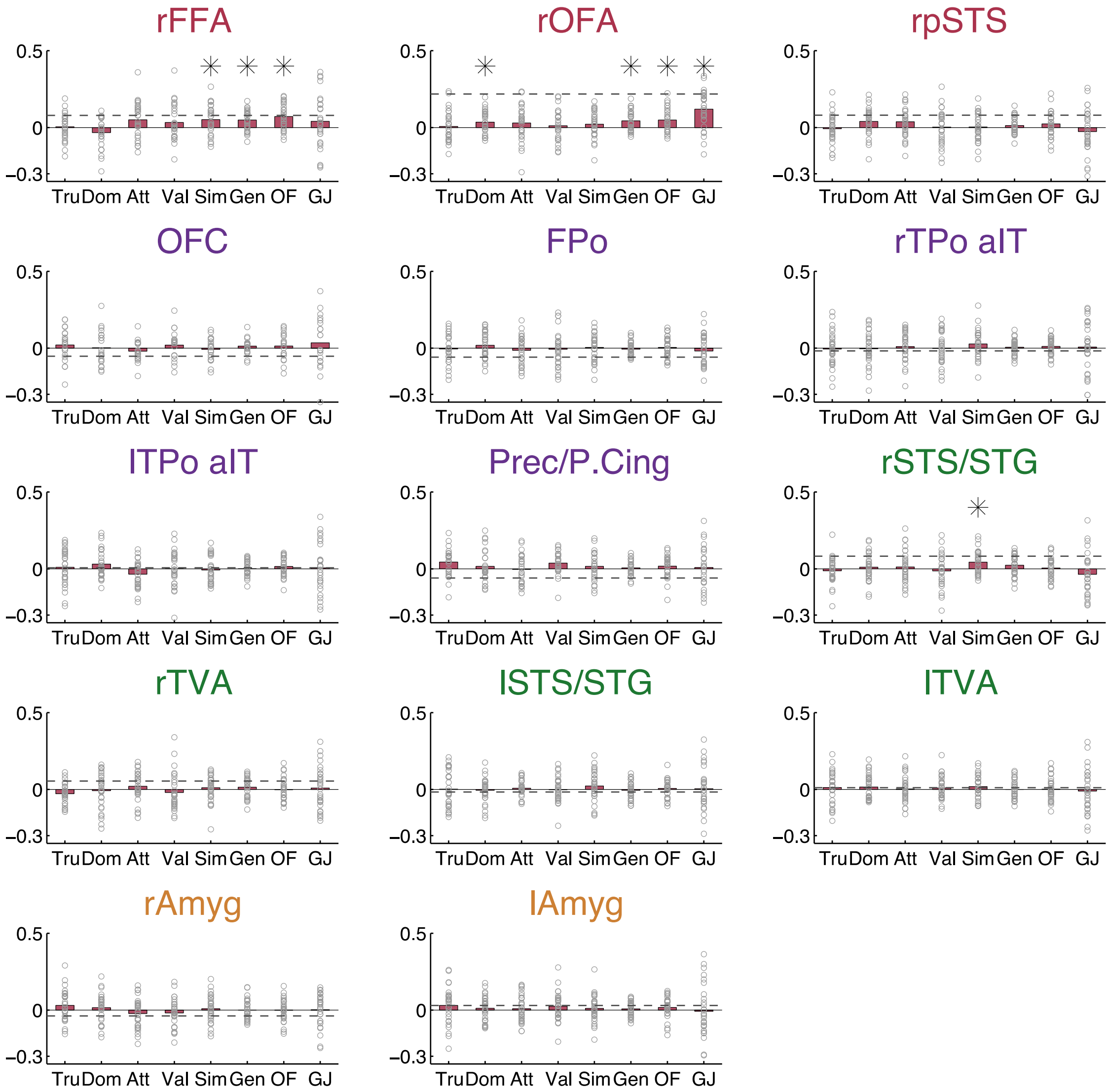
4.2.1 RSA comparing brain and candidate RDMs

Figure 4.6A shows correlations between brain RDMs for faces and candidate model RDMs in different brain regions. Significant correlations were found in the face-selective rFFA for perceived pairwise visual similarity (*mean r* = .05, *Z* = 2.585, *p* =.0097), gender (*mean r* = .05, *Z* = 3.074, *p* =.0021), and OpenFace similarity (*mean r* = .07, *Z* = 3.564, *p* =.0004), and in the rOFA for dominance (*mean r* = .03, *Z* = 2.325, *p* =.0201), gender (*mean r* = .04, *Z* = 2.854, *p* =.0043), OpenFace similarity (*mean r* = .05, *Z* = 2.887, *p* =.0039), and Gabor-Jet similarity (*mean r* = .12, *Z* = 3.882, *p* =.0001). In addition, there was a correlation with perceived pairwise visual similarity in the voice-selective rSTS/STG (*mean r* = .04, *Z* = 2.736, *p* =.0062). The noise ceiling, indicated by the dashed lines in Figure 4.6A, was low for all ROIs. From the correlations that were significantly greater than zero, only the correlation with OpenFace similarity in the rFFA approached the noise ceiling (*τ* = .08).

Figure 4.6B shows the correlations between brain RDMs for voices and candidate model RDMs. Significant correlations with gender, f0, and AVTL were found in all four voice-selective regions (bilateral TVAs and STS/STG), and significant correlations with perceived pairwise auditory similarity were found in all voice-selective regions apart from the lTVA. Specifically, for the rSTS/STG, perceived similarity: *mean r* = .07, *Z* = 4.103, *p* <.0001; gender: *mean r* = .08, *Z* = 4.401, *p* <.0001; f0: *mean r* = .10, *Z* = 3.909, *p* <.0001; AVTL: *mean r* = .10, *Z* = 3.877, *p* =.0001. For the rTVA: perceived similarity: *mean r* = .06, *Z* = 3.754, *p* =.0002;

gender: *mean r* = .05, *Z* = 3.610, *p* =.0003; f0: *mean r* = .08, *Z* = 3.363, *p* =.0008; AVTL: *mean r* = .07, *Z* = 3.240, *p* =.0012.For the lSTS/STG: perceived similarity: *mean r* = .05, *Z* = 3.445, *p* =.0006; gender: *mean r* = .08, *Z* = 4.309, *p* <.0001; f0: *mean r* = .10, *Z* = 4.289, *p* <.0001; AVTL: *mean r* = .11, *Z* = 4.042, *p* <.0001.For the lTVA: gender: *mean r* = .04, *Z* = 3.209, *p* =.0013; f0: *mean r* = .06, *Z* = 2.962, *p* =.0031; AVTL: *mean r* = .04, *Z* = 2.447, *p* =.0144. There was also a correlation with perceived pairwise auditory similarity in the multimodal rTP-aIT (*mean r* = .06, *Z* = 3.672, *p* =.0002). Similar to the faces, the noise ceiling was low for all ROIs. Only the average correlation with perceived similarity in rTP-aIT, and the correlation with f0 in the lTVA surpassed the noise ceiling (rTP-aIT: *τ* = 02, lTVA: *τ* = 05), whereas the correlation with AVTL in the lTVA approached the ceiling but did not surpass it.

A. Faces


rFFA
rOFA
rpSTS
OFC
FPo
rTPo aIT
lTPo aIT
Prec/P.Cing
rSTS/STG
rTVA
lSTS/STG
lTVA
rAmyg
lAmyg
0.5
0
−0.3
Tru Dom Att Val Sim Gen OF GJ

B. Voices

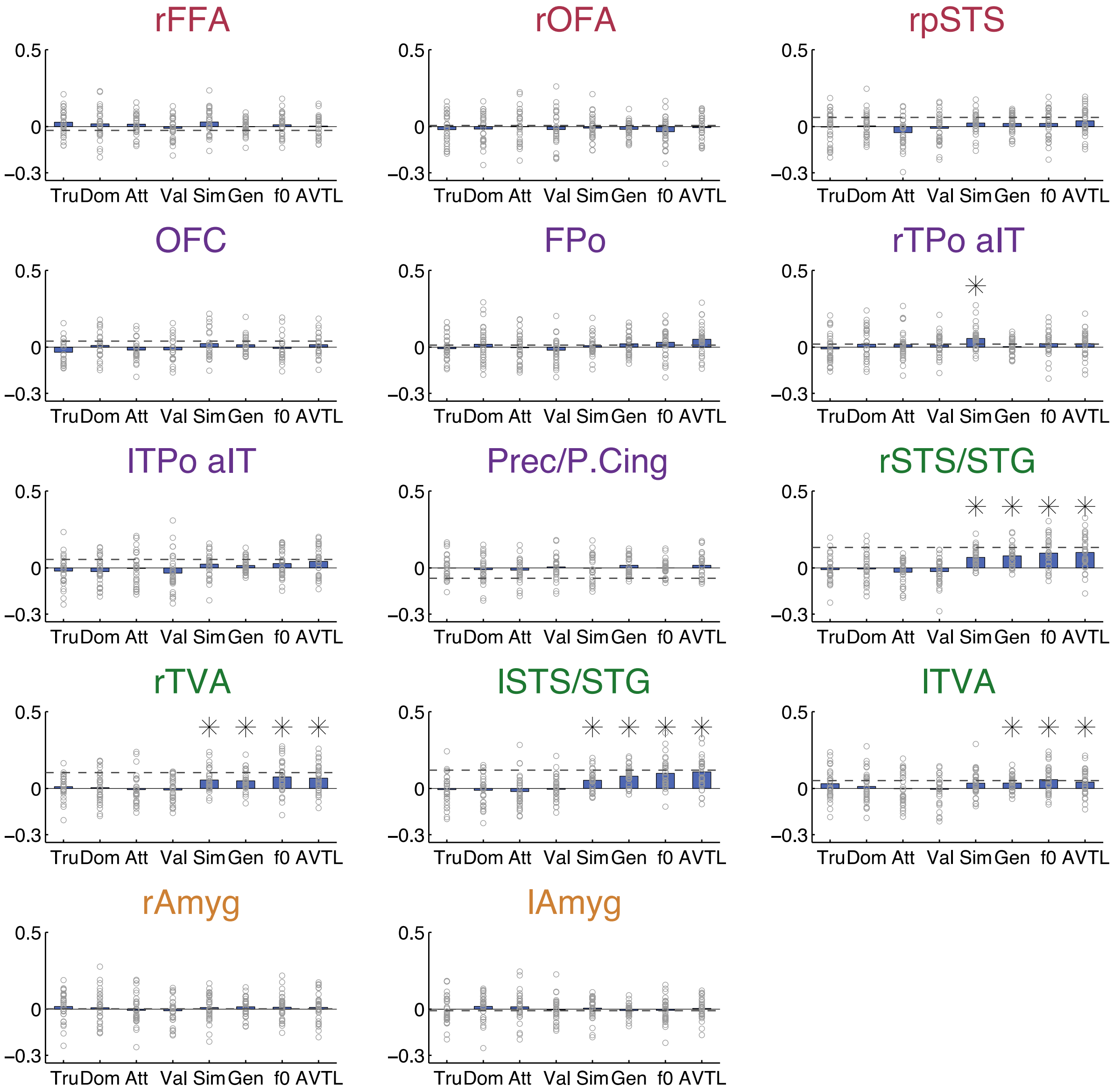


Figure 4.6: **Correlations between brain RDMs for faces (A) and voices (B) and candidate model RDMs**. The candidate model RDMs are trustworthiness (Tru), dominance (Dom), attractiveness (Att), pairwise visual/auditory similarity (Sim), stimulus gender (Gen), OpenFace similarity (OF) and Gabor-Jet similarity (GJ) for faces, and fundamental frequency (f0) and apparent vocal tract length (AVTL) for voices. Bars show average correlations across participants, and circles show individual participants. Stars indicate statistical significance at *q*≤.05. Horizontal dashed lines show the lower bound of the estimated noise ceiling (i.e. inter-subject reliability of brain RDMs). Different ROIs are colour-coded based on whether they

were originally defined as face-selective (magenta), voice-selective (green), multimodal (purple), or based on anatomical masks (orange). Note that the rpSTS was originally defined as face-selective, but was found to demonstrate multimodal properties (see Chapter 3).

Although the candidate model RDMs describe different information extracted from faces and voices, it is likely that some of these types of information are associated with each other. For example, voice gender discrimination has previously been associated with voice pitch (Pernet & Belin, 2012). To determine to what extent the different candidate RDMs shared information with each other, we computed Kendall tau a correlations between all of the face models and all of the voice models. Correlations were computed separately for each participant. We then averaged the correlations across participants (we computed the average across participants after Fisher transforming the single-subject correlations, and reverse transformed the resulting values). Correlations between models that were based on stimulus properties, rather than on participant ratings, were the same for all participants, and were thus excluded from inference testing. For all other comparisons, we compared the single-subject correlations against zero using one-sample Wilcoxon signed-rank tests, and corrected for multiple comparisons using FDR with $q \leq .05$.

Figure 4.7 shows that the candidate RDMs that explained the brain RDM for faces in the rFFA, namely perceived similarity, gender, and OpenFace, were also correlated with each other, suggesting that these models capture similar information about faces. For candidate RDMs that explained the brain RDM for faces in the OFA, apart from the correlation between gender and OpenFace distance, there do not appear to be strong associations for the other models, i.e. there is not a significant correlation between dominance and Gabor-Jet similarity. Therefore, it is likely that the RDMs for dominance and Gabor-Jet similarity capture distinct types of face information in the OFA. Finally, Figure 4.7 shows that the models that were associated with responses to voices in the voice-selective regions, i.e. perceived similarity, gender, f0, and AVTL, were correlated with each other, suggesting that they explain similar information about voice representations in these regions.

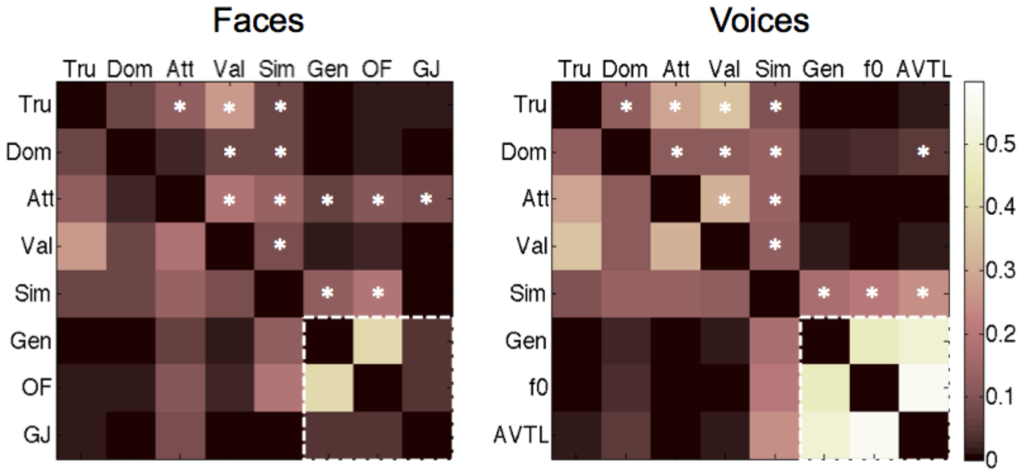


Figure 4.7: Correlations between candidate RDMs for faces and for voices. The matrix is symmetric around a diagonal of zeros, and each cell shows the correlation between the models stated in the corresponding row and column. Correlations have been averaged across participants in all cells except for those enclosed in white dashed lines, for which the RDMs were the same for all participants. Stars in the upper triangle indicate correlations that survived correction for multiple comparisons (FDR corrected p value for faces: p≤.0256, and for voices: p≤.0148).

To summarise, our results show that the face-selective rFFA and rOFA represent information relating to facial appearance, and voice-selective regions represent information relating to vocal appearance, as described by both perceived and objective measures of similarity. From the multimodal regions, only the rTPo-aIT showed a significant correlation with perceived similarity for voices, and this is likely to be due to the overlap between this region and the voice-selective regions in the right hemisphere. Our results also show that the rOFA represents information about perceived facial dominance. However, we find no significant correlations between face or voice brain RDMs with candidate RDMs for perceived trustworthiness, attractiveness, and valence in any of our ROIs after correcting for multiple comparisons.

4.2.2 RSA comparing brain RDMs across different ROIs

Figure 4.8 shows the Spearman correlations between face RDMs and voice RDMs in different ROIs, averaged across participants. The upper triangle shows correlations with p-values that survived FDR correction for multiple comparisons at $p \leq .0115$. Note that some ROIs showed large spatial overlap (rTPo-aIT, rpSTS, rSTS/STG, and rTVA with each other, and lTPo-aIT, lSTS/STG and the lTVA with each other), and the correlations between them are likely due to overlapping voxels. The results show significant correlations between face RDMs in face-selective regions, and between voice RDMs in voice-selective regions across different hemispheres, suggesting that these regions share information in the respective modality. Within each modality, there were also significant correlations across face-selective, voice-selective, and multimodal brain regions, demonstrating that information may be shared across these different regions. In addition, the results show low, but significant correlations between voice RDMs in face-selective regions, and between face RDMs in voice-selective regions. These findings suggest that face- and voice-selective regions may process and share some information relating to the other modality.

In the amygdala ROIs, which were anatomically defined, we show correlations of the face RDMs with face RDMs for face-selective and multimodal regions, and of the voice RDMs with voice RDMs for multimodal regions and the voice-selective rSTS/STG. Even though we did not find correlations of face or voice RDMs in the amygdala with any of our candidate models in the previous analysis, these results suggest that representations of faces and voices in this region share information with representations in both unimodal and multimodal (face- and voice-responsive) regions.

Finally, a small number of correlations between RDMs across modalities survived multiple comparisons correction. A correlation between the face and voice RDM was found in the rTVA (*mean r* =.09, Z = 2.561, p =.0104). In Chapter 3, there were no significant correlations between face and voice RDMs in any of the ROIs when comparing RDMs across different scanning sessions. The difference here is that the RDMs were averaged across scanning sessions before being compared between the two modalities. Correlations between face and voice RDMs were also found across different brain regions. However, given the small size of the correlations and the high number of comparisons across regions and modalities, these results should be interpreted cautiously.

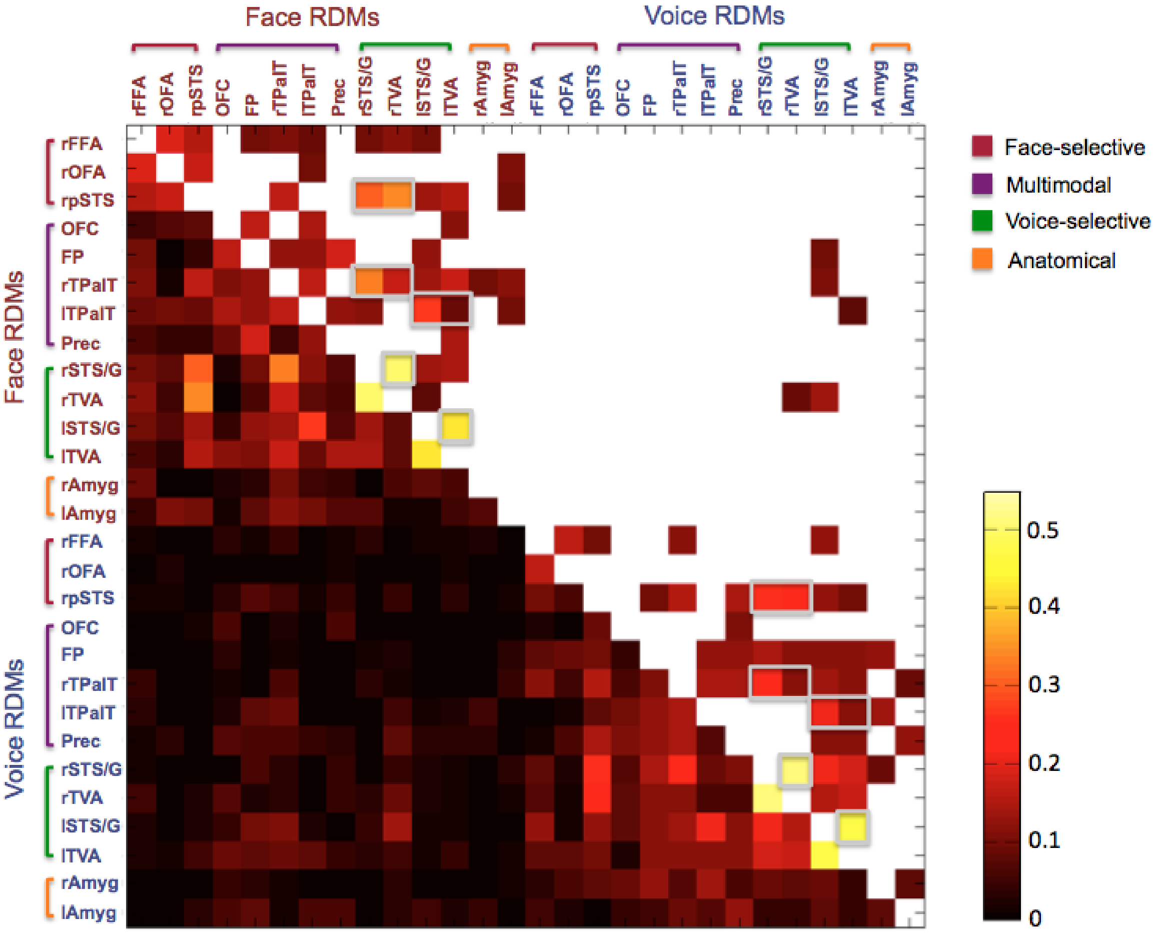
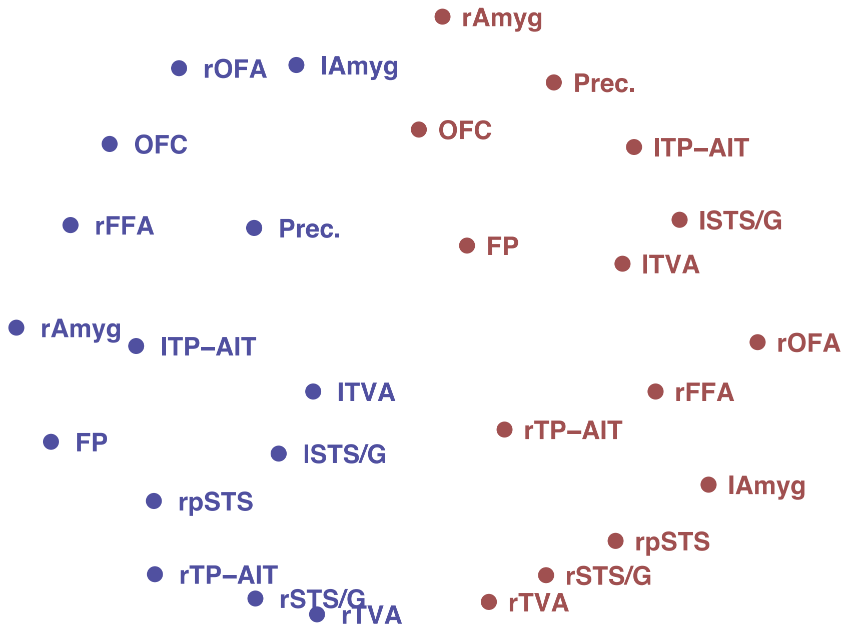


Figure 4.8: **Correlations between face and voice RDMs in face-selective, multimodal, voice-selective, and anatomically defined ROIs, averaged across all participants.** Each cell shows the average correlation between the ROIs in the corresponding row and column. The lower triangle shows all correlations, and the upper triangle only shows correlations that survived correction for multiple comparisons (FDR corrected p value: *p*≤.0115). Grey outlines around cells in the upper triangle show correlations between brain regions that are likely due to large overlap. Different ROIs are colour-coded based on whether they were originally defined as face-selective (magenta), voice-selective (green), multimodal (purple), or based on anatomical masks (orange). Note that the rpSTS was originally defined as face-selective, but was found to demonstrate multimodal properties.

Finally, we used multidimensional scaling (MDS) to visualise the similarity between face and voice RDMs in different ROIs. The Spearman correlations were first transformed into correlation distances using 1-correlation. Two-dimensional MDS was computed in Matlab using the metric stress criterion and 1000 iterations. Figure 4.9A shows the relationship between face RDMs in each ROI (red) and voice RDMs in each ROI (blue), and demonstrates a clear distinction between the two modalities. Figure 4.9B shows the relationship between face RDMs only, and Figure 4.9C shows the relationship between voice RDMs only. Both of these plots show a distinction between face-selective and voice-selective brain regions, and a similar organisation of these regions in both modalities. For face-selective regions, both plots show a posterior to anterior organisation of face-selective regions (rOFA-rFFA-rpSTS).

A) Faces & Voices



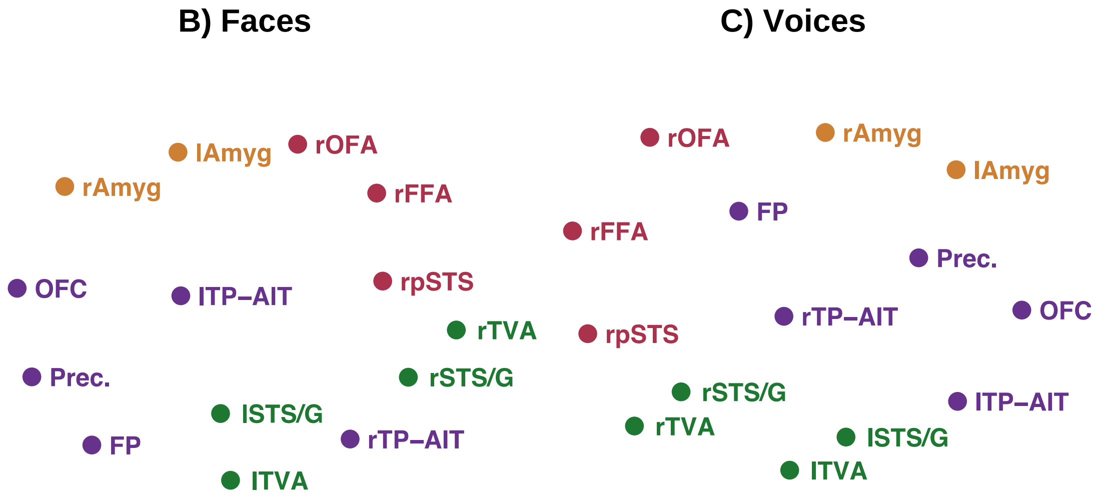


Figure 4.9: **Similarity of face RDM and voice RDMs across different ROIs.** The plots show the results of 2D MDS performed on the dissimilarities between face and voice RDMs (A), on the dissimilarities between face RDMs only (B), and on the dissimilarities for voice RDMs only (C). Regions that are close together (or even overlap) show similar representational geometries, and regions that are further apart show dissimilar representational geometries. A shows clustering of RDMs based on modality, with voice RDMs in blue and face RDMs in red, with a clear distinction between the two modalities along the first dimension (x axis). In B and C different ROIs are colour-coded based on whether they were originally defined as face-selective (magenta), voice-selective (green), multimodal (purple), or based on anatomical masks (orange). Note that the rpSTS was originally defined as face-selective, but was found to demonstrate multimodal properties. Both B and C show a distinction between face-selective and voice-selective regions across the second dimension (y axis), and a posterior to anterior organisation of face-selective regions.

4.3 Discussion

The study presented in this chapter showed that brain representations of face identities in the rFFA and rOFA were associated with information relating to the objective visual similarity and gender of these identities, and brain representations of face identities in the rFFA were also associated with information about the perceived visual similarity of faces. In addition, brain representations of voice identities in voice-selective regions (bilateral TVAs and STS/STG) were associated with information relating to the perceived and the objective auditory similarity and gender of these identities. These findings supported the hypothesis that face-selective and voice-selective regions would process information relating to physical stimulus properties. In contrast, the findings did not provide support for the hypotheses that social information from faces and voices would be represented in multimodal brain regions, and that the amygdala would represent both social and physical information in faces. Instead, a model of perceived facial dominance correlated with brain representations in the face-selective rOFA, and no other models of social information in faces or voices correlated with brain representations in any other ROIs. No significant correlations with any of the candidate models were found in the amygdala or in multimodal regions, with the exception of auditory similarity for voices in the rTPo-aIT.

Finally, a further analysis comparing brain representational geometries across different brain regions showed similarities in face representations across different face-selective regions (rFFA, rOFA, and rpSTS), and similarities in voice representations across voice-selective regions, within and across both hemispheres, suggesting commonalities in the informational content of these regions.

Informational content of brain representations of faces

Information about the perceived similarity of face identities, i.e. their subjective visual appearance, was associated with brain representations of these identities in the rFFA, and in the voice-selective rSTS/STG. Specifically, pairs of identities that were rated as being similar in terms of their visual appearance also elicited similar multivoxel response patterns in these regions. It is particularly interesting that such a representation was found in the rFFA. While many studies had shown that the rFFA contributes to discriminating between different face identities (Anzellotti et al., 2014; Axelrod & Yovel, 2015; Nestor et al., 2011; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017), the current findings further show that brain representations in this region are also related to the subjective judgment of visual similarity. While previous studies have compared brain responses with perceived visual similarity of objects (Charest & Kriegeskorte, 2015; Mur et al., 2013), with perceived similarity of animals from different biological classes (Connolly et al., 2012), and with perceived similarity of different emotional facial expressions (Said et al., 2010; Sormaz et al., 2016) or emotional concepts (Saarimaki et al., 2015; Saarimäki et al., 2018), to the best of my knowledge this is the first study to show correlations between brain responses and the perceived similarity of different face identities. This is important given that, although a great deal is known about the selective responses to faces in FFA, much less is known about the computations in this region. The results of the present study suggest that these computations may be related to the perceived visual similarity of faces.

The correlation with perceived visual similarity in the voice-selective rSTS/STG could be due to the presence of a face-selective region (or regions) contained within this relatively large ROI. One possibility could be that the overlap of the rSTS/STG with the rpSTS contributed to these results, but the results showed no significant correlations with perceived similarity in the rpSTS. However, face-selective regions have been identified in other regions of the STS, such as the anterior part (Fox et al., 2009; Pinsk et al., 2008; Pitcher et al., 2011), and it is possible that these regions could be driving the correlation with perceived visual similarity in the rSTS/STG.

The brain representations of faces in both the rFFA and the rOFA were also related to objective models of similarity that were computed based on visual image properties. Whereas the brain representations in rFFA and rOFA related to visual similarity computed using the OpenFace program, only the rOFA showed a significant correlation with visual similarity based on the Gabor-Jet model. The absence of a significant correlation in the rFFA is in contrast with studies showing that the Gabor-Jet similarity of pairs of faces is associated with the magnitude of the adaptation to the faces in the FFA (Xu & Biederman, 2010; Xu et al., 2009). However, in contrast to these studies, the present study used stimuli that were naturalistically variable, even within the same individuals, and this may have contributed to the differences in findings. The present results suggest that the Gabor-Jet model does not generalise well to different tokens of the same identity, and this may account for why this model can better explain the brain representations in the rOFA compared to the rFFA.

The analysis looking at the relationship between the different models showed that whereas OpenFace similarity was associated with face gender and perceived similarity, Gabor-Jet similarity was not associated with any of the other candidate models (with the exception of perceived attractiveness). Therefore, it appears that these two models capture different types of visual information that distinguishes different face identities. Based on the RDMs computed at the stimulus level, prior to averaging across different stimuli for the same identities, it seems that the OpenFace program performed much better than the Gabor-Jet model at ‘telling together’ different stimuli that belonged to the same identity. Thus, the OpenFace similarity model may be related to facial characteristics that are used to distinguish between different face identities, but can also ‘tell together’ images of the same person. However, the characteristics used by the OpenFace neural network to create face descriptors are not defined, and therefore any interpretation of its computations is largely speculative. In contrast, the Gabor-Jet model was designed to simulate response properties of cells in the primary visual cortex, and therefore describes low-level image properties. The correlation of the Gabor-Jet model with activity in the OFA suggests that the OFA may process primarily low-level characteristics of images. This would explain why no correlation was found between the OFA and perceived visual similarity, which is most likely based on higher-level face properties such as gender, age, or race.

Brain representations in both the rFFA and rOFA were also related to face gender. Specifically, the results showed that different-gender face identities were associated with more dissimilar response patterns compared with same-gender face identities. These findings support previous studies showing representations of face gender in the FFA (Contreras et al., 2013; Freeman et al., 2010; Kaul et al., 2011) and in the OFA (Kaul et al., 2011). However, whereas Kaul et al. (2011) also found representations of gender in the STS and OFC, there was no evidence of gender representations in these regions.

The analysis comparing different models showed that the OpenFace, perceived similarity, and gender models capture similar information about faces. This is unsurprising, given that behavioural judgements of similarity are likely to be based on physical facial features. Moreover, given that facial features are used to distinguish between faces of different genders (Bruce et al., 1993), it is likely that the gender model also reflects differences between faces in terms of visual features. Due to the similarities between the OpenFace, perceived similarity, and gender models, it is not possible to draw conclusions regarding the independent contributions of these models in explaining brain representations in the present study.

The only representation of social trait information was found in the rOFA for perceived dominance. Dominance, sometimes referred to as power, has been shown to be one of the main dimensions that describe first impressions of faces, and has been associated with facial masculinity (Oosterhof & Todorov, 2008; Sutherland et al., 2013). Todorov et al. (2011) have previously shown a relationship between the level of perceived facial power and the magnitude of response in a region of the right occipital cortex, the peak of which overlaps with a probabilistic map of the rOFA in the present study. Specifically, Todorov et al. (2011) found that this region showed stronger responses to faces perceived as very high or very low in power, compared with faces perceived to have a medium level of power. However, they failed to replicate this effect in a second experiment that used an orthogonal task, as opposed to an approach-avoidance decision task, suggesting that responses to dominance in the right occipital regions may be task-dependent. The current study, using a task unrelated to social judgements, showed that the representation of face identities in the rOFA contains some information about the perceived dominance of the faces. The facial dominance model was not correlated with any of the other models that were correlated with the brain RDM in the rOFA. This suggests that the dominance information in the rOFA could not be explained by our models of visual appearance, such as OpenFace and Gabor-Jet similarity, or by differences in the gender of the faces. Therefore, it seems that the rOFA may represent both perceptual and social information about faces.

To summarise, Chapter 3 showed that both the rFFA and rOFA could discriminate between different face identities. The current findings suggest that the rFFA may use information related to both the perceived and objective visual similarity between faces to distinguish between the faces of different people, whereas the rOFA may rely more on the objective image-based similarity between faces. Moreover, both the rFFA and rOFA may use gender information to distinguish between different identities, and the rOFA may additionally process information about facial dominance. Models that correlated with brain representations of faces in the rFFA, namely perceived similarity, gender and OpenFace similarity, were correlated with each other, suggesting that they describe similar information. In contrast, low correlations were found between the models that correlated with brain representations in the rOFA, namely perceived dominance, gender, OpenFace similarity, and Gabor-Jet similarity (with the exception of gender and OpenFace similarity), suggesting that these models describe distinct information. Although Chapter 3 also revealed representations of face identity in the rpSTS, none of the models described in the present chapter explained the representational geometry of face identities in this region.

Informational content of brain representations of voices

Brain representations in the rTVA and the bilateral voice-selective STS/STG were associated with information about the perceived auditory similarity of voice identities, and brain representations in all voice-selective regions were associated with information about voice gender, f0, and AVTL. The multimodal rTP-aIT, which overlaps with the voice-selective regions in the right temporal lobe, showed a significant correlation with perceived similarity only. Given the extremely low noise ceiling in this region, these results are interpreted with caution. In addition, the observed correlation in this region is most likely due to the overlap with the voice-selective regions. No significant correlations were found between brain representations and candidate models in any other multimodal RDMs. Moreover, none of the ROIs represented social trait information in voices.

For measures of objective similarity, the finding that the bilateral STS/STG and the TVAs represent information about f0 is in line with Formisano et al. (2008), who showed that the distances between the brain representations of individual voice identities in the STS/STG were correlated with the distances between the f0 of the voices. While Formisano et al. (2008) used vowel sounds as stimuli, the present work extends this finding to representations of longer, naturalistically varying speech stimuli. Moreover, in contrast to Formisano et al. (2008), this chapter showed that f0 information is used to distinguish between voices within independently localised voice-selective regions. The finding that the bilateral STS/STG and the TVAs represent information about AVTL is in agreement with previous studies showing that the left STS/STG is sensitive to VTL information in voices (von Kriegstein et al., 2007, 2010), and extends this finding to include the right hemisphere. Lastly, a previous study found that a region in the aSTS that overlapped with the right TVA, but not the right TVA itself, responded more to very masculine and very feminine voices than to gender-ambiguous voices, suggesting that only a sub-region of the right TVA processes information about voice gender (Charest et al., 2013). In contrast, the present study showed that both the right and left TVAs use gender information to discriminate between individual voices.

To the best of my knowledge, this is the first time that activity in voice-selective regions has been associated with perceived vocal similarity. The present findings showed that pairs of identities that were rated as being similar in terms of the way that they sound also elicited similar multivoxel response patterns in voice-selective regions.

The analysis comparing the similarity between different candidate models showed that perceived auditory similarity, voice gender, f0, and AVTL were correlated with each other, suggesting that they capture similar information from voices. This finding is in line with studies comparing MDS of perceived similarity ratings of voices with different acoustic measures, which have shown that perceptual judgements are associated with f0 and with formant frequencies, which are used to calculate AVTL (Baumann & Belin, 2009; Nolan, Mcdougall, & Hudson, 2011). As with faces, it is not surprising that judgements of perceived similarity of voices are likely to be based on acoustic features of the voice. Moreover, voice gender discrimination has been previously associated with both f0 and formant frequencies (Fitch & Giedd, 1999; Hillenbrand & Clark, 2009; Pernet & Belin, 2012; Poon & Ng, 2011; Titze, 1989). Ultimately, the similarities between the different models make it almost impossible to disentangle the contributions of each of the models to the brain representations of voices.

To summarise, Chapter 3 showed that the bilateral STS/STG and TVAs could discriminate between individual voice identities, and the present study suggests that these regions may use information about the subjectively and objectively defined similarity between voice identities, as well as voice gender, to discriminate between different identities.

Multimodal regions and the amygdala

It was predicted that social information could be represented in multimodal regions, but no evidence was found to support this prediction. With the exception of the association between perceived dominance and brain representations of faces in the rOFA, there was no evidence of representations of social information from faces and voices in any other ROIs. Furthermore, no correlations were found between brain representations for faces and voices in the rpSTS and any of the candidate models. Therefore, while the previous chapter showed that the rpSTS could discriminate between face identities, voice identities, and person identities, it is not known what type of information is used by this region to distinguish between the different identities. It has been proposed that the STS may integrate dynamic information that is extracted from the faces and voices of familiar people and that is unique to each individual (Yovel & O'Toole, 2016). Exposure to faces and voices during social interactions is usually concurrent, and it is likely that dynamic aspects of person's face would be automatically associated with their voice and manner of speech (Yovel & O'Toole, 2016). Given that the pSTS has been shown to be sensitive to dynamic information in faces (Bernstein, Erez, Blank, & Yovel, 2018; Fox et al., 2009; Pitcher et al., 2011), it is possible that it also integrates dynamic information from faces and voices at a person identity level. A second possibility is that the pSTS represents the degree of familiarity with different identities (Parkinson, Liu, & Wheatley, 2014), which may be idiosyncratic for each participant. Parkinson et al. (2014) have shown that the pSTS represents the 'social distance' between faces in terms of familiarity, according to which highly familiar faces are perceived as being 'closer' to a person whereas less familiar faces are perceived as being 'further away' (Parkinson et al., 2014). It is likely that participants in the present study would have had different levels of exposure to different identities, resulting in different levels of perceived familiarity. Thus, it is possible that the observed person identity representation in the pSTS reflects the different degrees of familiarity between different identities, regardless of whether this information is extracted from the face or from the voice. Future work should attempt to compare brain representations of faces and voices in the pSTS with models of dynamic face and voice properties and of face and voice familiarity.

No evidence was found to support the prediction that the amygdala would represent both physical and social characteristics of faces. This prediction was based on univariate studies showing sensitivity to both social and non-social facial characteristics in the amygdala (Bzdok et al., 2011; Mende-Siedlecki et al., 2013). It may be that while this sensitivity is evident when examining at the magnitude of response in the amygdala, it is not evident when distinguishing between multivoxel activity patterns in response to individual face identities. It is also possible that familiar faces and voices do not automatically engage social processing in the amygdala. Gobbini et al. (2004) showed that the amygdala responded more to unfamiliar faces compared to familiar faces during a one-back task that did not explicitly encourage the processing of social information. It may be that using an experimental task that encourages social processing would activate representations of social information from familiar faces and voices, and future studies should investigate this possibility.

Shared information between face-selective, voice-selective, and multimodal regions

This study presented a further exploratory analysis comparing representational geometries for face identities and for voice identities across different brain regions, which revealed that face-selective regions share face information with each other, and voice-selective regions share voice information with each other even across hemispheres. These findings support the results of the analysis comparing brain RDMs to candidate models RDMs within each of these regions, which showed that brain representations in both the rFFA and rOFA are associated with face gender and objective facial similarity (computed from the OpenFace program), and that brain representations in all voice-selective regions are associated with information about voice gender, perceived similarity (except for the lTVA), f0, and AVTL. Despite finding no correlations between representations of faces and voices in the rpSTS and models of informational content in the previous analysis, the present analysis suggested that the rpSTS shares some (undefined) information with the rFFA and the rOFA. The finding of similar representational geometries between the FFA and OFA is also in agreement with studies showing white matter structural connections (Gschwind, Pourtois, Schwartz, Van De Ville, & Vuilleumier, 2012; Pyles, Verstynen, Schneider, & Tarr, 2013) and functional connectivity between the two regions (Davies-Thompson & Andrews, 2012; O'Neil, Hutchison, McLean, & Köhler, 2014). Moreover, two effective connectivity studies showed feed-forward connections from the OFA to the FFA and STS (Fairhall & Ishai, 2007), and from the FFA and TVA to the pSTS (Davies-Thompson et al., 2018). Finally, one study showed functional connectivity between the pSTS and FFA (Turk-Browne, Norman-Haignere, & McCarthy, 2010).

This analysis also showed that face-selective regions, voice-selective regions, multimodal regions, and the amygdala share information from both faces and voices within each modality. This suggests that these regions may exchange information, and future work should further investigate the nature of the information shared across the different regions.

Finally, in terms of the organisation of representations in different regions, a two-dimensional MDS solution for face and voice RDMs across all ROIs revealed a distinction, in both modalities, between representations in face-selective and voice-selective brain regions. This suggests that the face and voice representations across different brain regions are more similar to each other within the same modality than across modalities regardless of the type of brain region (face-selective, voice-selective, or multimodal). Moreover, MDS solutions computed separately for face and voice representations showed a posterior to anterior organisation of face-selective regions from the rOFA to the rFFA and the pSTS for representations in both modalities, which mirrors their location in the brain. In other words, regions that are closer to each other in the brain had more similar brain representations than regions that are farther apart.

Limitations and future directions

An important limitation of this study is the low noise ceiling, computed as inter-subject reliability, in the ROIs. The current results showed that the mean inter-subject reliability of brain RDMs was in fact low (varying between -.07 and .22), and therefore future studies should consider ways to improve inter-subject reliability. On the other hand, it is possible that this measure of noise ceiling is not entirely suitable for the current analysis. Specifically, it is possible that the variability in face and voice representations across different participants was due to the idiosyncratic nature of the representations, which could be influenced by the level of familiarity and degree of exposure of each participant with the presented face and voice identities. If there is indeed a strong unique component to the representations, then inter-subject reliability may not be an ideal measure of the noise ceiling. This could be particularly relevant given that individual behavioural RDMs were correlated with individual brain RDMs. Another possibility is that the use of brain RDMs to compute inter-subject reliability does not reflect the reliability of the brain activity patterns themselves. An analysis reported in Chapter 3, which investigated the intra-subject reliability of brain activity patterns elicited by individual face and voice identities across the two scanning sessions, showed that activity patterns for faces in face-selective regions, and for voices in voice-selective regions were highly reliable. In contrast, the correlations between the RDMs that were computed based on these activity patterns across the two scanning sessions were very low. This raises the possibility that there was insufficient variability in the representational distances between the 12 identities within each modality, which resulted in the 12 identities being similarly distinct from each other. This would make any similarities between these RDMs difficult to detect using correlation. If this is the case, it is possible that similarities between brain activity patterns across different participants would also be underestimated when comparisons are made at the RDM level. However, due to the use of participant-specific ROIs (with different numbers of voxels) it is not possible to test the inter-subject reliability of the activity patterns themselves. Future studies could focus on optimising protocols for more reliable (intra-subject and inter-subject) RDMs.

Lastly, it should be noted that many of the candidate models that showed significant correlations in the ROIs failed to approach the noise ceiling for those ROIs. This highlights the need to develop models that better explain representational geometries for face and voice identities in different brain regions.

Conclusion

To conclude, this study confirmed the prediction that brain representations of faces in face-selective regions would be associated with information about perceived and objective visual similarity of faces and face gender, and that brain representations of voices in voice-selective regions would be associated with information about perceived and objective auditory similarity of voices and voice gender. In contrast, the hypothesis that multimodal regions would represent social information from faces and voices was not supported by the results. Moreover, no representations of social or physical information from faces were found in the amygdala. It is important to note that many of the candidate models that were used to explain brain representational geometries were correlated with each other, suggesting that they capture similar information. Therefore, the independent contribution of these models in explaining the brain representational geometries cannot be disentangled in the present study. The findings from this chapter complement results from Chapter 3 showing that the rFFA and rOFA could discriminate between face identities, and that voice-selective regions could discriminate between voice identities, by revealing that these regions may use information related to physical properties of the stimuli to distinguish between identities in their preferred modality. However, no correlations between face and voice representations in the rpSTS and models of face and voice information were found in the present chapter. Therefore, the informational content of face and voice representations in the rpSTS remains unclear. Finally, an analysis comparing representations across different regions showed that face- and voice-selective regions share face and voice information, respectively, with each other, and that face and voice information is also shared between face-selective, voice-selective, and multimodal brain regions within each modality. The next chapter will investigate how information extracted from the face relates to information extracted from the voice using behavioural judgements of the faces and voices of the same identities.

Chapter 5

The relationship between perceived information from faces and voices

Chapter 4 investigated the informational content of face and voice representations in different brain regions by comparing these representations to models of objective and perceived face and voice characteristics. The current chapter focuses on how perceived characteristics, namely trustworthiness, dominance, attractiveness, positive-negative valence, and perceived visual/auditory similarity, compare between the face and the voice of the same person on a behavioural level. Thus, this chapter addresses the third aim of this thesis, which was to determine how information extracted from a person’s face relates to the information extracted from their voice. The extent to which person-related information is consistent across faces and voices may influence the way that it is represented in the brain, i.e. separately for each modality or independently from modality. Although none of the brain regions tested in Chapter 4 showed representations of the same type of information from both faces and voices, it is possible that these representations may exist in brain regions not included in this analysis.

While there is some evidence showing the face and the voice of the same person convey concordant information about perceived physical characteristics, such as masculinity-femininity, health, and height (Smith et al., 2016a), it is not clear whether faces and voices also convey concordant information about social characteristics. Specifically, while studies have shown that faces and voices convey similar information regarding social traits such a trustworthiness, dominance, and attractiveness (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zuckerman & Driver, 1989), the majority of studies have focused on one modality, and less is known regarding the relationship between the social evaluation of a person’s face and the evaluation of their voice. Studies that have compared ratings of faces to ratings of voices of unfamiliar people have shown inconsistent results in regard to attractiveness and dominance. Correlations between face and voice attractiveness were often dependent on stimulus and/or participant gender, in that correlations were found only when rating opposite-sex stimuli (Lander, 2008), only for male participants rating female stimuli (Abend et al., 2015; Valentova et al., 2017; Wells et al., 2013), or only when considering male and female stimuli together, rather than separately (Rezlescu et al., 2015; Saxton et al., 2009). Moreover, one study found no correlation between facial and vocal attractiveness (Oguchi & Kikuchi, 1997). For dominance, one study showed a negative correlation between face and voice ratings (Rezlescu et al., 2015), while another showed a positive correlation (Han et al., 2017). Finally, one study showed a correlation between facial and vocal trustworthiness only when male and female stimuli were analysed together, but not when they were analysed separately (Rezlescu et al., 2015).

The inconsistencies in the findings of studies comparing ratings of faces and voices on social judgements may be due to these studies relying on ratings of a single token of the face and voice of each person. It has been shown that there is large variability in social judgements for different face images of the same person (Sutherland et al., 2017; Todorov & Porter, 2014), which may be influenced by cues extracted from changeable aspects of the face, such as emotional expression and face viewpoint (Sutherland et al., 2017). Given the variability in ratings of different images of the same face within modality, it is likely that variability in ratings of different face and voice tokens from the same person across modalities would be even greater.

Previous studies that compared ratings of faces and voices on physical (Smith et al., 2016a) or social information (Rezlescu et al., 2015) used unfamiliar faces and voices. Therefore, virtually nothing is known regarding the relationship between information perceived from the face and information perceived from the voice of familiar people. Moreover, the influence of familiarity on the relationship between information extracted from a person's face and information extracted from a person's voice has not yet, to the best of my knowledge, been investigated. For unfamiliar people, it is possible that the different nature of the cues extracted for social evaluation results in independent judgements of faces and voices. However, it is also possible that face and voice characteristics that tend to co-occur across different people are jointly associated with certain social traits through experience. This is particularly likely to be the case for social traits that are associated with physical characteristics that are conveyed consistently by the face and the voice, such as masculinity-femininity (Smith et al., 2016a). For familiar people, given that the faces and voices are naturally associated through experience (Yovel & Belin, 2013) and convey semantic information (Damjanovic & Hanley, 2007), it is likely that prior knowledge of a familiar person would result in similar judgements of their face and voice.

The present study compared ratings of faces and voices on trustworthiness, dominance, attractiveness, positive-negative valence, and perceived visual/auditory similarity for familiar people (Experiment 1 – presented in Chapter 4) and unfamiliar people (Experiment 2 – presented here). Unfamiliar identities in Experiment 2 were matched to the familiar identities in Experiment 1 in terms of demographic characteristics and facial and vocal appearance. The relationship between face and voice ratings in each experiment was compared between experiments to determine the influence of familiarity. The experimental paradigms used to test judgements of the faces and voices (presented in Chapter 4) attempted to address the issue of variability in the evaluation of different tokens of the face and voice of the same person by collecting ratings based on multiple different tokens of each face or voice, as opposed to a single face or voice stimulus. The aim of these paradigms was to encourage participants to make face and voice judgements by generalising across different tokens of the face and voice of each person, as opposed to making judgements based on a single token of the face or voice. Moreover, the different face videos and voice recordings of each identity, for both familiar and unfamiliar people, were highly variable in terms of changeable face and voice characteristics such as facial expression and vocal intonation, and were obtained from different original videos, with the aim to sample the variability of visual and auditory appearance encountered in everyday life (Burton, 2013; Lavan, Burton, et al., 2018).

It was predicted that judgements of the face and voice of the same person would be more similar for familiar people compared with unfamiliar people, due to prior knowledge of the person's character. An additional, exploratory analysis was performed, in which ratings of different judgements were compared with each other within modality to determine how any relationships between different judgements compare between faces and voices. Lastly, ratings of faces and voices on the social judgements were compared with perceived similarity, to determine to what extent different judgements are influenced by perceptual face and voice features.

5.1 Methods

Throughout this chapter, the experiment involving ratings of the faces and voices of familiar people, which was presented in Chapter 4, will be referred to as ‘Experiment 1’. Given that this experiment was described in detail in Chapter 4, this section focuses on the experiment involving ratings of the faces and voices of unfamiliar people, which will be referred to as ‘Experiment 2’.

5.1.1 Participants

Participants were recruited at Brunel University London. The study was approved by the Ethics Committee of Brunel University London (see Appendix B). 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 match the participant sample in Experiment 1. Thirty-six participants were recruited to take part in the study. Participants completed a Recognition Task (see below) to determine whether they recognised the identities whose faces and voices were presented in the experiment. Six participants were excluded because they recognised one or more of the 12 identities. The final sample consisted of 30 participants (six males) with mean age of 18.63 (SD=1, range=18-22). Participants provided written informed consent and received course credit for their participation.

Participants completed a face and voice Recognition Task to determine whether they recognised any of the 12 individuals whose faces and voices were included in the study. The task took place at the start of the experiment, immediately prior to the Familiarity task. Face and voice stimuli for this task were created using the same procedure that was described in Experiment 1, and were different from the stimuli used in main experiment. Stimuli were presented using Microsoft PowerPoint. For each stimulus participants were asked to verbally identify the person shown in the picture or the person speaking (by providing their name or other uniquely identifying biographical information).

5.1.2 Stimuli

Silent, non-speaking videos of moving faces, and recordings of voices belonging to 12 people (six female, six male) were obtained from videos on YouTube. These 12 people were selected to approximately match the 12 famous people whose faces and voices were presented in Experiment 1 on gender, age, race, general facial and vocal appearance, and accent. These individuals appear in multiple videos on YouTube but are not widely featured in British popular culture. The set of people included video bloggers, a video blogger guest, a creator of a make-up brand for women over sixties, a YouTube film actor, a businessman, a professional bodybuilder, and a comedian: Jessica Pettway, Suzie Bonaldi, Hollie Wakeham, Tricia Cusden, Daniel Trevenna, Robert F. Smith, Lawrence Brown, Dennis Wolf, Andrew Maxwell, John Adams, Crystal Conte, Amanda Deyes.

Six videos of the face of each person and six recordings of their voice were created using the same methods that were used to create the stimuli for Experiment 1 (for details please refer to Chapters 3 & 4). The duration of the face videos and the voice recordings was 1500ms. There were 72 face stimuli and 72 voice stimuli in total.

5.1.3 Procedure

The main experimental tasks took place over two days. On the first day, prior to the main experiment participants completed a Familiarity task in which they rated all face and voice stimuli on perceived familiarity. Then participants completed the trustworthiness, dominance, attractiveness, and positive-negative valence rating tasks. On the second day participants completed the pairwise visual and auditory similarity tasks.

In the Familiarity Task, Participants rated each individual face and voice stimulus on perceived familiarity. This task aimed to familiarise participants with the stimuli prior to the main experiment, and to confirm that participants were unfamiliar with the presented identities. The procedure for this task was the same as described in Chapter 4 for Experiment 1. The design and procedure for the trustworthiness, dominance, attractiveness, valence, and pairwise visual/auditory similarity tasks was the same as described in Chapter 4 for Experiment 1. The order of the tasks in was counterbalanced across participants.

5.2 Results

For all analyses, except for the within-subject reliability analysis, ratings were averaged across the two trials in which each identity/identity pair was presented. When averaging ratings across participants, comparing ratings across modalities, or computing between-subject reliability, ratings were first z-scored for each participant. Mean correlations were calculated by Fisher-transforming the single-subject correlations, averaging them, and then reverse-transforming the resulting values.

5.2.1 Analyses of differences in the stimuli across both experiments

First, we wanted to confirm that the familiar faces and voices in Experiment 1 were rated as more familiar than the unfamiliar faces and voices presented in Experiment 2. For each identity we averaged each participant's ratings of the six different tokens of the face, and the six different tokens of the voice, and then averaged the mean ratings for the face and the mean ratings for the voice across all participants. We compared average ratings of familiar face and voice identities with ratings of their corresponding matched unfamiliar identities using independent-sample t-tests. Familiarity ratings for familiar people (Exp.1) were significantly higher than for unfamiliar people (Exp.2) for their faces ($t(22) = 16.34$, $p<.0001$) and for their voices ($t(22) = 15.55$, $p<.0001$) (Figure 5.1), demonstrating that stimuli in Experiment 1 were indeed more familiar to participants than stimuli in Experiment 2.

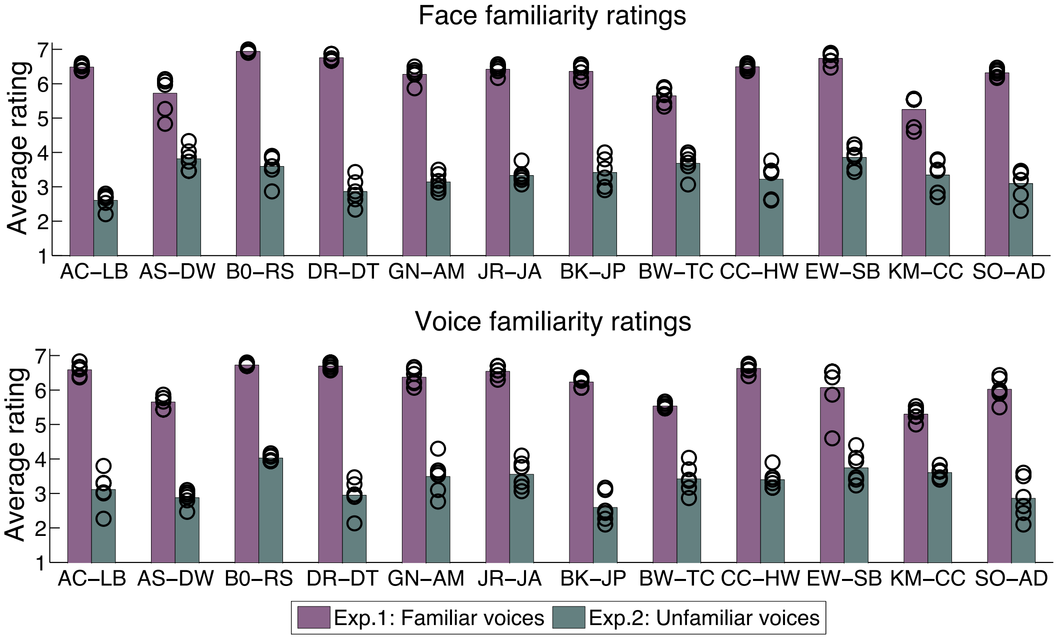


Figure 5.1: Familiarity ratings of the faces and voices of familiar people and matched unfamiliar people. Ratings were averaged across all 30 participants. Bars show ratings averaged across the six tokens of each face/voice, and circles show ratings of individual tokens.

Second, we investigated whether the unfamiliar identities that were presented in experiment 2 were well matched to the familiar identities presented in experiment 1, in terms of the perceived visual appearance of their faces, and the perceived auditory appearance of their voices. In each experiment, visual and auditory similarity ratings of faces and voices were averaged across all participants. We then compared the average visual and auditory similarity ratings of familiar identity pairs in experiment 1 with their corresponding matched unfamiliar identity pairs in experiment 2 using Spearman correlation. If two familiar faces in a pair both look similar to their matched pair of unfamiliar faces, the rating of perceived similarity of the faces within each pair should be similar between the familiar and the unfamiliar pair. The results showed high correlations between experiment 1 (familiar people) and experiment 2 (unfamiliar people) for ratings of face identities on visual similarity (*rho*=.72, *p*<.0001; Figure 5.2) and for ratings of voice identities on auditory similarity (*rho*=.76, *p*<.0001; Figure 5.2). Therefore, familiar and unfamiliar identities were well matched in terms of their perceived facial and vocal appearance.

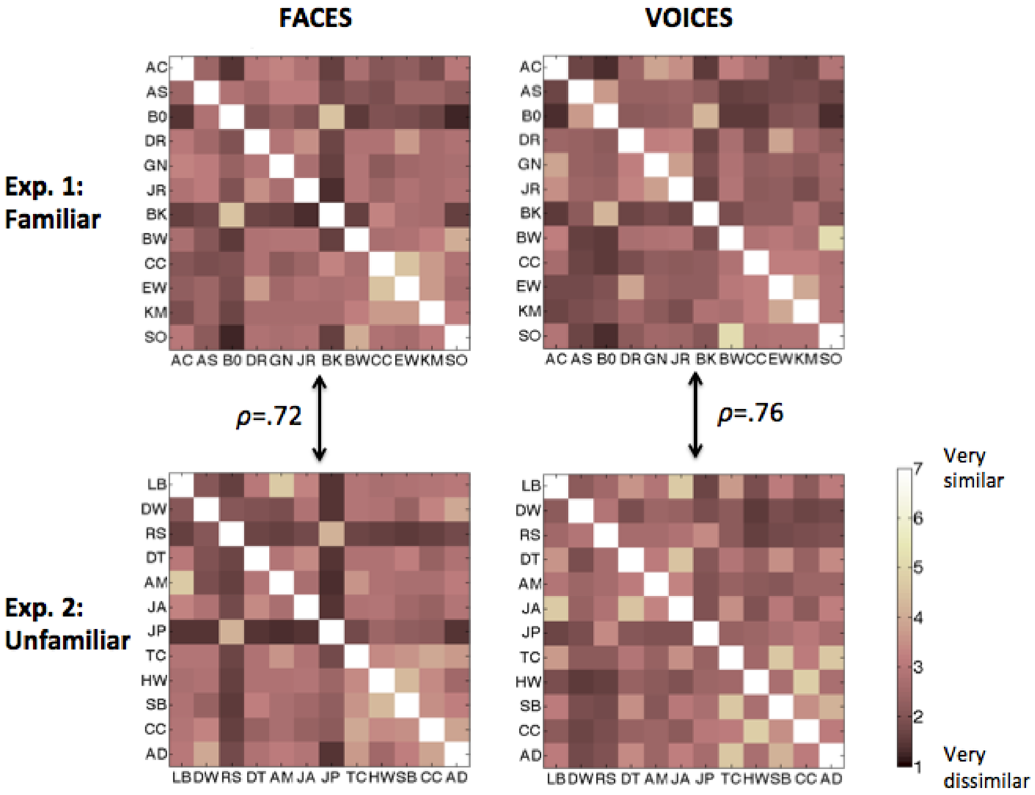


Figure 5.2: Representational similarity matrices showing average ratings of faces on visual similarity and ratings of voices on auditory similarity for familiar people and unfamiliar people. Ratings were averaged across all 30 participants. Matrices are symmetric around a diagonal of 7s (for illustration purposes – same-identity pairs were not rated in the task). Each cell represents the average similarity ratings between the identities in the corresponding row and column. Correlations between matrices were significant at *p*<.0001.

5.2.2 Comparison between ratings of faces and voices on social judgements

We compared ratings of faces and voices on trustworthiness, dominance, attractiveness, and positive-negative valence for familiar people (Experiment 1) and unfamiliar people (Experiment 2). For each participant and for each task, we calculated the Spearman correlations between the ratings of the faces and the ratings of the voices of the 12 identities (Figure 5.3). To test whether the single-subject correlations between face and voice ratings were significantly greater than zero, for each task we computed one-sample Wilcoxon sign-rank tests across all 30 participants (a non-parametric test was chosen due to correlation values not being normally distributed). The results showed that correlations between face and voice ratings were significantly greater than zero for all judgements, and for both familiar and unfamiliar people (Table 5.1).

Next, we tested whether familiarity influences the strength of the correlations between face and voice ratings, by comparing face-voice correlations across experiments, using Wilcoxon rank-sum tests. The results showed that for all tasks (trustworthiness, dominance, attractiveness, and valence) face-voice rating correlations were significantly higher for familiar people compared with unfamiliar people, confirming the hypothesis (Table 5.2). Finally, to visualise the relationship between the ratings of the face and voice of each identity, for each experiment we averaged ratings across participants and plotted the ratings of the faces against the ratings of the voices (Figure 5.4).

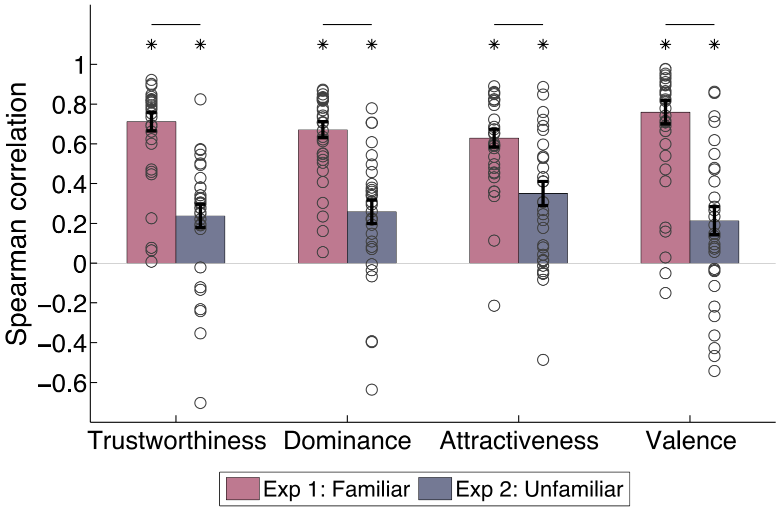


Figure 5.3: Correlations between face and voice ratings on trustworthiness, dominance, attractiveness, and positive-negative valence for familiar people and unfamiliar people. Bars show mean Spearman correlations across participants, error bars show standard error, and circles show individual participants. Stars show correlations that were significantly greater than zero at *p*≤.0256 (FDR corrected for all 8 comparisons). Horizontal lines above the bars show correlations that were

significantly different between familiar people and unfamiliar people at *p*≤.0009 (FDR corrected for all 4 comparisons).

Table 5.1: Average correlations between face and voice ratings on trustworthiness, dominance, attractiveness, and positive-negative valence, and one-sample Wilcoxon sign-rank test results for familiar people and unfamiliar people. Stars indicate correlations that were significantly greater than zero at *p*≤.0256 (FDR corrected for all 8 comparisons).

	Exp.1: Familiar				Exp.2: Unfamiliar			
	Mean rho	*SD*	*Z*	*p*	*Mean rho*	*SD*	*Z*	*p*
Trustworthiness	.71	.26	4.782	.0001*	.24	.32	3.301	.0010*
Dominance	.67	.22	4.782	.0001*	.26	.32	3.260	.0011*
Attractiveness	.63	.24	4.660	.0001*	.35	.33	3.774	.0002*
Valence	.76	.32	4.679	.0001*	.21	.39	2.232	.0256*

Table 5.2: Wilcoxon rank-sum test results for the comparison of the face-voice correlations between familiar people and unfamiliar people for each task. Stars indicate correlations that were significantly greater than zero at *p*≤.0009 (FDR corrected for all 4 comparisons).

Exp.1: Familiar VS. Exp.2 Unfamiliar		
	Z	*p*
Trustworthiness	4.827	.0001*
Dominance	4.753	.0001*
Attractiveness	3.328	.0009*
Valence	4.339	.0001*

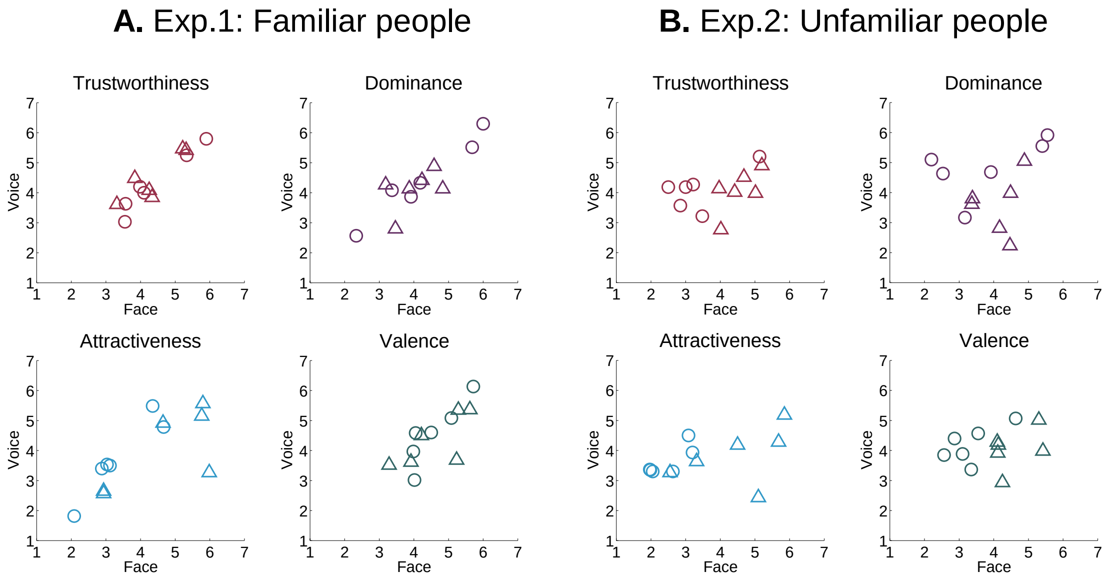


Figure 5.4: Relationships between the average ratings of the face and the voice of each of the 12 identities for familiar people and unfamiliar people. Ratings were averaged across all 30 participants. Triangles show female identities and circles show male identities. On the rating scale 1=very untrustworthy/non-dominant/unattractive/negative) and 7=very trustworthy/dominant/attractive/positive.

5.2.3 Comparison between ratings of faces and voices on pairwise perceptual similarity

We next compared ratings of face pairs on perceived visual similarity and ratings of voice pairs on perceived auditory similarity. For each participant and for each task, we calculated the Spearman correlations between the similarity ratings of the 66 face identity pairs and the 66 pairs (Figure 5.5). We used one-sample Wilcoxon sign-rank tests to test whether the single-subject correlations between face identity pair ratings and pair ratings in each experiment were significantly greater than zero. The results showed that correlations between face and voice ratings were significantly greater than zero for familiar people (mean *rho*=.42, *Z*=4.782, *p*<.0001) and for unfamiliar people (mean *rho*=.23, *Z*=4.432, *p*<.0001).

To test whether familiarity influences the strength of the correlations between visual similarity and auditory similarity ratings, we compared the correlations across experiments using Wilcoxon rank-sum tests. We found that correlations were significantly higher for familiar people compared with unfamiliar people (*Z*= 3.408, *p*=.0007). Finally, we visualised the relationship between average ratings (across participants) of visual similarity and ratings of the auditory similarity for each of the 66 identity pairs in experiments 1 and 2 (Figure 5.6).

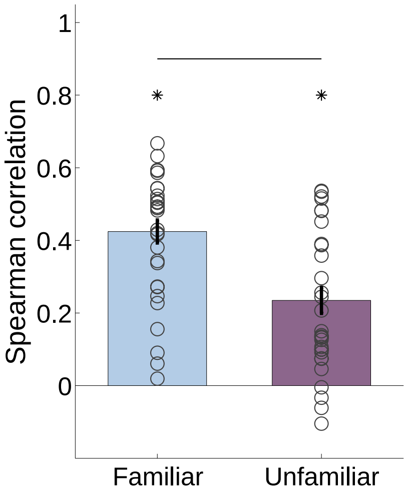


Figure 5.5: Correlations between ratings of faces on visual similarity and ratings of voices on auditory similarity for familiar people and unfamiliar people. Bars show mean Spearman correlations across participants, error bars show standard error, and circles show individual participants. Stars show that correlations were significantly greater than zero (both *p*<.0001). The horizontal line above the bars indicated that the correlations for familiar people were significantly higher than the correlations for unfamiliar people (*p*=.0007).

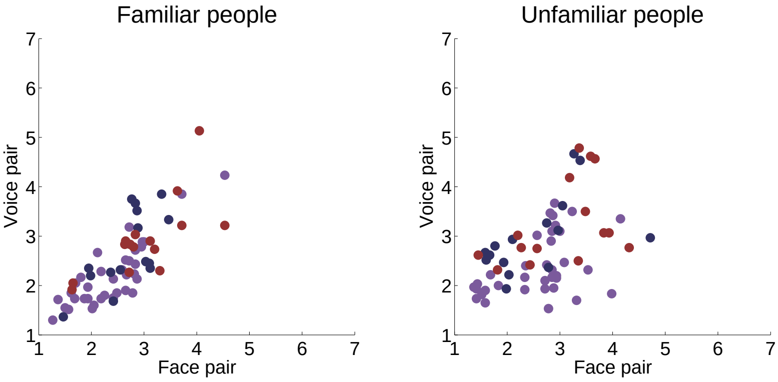


Figure 5.6: Relationships between the average visual similarity ratings of face pairs and the average auditory similarity ratings of voice pairs of each of the 66 identity pairs for familiar people and unfamiliar people. Ratings were averaged across all 30 participants. Female identity pairs are shown in red, male identity pairs are shown in blue, and male-female identity pairs are shown in purple. On the ratings scale 1=very dissimilar to 7=very similar.

5.2.4 Reliability of face and voice ratings

We next investigated the reliability of face and voice ratings both within-subjects and across-subjects, for familiar people (Experiment 1) and unfamiliar people (Experiment 2). Reliabilities were calculated for each task and each modality. This analysis is not only important to understand the consistency in ratings within and across participants, but also to provide upper bounds for the expected correlations we could find across faces and voices in the previous analyses.

To calculate the within-subject reliability, for each participant, we compared the ratings between the two trials in which each identity, or identity-pair, was presented. For the judgment tasks, the two trials featured different stimuli for each identity. For the similarity tasks, the two trials presented the same identity-pair, but did not necessarily present different stimuli. We computed the Spearman correlation between the ratings in the two trials, and the single-subject correlations were then averaged across all 30 participants. The results are shown in Table 5.3, and show high within-subject reliabilities for all judgements, especially for familiar stimuli, but also unfamiliar faces. The ratings of pairwise similarity were less consistent, and perhaps more dependent on the individual stimuli that were presented in each trial.

To calculate the between-subject reliability we calculated inter-rater agreements. For each participant, we calculated the Spearman correlation between their ratings and the average of the ratings of all other (29) participants. Finally, the single-subject correlations were averaged across all 30 participants. The results are shown in Table 5.4.

Table 5.3: Correlations of face and voice ratings between the two trials in which each identity was featured for familiar people and unfamiliar people. Values show Spearman correlations averaged across participants.

	Exp.1: Familiar		Exp.2: Unfamiliar	
	Faces	*Voices*	*Faces*	*Voices*
Trustworthiness	.79	.87	.81	.63
Dominance	.86	.79	.78	.67
Attractiveness	.93	.91	.92	.71
Valence	.85	.88	.76	.53
Pairwise similarity	.55	.53	.54	.42

Table 5.4: Inter-rater agreement of face and voice ratings for familiar people and unfamiliar people. Values show Spearman correlations averaged across participants.

	Exp.1: Familiar		Exp.2: Unfamiliar	
	Faces	*Voices*	*Faces*	*Voices*
Trustworthiness	.56	.57	.65	.44
Dominance	.70	.55	.67	.76
Attractiveness	.82	.74	.79	.39
Valence	.50	.59	.60	.44
Pairwise similarity	.61	.64	.60	.55

5.2.5 Relationship between ratings of different social judgements within modality

We next conducted a number of exploratory analyses to investigate the relationships between the ratings of different judgements within each modality (i.e. separately for faces and voices), for familiar people (Experiment 1) and unfamiliar people (Experiment 2), and how they differ between faces and voices. For each participant we compared the ratings of the 12 identities (in each modality) between all pairwise combinations of the four judgements (6 comparisons). Comparisons were made using Spearman correlation. At the group level, for each comparison between tasks, we tested whether the single-subject correlations were significantly greater than zero using one-sample Wilcoxon sign-rank tests across all 30 participants.

Figure 5.7 shows the correlations between each pair of judgements, averaged across participants. The results showed that, in both experiments, and for both faces and voices, correlations between different judgements were significantly greater than zero (FDR corrected *p*≤.0428), except for the correlations between ratings of familiar faces on dominance and each of the other judgements (Table 5.5).

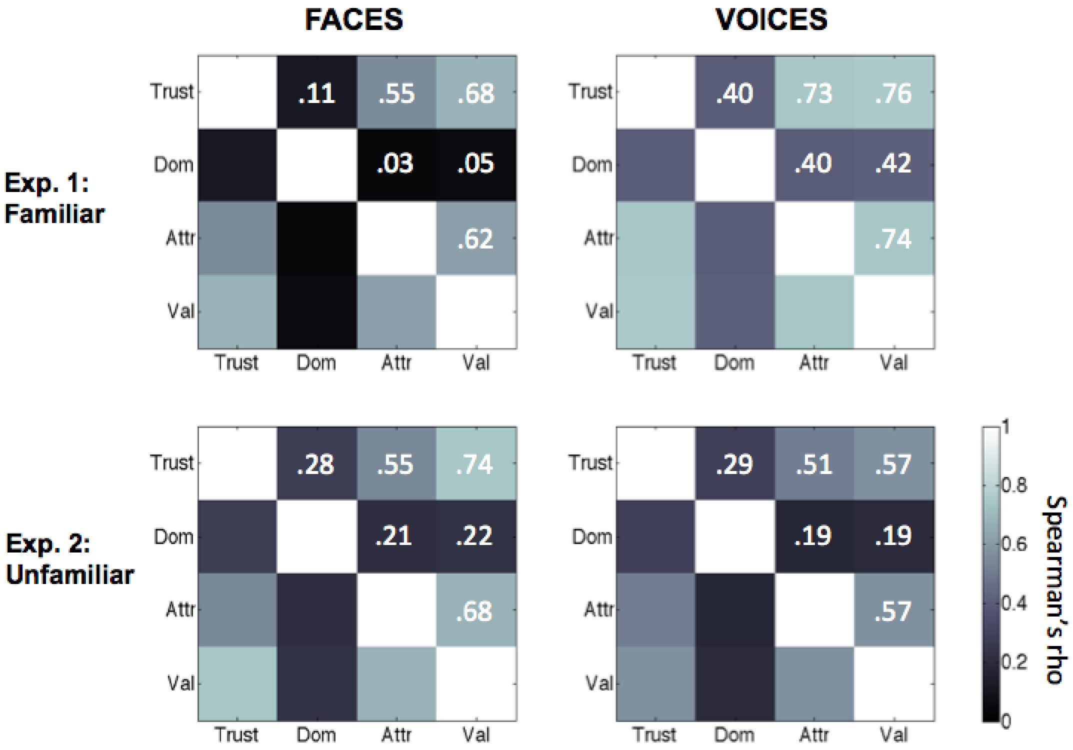
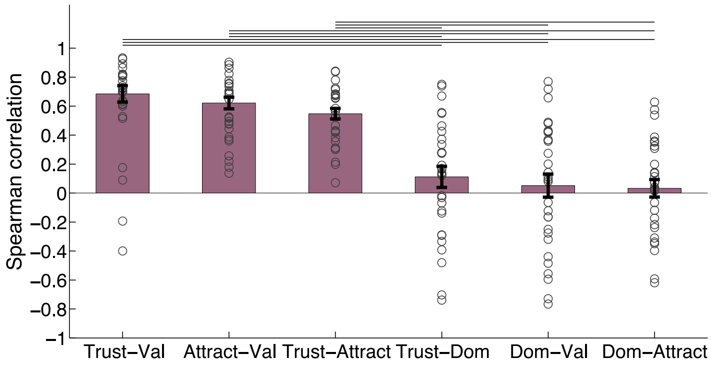


Figure 5.7: Average correlations between different judgements within each modality for familiar people and unfamiliar people. Matrices are symmetric around a diagonal of 1s. Each cell shows the Spearman correlation (averaged across all 30 participants) between the judgements in the corresponding row and column.

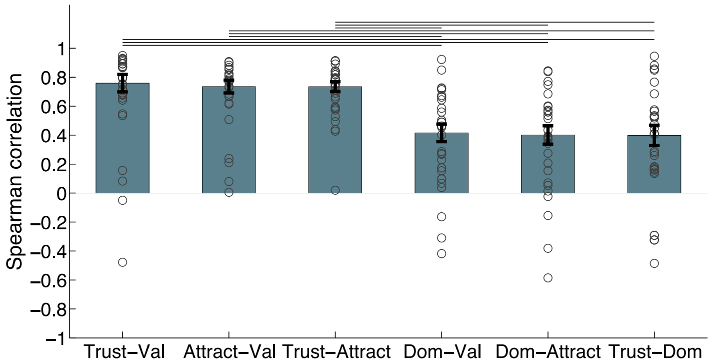
Table 5.5: Wilcoxon sign-rank test results of analysis comparing the correlations between different judgements within-modality against zero, for familiar people and matched unfamiliar people. Stars indicate correlations that were significantly greater than zero at *p*≤.0428 (FDR corrected for all 24 comparisons).

	Exp.1: Familiar people				Exp.1: Unfamiliar people			
	Faces		Voices		Faces		Voices	
	Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
Trust-Dom	1.409	.1589	3.363	.0008*	2.581	.0098*	3.013	.0026*
Trust-Attr	4.703	.0001*	4.782	.0001*	4.494	.0001*	4.350	.0001*
Trust-Val	4.638	.0001*	4.679	.0001*	4.741	.0001*	4.700	.0001*
Dom-Attr	0.596	.5521	3.898	.0001*	2.499	.0125*	2.293	.0218*
Dom-Val	0.586	.5577	4.083	.0001*	2.026	.0428*	2.293	.0218*
Attr-Val	4.703	.0001*	4.782	.0001*	4.782	.0001*	4.165	.0001*

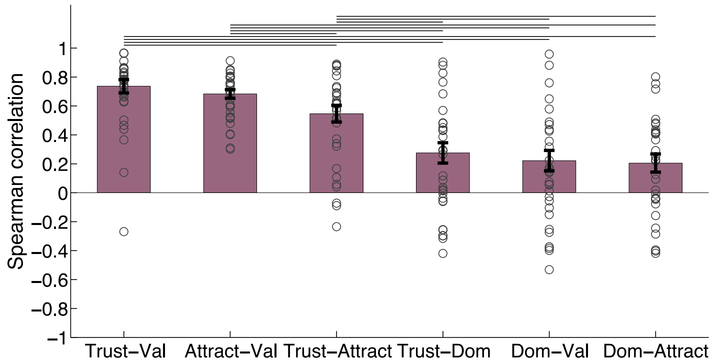
Figure 5.7 shows a tendency for trustworthiness, attractiveness, and valence ratings to be more correlated with each other than with dominance, for both faces and voices, and for both familiar and unfamiliar people. To investigate this further, we tested whether correlations between different pairs of judgements were significantly different from each other using Wilcoxon sign-rank tests. Figure 5.8 shows the correlations for each experiment and modality, sorted from highest to lowest, and indicates which correlations were significantly different. For both familiar and unfamiliar people, and for both modalities, the three pairwise correlations between dominance and the other judgments were significantly lower than the three pairwise correlations between trustworthiness, attractiveness, and valence (with the exception of the comparison between attractiveness-valence and trustworthiness-dominance in unfamiliar voices) at *p*≤.0207 (FDR corrected for all 60 comparisons). Furthermore, the ranking of the correlations is highly consistent across experiments and modalities, with the correlation between trustworthiness and valence being the highest, and the correlations between dominance and attractiveness/trustworthiness being the lowest.



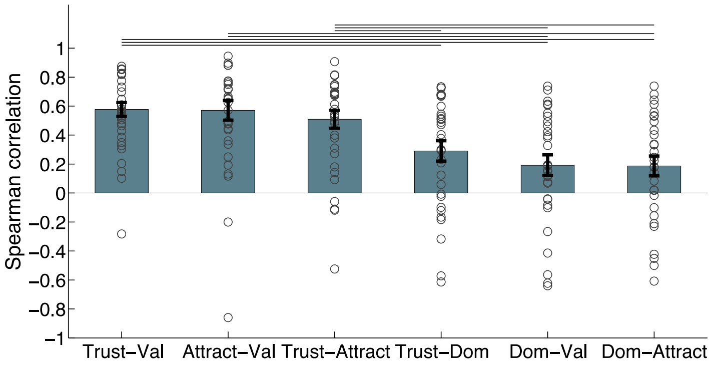
Exp. 1: Familiar Faces



Exp. 1: Familiar Voices



Exp. 2: Unfamiliar Faces



Exp. 2: Unfamiliar Voices

Figure 5.8: Pairwise correlations between ratings of trustworthiness, dominance, attractiveness, and positive-negative valence from faces and voices, in familiar people and unfamiliar people, sorted from highest to lowest. Bars show mean Spearman correlations across participants, error bars show standard error, and circles show individual participants. Correlations are presented sorted from highest to lowest in each modality. Horizontal lines above the bars show correlations that were significantly different from each other at *p*≤.0.0207 (FDR corrected for all 60 comparisons).

5.2.6 Comparison between ratings of faces and voices on pairwise perceptual similarity and their social evaluation

We also investigated the relationship between ratings of faces/voices on pairwise visual/auditory similarity, and the ratings of faces/voices on trustworthiness, dominance, attractiveness, and valence. In each experiment, for each participant and for each task, we calculated the Euclidean distance between their ratings of each possible pair of identities, separately for faces and voices. Thus, for each judgement we obtained a vector of rating distances between 66 identity pairs, which were compared to perceptual similarity ratings of the same 66 identity pairs using Spearman correlation, separately for faces and voices.

For each judgement, in each modality, we then tested whether the correlation with perceptual similarity was significantly greater than zero using Wilcoxon sign-rank tests across all 30 participants. The results showed low, but significantly greater than zero, correlations between all four judgements and perceptual similarity for both faces and voices, and for both familiar and unfamiliar people (with the exception of valence for unfamiliar voices; Figure 5.9, Table 5.6).

Finally, we tested whether some judgements were more correlated with perceptual similarity than others using Wilcoxon sign-rank tests, separately for faces and voices and for familiar and unfamiliar people. Attractiveness in familiar faces was significantly more correlated with perceptual similarity compared with trustworthiness

(*Z*=3.060, *p*=.0022) and dominance (*Z*=2.714, *p*=.0067). Valence in unfamiliar voices was significantly less correlated with perceptual similarity compared with dominance (*Z*=-2.705, *p*=.0068) and attractiveness (*Z*=-2.746, *p*=.0060).

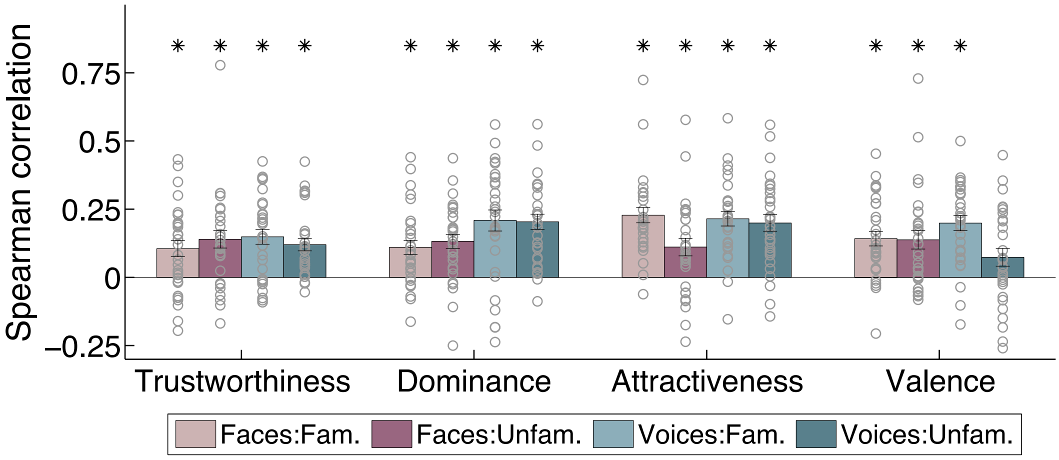


Figure 5.9: Correlations between ratings of face or voice identity pairs on perceptual similarity and the Euclidean distances between their ratings on trustworthiness, dominance, attractiveness, and valence, for familiar people and unfamiliar people. The direction of the correlations is reversed for ease of interpretation. Bars show mean Spearman correlations across participants, error bars show standard error, and circles show individual participants. Stars show significant tests at *p*<.0039 (FDR corrected for all 16 comparisons).

Table 5.6: Average correlations between ratings of face or voice identity pairs on perceptual similarity and the Euclidean distances between their ratings on trustworthiness, dominance, attractiveness, and valence, and one-sample Wilcoxon sign-rank test results for familiar people and matched unfamiliar people. The direction of the correlations is reversed for ease of interpretation. Stars indicate correlations that were significantly greater than zero at *p*<.0039 (FDR corrected for all 16 comparisons).

Faces

	Exp.1: Familiar				Exp.2: Unfamiliar			
	Mean rho	*SD*	*Z*	*p*	*Mean rho*	*SD*	*Z*	*p*
Trustworthiness	.11	.16	2.890	.0039*	.14	.18	3.692	.0002*
Dominance	.11	.14	3.548	.0004*	.13	.14	3.815	.0001*
Attractiveness	.23	.15	4.638	.0001*	.11	.18	2.972	.0030*
Valence	.14	.15	4.062	.0001*	.14	.18	3.260	.0011*

Voices

	Exp.1: Familiar				Exp.2: Unfamiliar			
	Mean rho	*SD*	*Z*	*p*	*Mean rho*	*SD*	*Z*	*p*
Trustworthiness	.15	.15	3.857	.0001*	.12	.12	4.309	.0001*
Dominance	.21	.21	3.836	.0001*	.20	.15	4.597	.0001*
Attractiveness	.22	.15	4.556	.0001*	.20	.17	4.289	.0001*
Valence	.20	.15	4.371	.0001*	.07	.18	1.923	.0545

5.3 Discussion

The study in this chapter showed that judgements of trustworthiness, dominance, attractiveness, positive-negative valence, and perceived visual/auditory similarity were more similar between faces and voices for familiar people compared with unfamiliar people. This finding confirms the hypothesis that the relationship between judgements of the face and voice would be stronger for familiar people, compared with unfamiliar people, due to prior knowledge of a person's character. However, despite correlations across modalities being lower for unfamiliar people, significant correlations were observed for all judgements. This suggests that prior knowledge of a person may not be the only reason for similar face-voice evaluation.

This chapter also showed that, within modality, the correlations between the different judgements were similar for both faces and voices, and for both familiar and unfamiliar people. Specifically, the results showed moderate to high correlations between ratings of trustworthiness, attractiveness, and valence, but low or non-significant correlations between dominance and the other judgements, suggesting that these judgements are conceptualised in a similar way regardless of modality and familiarity level. Finally, this chapter showed that the perceptual similarity of faces and voices explained only a small amount of the variance in social judgements for both familiar and unfamiliar people.

Similar social evaluation of faces and voices, for familiar and unfamiliar people

For familiar people, the highly similar evaluation of faces and voices is likely to be due to the prior knowledge of the person having a major influence on the evaluation of their face and voice. This semantic knowledge most likely overrides the influence of perceptual cues in faces and voices, which have been shown to influence face and voice judgements in unfamiliar people (McAleer et al., 2014; Sutherland et al., 2017; Todorov & Oosterhof, 2011). However, similar evaluations of faces and their corresponding voices were also observed for unfamiliar people (although correlations were much lower compared to familiar people), suggesting that semantic knowledge may not be the only factor that drives similar evaluations in faces and voices.

The finding of similar evaluations of faces and voices of unfamiliar people is in agreement with studies showing low to moderate correlations between ratings of faces and voices in terms of attractiveness (Abend et al., 2015; Lander, 2008; Rezlescu et al., 2015; Saxton et al., 2009; Valentova et al., 2017; Wells et al., 2013), dominance (Han et al., 2017), and trustworthiness (Rezlescu et al., 2015). Some studies have shown that ratings of faces and voices on attractiveness and trustworthiness were no longer associated when male and female stimuli were analysed separately (Rezlescu et al., 2015; Saxton et al., 2009). However, given the relatively small number of identities presented in the current experiments (six male and six female), it was not possible to conduct separate analyses for male and female identities.

An important difference between the present study and previous studies is the use of an experimental paradigm that aimed to encourage participants to base their judgements on multiple, naturalistically varying tokens of the face/voice of each identity. Correlations between ratings of trials featuring different tokens of the same identity were high across participants, for both faces and voices, and for both familiar and unfamiliar people (although correlations were slightly lower for unfamiliar voices). This suggests that this paradigm encouraged participants to base their ratings on cues that are largely consistent across different images of the face, or recordings of the voice, of the same person, rather than on changeable aspects of the face and voice that are specific to an image or recording.

It has been suggested that people form consistent associations between faces and social traits that are driven by similar cultural experiences and exposure to stereotypes (Freeman & Johnson, 2016; Over & Cook, 2018; Stolier et al., 2018). However, there is evidence that the personal attitudes and experiences of each individual also play a part in shaping judgements of social traits in faces (Hehman, Sutherland, Flake, & Slepian, 2017; Kramer, Mileva, & Ritchie, 2018; Stanley, Sokol-Hessner, Banaji, & Phelps, 2011; Watkins, Jones, & DeBruine, 2010). Unlike the majority of previous studies examining social judgements of faces and voices, in the present study all analyses were performed at the level of each individual participant, rather than at the stimulus level. Specifically, correlations across ratings of all faces and ratings of all voices were computed for each participant, as opposed to computing correlations between each individual face and voice across all participants. Thus, this approach takes into account individual differences across participants in their ratings of the face and the voice of each identity. The lack of such an approach in previous studies comparing rating of faces and voices may have contributed to their inconsistent results

A possible explanation for the finding of similar evaluations of the faces and voices of unfamiliar people is the formation of joint associations between specific facial and vocal features and certain social traits that appear to co-occur across different people (Over & Cook, 2018). Specifically, it is possible that multisensory representations of social traits that include information about what a person looks like and sounds like are formed through experience. This is particularly likely to be the case for social judgements that are associated with physical characteristics that are conveyed by both the face and the voice, such as masculinity and health (Smith et al., 2016). For example, dominance has been associated with masculinity (McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013), and attractiveness has been associated with measures of health (Gangestad et al., 1994; Hughes et al., 2002) in both faces and voices. Thus, if people that are perceived as dominant typically have masculine-looking faces and masculine-sounding voices, then facial and vocal cues that signal masculinity will be associated with dominance, and a multisensory representation of dominance may be formed. Therefore, if a newly encountered unfamiliar person conforms to this stored representation of dominance, both their face and their voice will be judged as dominant even if the face and voice are judged independently.

Similar relationship between social judgements for faces and voices in familiar and unfamiliar people

A comparison of the ratings of faces and voices on trustworthiness, dominance, attractiveness, and positive-negative valence within each modality showed that the relationships between the different judgements were strikingly similar for both faces and voices, in familiar and unfamiliar people. Specifically, while trustworthiness, attractiveness, and valence were well correlated with each other, correlations between dominance and the other three judgements were significantly lower, for both faces and voices, and for both familiar and unfamiliar people. These findings are in line with those of previous studies that have used dimensionality-reduction techniques for ratings of unfamiliar faces and voices on multiple judgements, and have shown that valence/trustworthiness and dominance form orthogonal dimensions for faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and for voices (McAleer et al., 2014). Previous studies have also shown a relationship between valence/trustworthiness and attractiveness in faces (Oosterhof & Todorov, 2008; Rezlescu et al., 2015), although in one study this was only found for male faces (Rezlescu et al., 2015). In voices, the finding is in agreement with Rezlescu et al. (2015), who found a high correlation between attractiveness and trustworthiness. McAleer et al. (2014) also showed that attractiveness was strongly correlated with trustworthiness, but only in female voices.

The finding of similar relationships between the different judgements across faces and voices suggests that these judgements may be conceptualised in a similar way for faces and voices. It has previously been shown that concepts for different emotions are also similar across faces and voices (Kuhn et al., 2017), suggesting that representations of high-level information that is conveyed by both faces and voices are modality-general. Furthermore, the present study showed that the relationships between different judgements are not influenced by familiarity, and contradict the idea that trait space varies based on familiarity (Stolier et al., 2018). For example, a person perceived as attractive will also tend to be perceived as trustworthy and likeable, regardless of whether the person is familiar or not, and regardless of whether judgements are based on the face or the voice.

Perceptual similarity ratings and relationship with social judgements

Significant correlations were found between ratings of the perceptual similarity of faces and voices, for both familiar and unfamiliar people. In line with the social judgements, these correlations were significantly higher for familiar people, compared with unfamiliar people. For familiar people, these correlations could be due to participants being influenced by their knowledge of the people's character, despite being instructed to base their judgements solely on the similarity of the physical appearance of the face and voice. For unfamiliar people, the correlations between faces and voices may be explained by redundant information in faces and voices about certain physical characteristics, such as age, gender, race, nationality, and masculinity/femininity. This would be in line with the findings of Smith et al. (2016) showing concordant cues in faces and voices for masculinity/femininity and age.

Finally, this study showed that although perceptual similarity of faces and voices was significantly correlated with the Euclidean distance between ratings of faces and voices on social judgements (with the exception of valence in unfamiliar voices), all correlations were very low. This suggests that perceived visual/auditory similarity explains a small amount of the variance in social judgements of faces and voices. However, it is possible that social judgements are influenced by more subtle variations in facial or vocal features that were not used as cues for judging the similarity between two different faces/voices in the perceptual similarity task.

Limitations and future directions

One potential limitation of the current study concerns the large variability across the different identities due to differences in gender, age, accent, race, and nationality. For the unfamiliar identities, which were matched to the familiar identities in regard to these properties, this variability may have lead participants to realise that the faces and voices belonged to the same people (this was not mentioned in the instructions). In this case, participants may have attempted to predict which face corresponded to which voice, and as a consequence may have given similar ratings to faces and voices that they thought corresponded to the same people. This may, at least in part, account the observed similarity in the ratings of the faces and voices of unfamiliar people. Future studies should attempt to replicate the current findings using a more homogenous set of unfamiliar identities.

A further potential limitation concerns the use of famous-familiar identities, as opposed to personally-familiar identities. Although participants reported being highly familiar with the famous people from both their faces and voices, it is unlikely that participants ever interacted with these people personally in a social context. Thus, any knowledge of their personality will have been obtained from third-parties, and most likely through the media. This level of familiarity with a person may not be sufficient to challenge potential stereotypes regarding the relationships between different social traits, e.g. that an attractive person is also trustworthy (Stolier et al., 2018), and may have contributed to the finding of similar relationships between the different social judgements for familiar and unfamiliar people. It is possible that personal familiarity with an individual would have stronger influence on the relationships between the different judgements (Stolier et al., 2018). For example, a personally familiar individual could be perceived as attractive but also as untrustworthy (e.g. in a romantic relationship situation). In future, it would be interesting to investigate the evaluation of people who are perceived as attractive but have been experimentally associated with behaviour that indicates untrustworthiness.

Finally, although an analysis comparing perceived visual and auditory similarity of faces and voices, respectively, between familiar and unfamiliar people showed that these people were well matched based on physical characteristics, the possibility that variables other than familiarity could have influenced the differences between familiar and unfamiliar people cannot be excluded. Therefore, this chapter cannot conclusively claim that gaining familiarity with a person through experience will result in the judgements of their face and voice becoming more similar. To overcome this limitation, future work could experimentally familiarise participants with the faces and voices of a set of identities, and investigate the evaluation of these faces and voices before and after familiarisation.

Conclusion

This chapter showed that judgements of perceived trustworthiness, dominance, attractiveness, positive-negative valence, and perceived visual/auditory similarity were more similar between faces and voices for familiar people compared with unfamiliar people. This was attributed to the influence of prior semantic and biographical knowledge of the person, and suggests that this information is represented largely independently from modality for familiar people. Therefore, although the previous chapter failed to identify representations of social information from familiar faces and voices in multimodal person-selective brain regions, it is possible that this information is represented independently from modality in a brain region that was not included as a ROI in this study, and future studies should further investigate this possibility. For unfamiliar people, although correlations were lower, similarities were also observed between judgements of the face and the voice. It was speculated that these similarities may be due to the learning of joint associations between frequently co-occurring face and voice features, on the one hand, and social traits, on the other hand. Finally, this chapter showed that the perceived visual similarity of faces and the perceived auditory similarity of voices explained very little variance in the judgements of social traits.

Chapter 6

General Discussion

6.1 Summary of main findings

Faces and voices both serve as sources of person identification, and they convey a wealth of information about a person, such as their gender, age, and whether they can be trusted (McAleer et al., 2014; Oosterhof & Todorov, 2008; Yovel & Belin, 2013). The first main aim of this thesis was to determine how the brain integrates information from the faces and voices of familiar people to represent person identity. This aim was addressed in the study described in Chapter 3, which used RSA to compare multivoxel activity patterns in response to the faces and voices of familiar people in face-selective, voice-selective, and multimodal regions. This chapter identified a multimodal region in the pSTS that integrates information from faces and voices in a crossmodal representation of person identity, providing support for the Multimodal Processing (MP) model of face and voice integration. Furthermore, stimulus-invariant representations of face identity and voice identity were identified in face-selective and voice-selective regions, respectively.

The second main aim of the thesis was to determine the informational content of face and voice representations in face-selective, voice-selective, and multimodal brain regions. To address this aim, the study described in Chapter 4 used RSA to compare brain representations to models of face and voice properties, and showed that brain representations in face- and voice-selective regions are associated with information about both the perceived and objective physical similarity between individual faces and voices. These findings suggest that face- and voice-selective regions primarily process visual face properties and auditory voice properties, respectively. No evidence was found of representations of social properties from faces and voices in multimodal brain regions.

The final main aim of this thesis was to determine how perceived information extracted from a person’s face relates to the information extracted from their voice, and how this relationship is influenced by familiarity. Therefore, the study in Chapter 5 compared ratings of faces and voices on social judgements, namely trustworthiness, dominance, attractiveness, and positive-negative valence, and ratings on perceived visual/auditory similarity, for familiar and unfamiliar people. This chapter showed that similar information is extracted from a person’s face and voice when the person is familiar and to a lesser extent when the person is unfamiliar. This suggests that having prior semantic knowledge about a person leads to similar judgements of their face and voice. Moreover, it suggests that some concordant information relating to social judgements may be available even in the faces and voices of unfamiliar people.

6.2 The integration of face and voice information in the brain

Previous work had largely investigated face and voice identity representations separately, and a consequence little is known regarding the integration of face and voice information in the brain to form representations of person identity that are independent from modality. As explained in the Introduction, two theoretical models of face and voice integration have been put forward (Blank et al., 2011; Campanella & Belin, 2007; Yovel & O’Toole, 2016). The Multimodal Processing (MP) model proposes that face and voice information is integrated in multimodal brain regions (e.g. A. W. Ellis et al., 1989; Shah et al., 2001), whereas the Coupling of Face and Voice Processing (CFVP) model proposes that face and voice information is also integrated though direct coupling between face- and voice-responsive regions (e.g. Blank et al., 2011; von Kriegstein et al., 2005). These two models are not mutually exclusive, and Chapter 3 tested the predictions of both models by using fMRI and RSA to compare multivoxel activity patterns in response to the faces and voices of the same identities in face-selective, voice-selective, and multimodal (both face-selective and voice-selective) brain regions. Based on the MP model, the hypothesis was that crossmodal person identity representations would be found in multimodal regions, whereas based on the CFVP, the prediction was that crossmodal person identity representations would be found in face-selective or voice-selective brain regions.

The results from the study described in Chapter 3 revealed a representation of person identity in the multimodal rpSTS, providing support for the MP model. Specifically, this study demonstrated that a region of the pSTS that selectively responded to both faces and voices could discriminate between different identities based on crossmodal information in the face and voice. Importantly, the rpSTS not only contained representations of person identity, but also showed stimulus-invariant representations of face and voice identity that were robust to within-person variability across different, naturalistically varying face videos and voice recordings of the same person.

The role of the pSTS may be equivalent to the person identity node (PIN) in cognitive models of face and voice recognition, in that it receives and integrates identity information from face recognition units (FRUs) and voice recognition units (VRUs) (Belin et al., 2004; Bruce & Young, 1986; Burton et al., 1990; Campanella & Belin, 2007). Chapter 3 showed that the rFFA and rOFA and the voice-selective regions in the temporal lobes distinguish between identities in their respective modality, but not across modalities, in line with the proposed role of the FRUs and VRUs in cognitive models (Belin et al., 2004; Bruce & Young, 1986; Burton et al., 1990; Campanella & Belin, 2007). Furthermore, Chapter 4 showed “representational connectivity”, i.e. similar representational geometries, between face representations in the pSTS and the FFA and OFA, and between voice representations in the pSTS and voice-selective regions, suggesting that these regions share informational content. Therefore, one possibility is that the pSTS receives and integrates information from face-selective regions, such as the FFA and OFA, and voice-selective regions, such as the TVAs. Previous studies using effective connectivity analyses have shown feed-forward connections between the FFA and TVA, on the one hand, and the multimodal pSTS, on the other hand (Davies-Thompson et al., 2018), and between the OFA and pSTS (Fairhall & Ishai, 2007). Moreover, one study showed evidence of functional connectivity between the FFA and the pSTS (Turk-Browne et al., 2010). However, other studies investigating the connectivity between the pSTS, on the one hand, and the FFA and OFA, on the other hand, have shown limited structural connectivity (Blank et al., 2011; Gschwind et al., 2012; Pyles et al., 2013) and functional connectivity (Davies-Thompson & Andrews, 2012; O'Neil et al., 2014) between these regions. Future work should focus on identifying possible functional and structural pathways in the brain through which visual and auditory information may be conveyed to the pSTS.

A second possibility is that face and voice information is both processed and integrated within the pSTS. Previous work has shown that the pSTS contains distinct cortical patches that respond preferentially to faces, voices, and audiovisual face-voice stimuli (Beauchamp et al., 2004). Moreover, it has been shown that the part of the pSTS that responds preferentially to audiovisual face-voice stimuli is located in between regions of the STS that are primarily face-selective (in the posterior pSTS) and primarily voice-selective (towards the mid STS) (Kreifelts et al., 2009). Based on these findings, it is possible that the pSTS contains sub-regions that are face-selective only, voice-selective only, or both face- and voice-selective (i.e. multimodal). Relating back to cognitive models of face and voice recognition, the pSTS may initially process face and voice information independently in face- and voice-selective sub-regions, corresponding to the FRUs and VRUs, and subsequently integrate that information in person-selective multimodal sub-regions, corresponding to the PIN. This mechanism does not require information to be conveyed to the pSTS from other brain regions. Duchaine & Yovel (2015) proposed that the pSTS may receive face input directly from early visual cortex, which would be consistent with this view. Future studies should use high-resolution brain imaging methods to investigate modality-specific and modality-general representations in sub-regions of the pSTS.

The involvement of the rpSTS in integrating face and voice information in a person identity representation has been suggested by two previous studies (Anzellotti & Caramazza, 2017; Hölig et al., 2017). Hölig et al. (2017) compared responses to voices that had been primed by faces of the same identity or faces of a different identity, and showed that the pSTS responded more to voices primed by identity-incongruent faces. Hölig et al. (2017) proposed that viewing a familiar person's face automatically activates a representation of the person's voice in the pSTS, and creates a prediction regarding the identity of the voice. The increased activation of the pSTS to incongurent face-voice pairings is thus attributed to a violation of expectations regarding the identity of the subsequently presented voice. The main limitation of this study was that, given that the authors trained participants to associate unfamiliar faces with their corresponding voices, rather than presenting identities that were familiar to participants, the pSTS may have been responding to the incongruence due to the violation of expectations derived from learned associations between faces and voices that could be entirely arbitatry and unrelated to person identity processing. However, the finding of a crossmodal person identity representation in the pSTS in Chapter 3 suggests that the pSTS may indeed have the capacity to engage identity representations from one modality after exposure to the other modality.

A crossmodal representation of person identity in the rpSTS was also identified by a previous MVPA study using crossmodal classification (Anzellotti & Caramazza, 2017). This study showed that the pSTS could discriminate between response patterns to a pair of familiar face identities based on the response patterns to their corresponding voices, and vice versa. Moreover, the authors demonstrated that this pSTS region responded selectively to both faces and voices, similar to the pSTS region described in Chapter 3. However, in contrast to the study presented here, in their study Anzellotti & Caramazza (2017) showed limited evidence that representations of faces and voices in the pSTS generalise to novel tokens of the face and voice of the same person. Moreover, the authors presented only two tokens of the face and voice of each of two identities, which were also constrained in terms of their natural variability. Therefore, the findings of the present thesis substantially extend the evidence of a person identity representation in the pSTS by showing that this region can distinguish between a much larger set of identities based on multiple, naturalistically varying tokens of their face and voice.

The current thesis did not show any evidence of representations of person identity in the ATL. The involvement of the ATL in representing person identity independently from modality has been proposed mainly by studies involving patients with lesions to this region who showed impairments in both face and voice recognition (A. W. Ellis et al., 1989; Gainotti, 2011). However, a meta-analysis of neuroimaging studies in healthy participants that tested responses to face and voice familiarity and recognition showed that the ATL was consistently involved in the processing of familiarity, but not recognition (Blank et al., 2014). Therefore, it is possible that the ATL is involved in the processing of person familiarity, rather than person identity. This notion is supported by a meta-analysis of studies involving patients with ATL lesions, and which showed that patients who showed deficits in face and voice recognition also showed impairments in face familiarity tasks (Gainotti, 2011). Thus, it is likely that intact familiarity processing is a necessary pre-requisite for person identity recognition, and that the ATL primarily engages in familiarity processing. Although multiple studies have shown evidence of representations of face identity in the ATL (Anzellotti & Caramazza, 2016; Anzellotti et al., 2014; Collins et al., 2016; Guntupalli et al., 2017; Verosky et al., 2013), it is notable that none of the fMRI studies investigating representations of person identity, which were reviewed in the Introduction chapter, found evidence of such representations in the ATL (Anzellotti & Caramazza, 2017; Hasan et al., 2016; Hölig et al., 2017; Joassin et al., 2011; Shah et al., 2001). Therefore, another possibility is that the ATL may contain representations of face identity, but not of voice identity or person identity. However, it is also possible that identity representations in the ATL were not identified in the present study and in previous fMRI studies due to the low signal-to-noise ratio that has been observed in this region when using standard fMRI sequences (Axelrod & Yovel, 2013). Future studies aiming to investigate person identity representations should attempt to optimise their scanning parameters to minimise the signal-to-noise ratio in the ATL, for example by using coronal slice orientation, as opposed to the standard axial slice orientation (Axelrod & Yovel, 2013).

An exploratory whole-brain searchlight analysis conducted in Chapter 3 with the aim of identifying brain regions outside the selected ROIs that may contain crossmodal representations of person identity revealed, among other regions, a cluster in the left hippocampus. Neurophysiological studies that conducted single-cell recordings in human brains have shown that the bilateral hippocampus contains multisensory neurons that respond to both the face and the name of a person (Quiroga et al., 2009, 2005). However, responses to voices were not tested in these studies. The current thesis suggests that the left hippocampus may contain representations of person identity that integrate information from faces and voices. An fMRI study previously showed that the right hippocampus responded more to newly-learned faces and voices that were presented simultaneously, than to the same faces and voices presented in isolation (Joassin et al., 2011). Although only the left hippocampus showed significant results in searchlight analysis presented in Chapter 3, it is possible that both the right and left hippocampus play a role in integrating face and voice information. Future studies investigating person identity representations should therefore include ROIs for both the right and left hippocampus to further probe person identity representations in this region.

Previous studies using fMRI have shown that the retrosplenial cortex (Shah et al., 2001) and the angular gyrus (Hölig et al., 2017; Joassin et al., 2011) may contain person identity representations. The multimodal precuneus ROI that was used in the study presented in Chapter 3 included the retrosplenial cortex, but did not show representations of person identity, despite being able to distinguish between both face and voice identities. Therefore, it is likely that this region represents face and voice identity independently without integrating information across the two modalities. The angular gyrus was not found to be face-selective or voice-selective, and was therefore not included as a ROI. However, the exploratory whole-brain searchlight analysis did not show crossmodal person identity representations in this region. The two studies implicating the angular gyrus in face and voice identity integration (Hölig et al., 2017; Joassin et al., 2011) both trained their participants to associate initially unfamiliar faces with their corresponding voices, and therefore could not rule out the possibility that the observed results were due to learned associations between faces and voices, which could be completely arbitrary and unrelated to identity processing. Therefore, it is possible that the angular gyrus engages in associative memory processing, rather than identity processing.

Finally, this thesis showed no supporting evidence for the CFVP model, which proposes that face-responsive and voice-responsive brain regions directly exchange crossmodal information (e.g. Blank et al., 2011; von Kriegstein et al., 2005). Specifically, there was no evidence of crossmodal discrimination of person identities in face-selective or voice-selective brain regions. Previous studies showed activation of the FFA during voice recognition and described functional connectivity between the FFA and the voice-selective STS/STG (Schall et al., 2013; von Kriegstein et al., 2008, 2006, 2005; von Kriegstein & Giraud, 2006). However, in contrast to the study described in Chapter 3, these studies used an explicit voice recognition task. It is possible that the crossmodal coupling between the FFA and voice-selective regions that was observed in these previous studies is contingent on explicit voice recognition, rather than automatic process. Moroever, given the absence of evidence of the activation of voice-selective regions during face recognition, crossmodal coupling between face and voice regions may be specific to voice recognition. It has been shown that voice recognition is facilitated by prior exposure to a face to a greater extent than face recognition is facilitated by prior exposure to a voice (Stevenage et al., 2012). Moreover, studies have shown that recognising a familiar person from their face is easier than recognising the person from their voice (Damjanovic & Hanley, 2007; Hanley & Damjanovic, 2009; Hanley et al., 1998). Thus, the coupling between the FFA and voice-selective regions during voice recognition may serve the purpose of enhancing voice recognition by activating a representation of a person’s face, but this coupling may only be observable during explicit voice recognition tasks.

6.3 The informational content of the person identity representation in rpSTS

Although the work in this thesis showed that the rpSTS contains crossmodal representations of faces and voices, the type of face and voice information that is processed and integrated in the rpSTS is unfortunately still unclear. Chapter 4 failed to show any correlations between face and voice representations in the rpSTS and the candidate models used in this study. Therefore, brain representations in the rpSTS did not correlate with representations of any of the perceived face or voice characteristics that were tested, such as social traits or perceived similarity, nor did they correlate with representations of objective face or voice characteristics that reflected image-based and acoustic-based similarity, respectively.

Future studies should investigate alternative possible models for the type of information that is represented in rpSTS. One possibility is that the STS integrates person-specific patterns of movement from faces, voices, and bodies to assist in person identity recognition (Yovel & O'Toole, 2016). Specifically, Yovel and O'Toole (2016) proposed that the STS extracts an idiosyncratic multisensory "dynamic signature" for every known familiar person. In contrast to voices, faces do not need to be dynamic to convey information. However, during social interactions faces are rarely static, and display both rigid facial movements, such as nodding and turnings of the head, and non-rigid movements, such as facial expressions and eye gaze direction (Yovel & O'Toole, 2016). Moreover, facial movements, and mouth movements in particular, are intrinsically associated with speech. Given that both faces and voices play a fundamental role in social interactions, it seems plausible that dynamic aspects of person's face would be automatically associated with their voice and manner of speech, and may be integrated in a crossmodal person identity representation in the brain. The pSTS has been shown to be particularly sensitive to dynamic information in faces (Bernstein et al., 2018; Fox et al., 2009; Pitcher et al., 2011) and to respond more to combined face-voice stimuli than to faces and voices presented in isolation (Kreifelts et al., 2007; Robins et al., 2009; Watson, Latinus, Charest, et al., 2014), and is therefore a good candidate region for the processing and integration of dynamic identity information from faces and voices. Moreover, findings showing that the pSTS is involved in the processing of visually presented social interactions (Isik, Koldewyn, Beeler, & Kanwisher, 2018; Walbrin, Downing, & Koldewyn, 2018) suggest that any dynamic identity information represented in the pSTS is likely to be associated with social interactions. Future studies should attempt to compare brain representations of faces and voices in the pSTS with models describing facial movement and speech patterns. Descriptions of facial movement could be obtained from face videos using facial motion tracking software, and descriptions of speech patterns could be obtained from speech recordings using speech analysis software.

A second possibility is that the pSTS represents the degree of familiarity with different identities (Parkinson et al., 2014), which may be idiosyncratic for each participant. It has been shown that the pSTS represents the perceived 'social distance' between faces in terms of familiarity, according to which highly familiar faces are perceived as being 'closer' to a person whereas less familiar faces are perceived as being ‘further away’, in a similar way that objects can be closer or further away in space, and a verbally described time point can be closer or further away in time (Parkinson et al., 2014). Although participants reported being highly familiar with the presented identities, it is likely that each participant was more familiar with some identities than others. Specifically, given that the presented identities were famous-familiar people, participants were likely to have had different degrees of exposure to these people through the media based on personal preferences. For example, a participant who is interested in international politics but is not a fan of the Harry Potter films would have had more exposure to the face and the voice of Barack Obama than to the face and voice of Emma Watson and Daniel Radcliffe, despite being able to recognise the faces and the voices of all three people. Therefore, it is possible that the pSTS of each participant may code their level of familiarity with different people, regardless of whether familiarity is judged from a person’s face or voice. If this is the case, identities with similar levels of familiarity should be less discriminable that identities with different levels of familiarity. For example, identities such as Emma Watson and Daniel Radcliffe are likely to elicit similar levels of familiarity due to appearing in the same films. Similar levels of familiarity across multiple identities could be also elicited due to identities sharing the same occupation or nationality. For example, a participant may watch a lot of talk shows, but not a lot of interviews of singers. Future work could investigate the possibility that some identities are more discriminable that others in the pSTS due to different levels of familiarity with the different identities at the level of each individual participant. This could be possible by designing a comprehensive form of familiarity assessment that would include a measure of the degree of each participant’s exposure to the different identities, as well as their personal interest and engagement with each identity.

6.4 Stimulus-invariant representations of face and voice identity

Although the current thesis did not have this specific aim, the study described in Chapter 3 also allowed the investigation of representations of face and voice identity that are invariant to different, naturalistically varying tokens of the face and voice. These results provided novel and interesting insights that add to the considerable literature on this topic. In fact, a multitude of studies have found that face identity representations in face-responsive regions are invariant to different viewpoints of the same face (Anzellotti et al., 2014; Collins et al., 2016; Guntupalli et al., 2017; Natu et al., 2010; Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017), different emotional expressions (Nestor et al., 2011), different parts of the face (Anzellotti & Caramazza, 2016), or different photographs taken on separate occasions (Axelrod & Yovel, 2015), and that voice identity representations in voice-responsive regions are invariant to different recordings of the same persons' voice (Bonte et al., 2014; Formisano et al., 2008). However, in the Introduction chapter it was argued that these studies showed three main limitations. First, with the exception of Anzellotti et al. (2014), Anzellotti & Caramazza (2016), Guntupalli et al. (2017), and Formisano et al. (2008), the majority of studies did not test whether identity representations generalised to different tokens of the face or voice of the same identities. Therefore, these studies were not able to show that the observed identity representations were not specific to the selection of face images/voice recordings that were presented. Second, most studies presented the faces and voices of unfamiliar identities that participants were experimentally familarised with to different extents (but see Axelrod & Yovel, 2015 and Visconti Di Oleggio Castello et al., 2017). Behavioural studies have shown that categorising different tokens of the face and the voice of the same person as belonging to the same identity, i.e. 'telling people together', which is an essential component of identification (Burton, 2013), is largely contingent on being familiar with that person (Jenkins et al., 2011; Lavan, Burston, & Garrido, 2018). Third, and related to the previous limitation, the vast majority of studies presented face and voice stimuli that were highly controlled in terms of their low-level properties, were often artificially-generated, and showed low variability across different tokens of the face and voice of the same person. Familiar identity recognition involves abstracting the variability that is present in different exposures to the same person's face or voice in everyday life (Burton et al., 2016; Lavan, Burton, et al., 2018), and in the behavioural literature there has been a move towards using more naturalistic and variable face stimuli to study face recognition (Burton et al., 2016; Burton, 2013; Jenkins et al., 2011) and voice recognition (Lavan, Burton, et al., 2018). In contrast, in neuroimaging experiments the ability to 'tell people together' based on different, naturalistically varying presentations of their face and voice has either being ignored or it has not being adequately captured because of the use of highly-controlled and sometimes artificial face or voice stimuli (Lavan et al., 2018; Burton, 2013).

The study described in Chapter 3 addressed the first limitation of previous studies, relating to the generalisability of identity representations to novel tokens of the face and voice, by presenting multiple tokens of the face and voice of each identity, and using different tokens to obtain and test pattern discriminants. This approach is comparable to studies that trained and tested pattern classifiers using different tokens of the face or voice of each identity (Anzellotti et al., 2014; Anzellotti & Caramazza, 2016; Formisano et al., 2008; Guntupalli et al., 2017). However, in contrast to these studies, the present thesis also addressed the second limitation, which related to the use of initially unfamiliar face and voice identities, by presenting identities that were highly familiar to participants from both their face and their voice. This practice ensured that participants were able to 'tell together' different tokens of the same person's face and voice (Burton, 2013; Lavan, Burton, et al., 2018). Although two previous studies investigated face identity representations using familiar faces (Axelrod & Yovel, 2015; Visconti Di Oleggio Castello et al., 2017), these studies did not demonstrate that representations generalise to novel tokens of the face. Lastly, the current thesis addressed the third limitation of previous studies, relating to the lack of natural variability across different images and recordings of the same person's face and voice, by presenting naturalistically varying face videos and voice recordings that were highly variable across different tokens of the same identity. Thus, the observed face and voice representations were robust to the natural variability that is present in the faces and voices encountered in everyday life (Burton, 2013). Although one study presented different face photographs for each identity that were taken on separate occasions (Axelrod & Yovel, 2015), these images were selected to have neutral facial expressions, were converted to grey-scale, and equated in terms of luminance and colour, and cannot therefore be considered naturalistically varying.

Representations of face identity were found in the face-selective rFFA and rOFA, in the voice-selective bilateral TVAs, and in the multimodal rpSTS and precuneus/posterior cingulate. Multiple studies have previously shown identity representations in the FFA that generalise across different face images (Axelrod & Yovel, 2015), different emotional expressions (Nestor et al., 2011), and different viewpoints of the face (Verosky et al., 2013; Visconti Di Oleggio Castello et al., 2017), with two studies showing that identity representations also generalise to novel viewpoints of the face (Anzellotti et al., 2014; Guntupalli et al., 2017). In contrast, only one study showed evidence of invariant face identity representations in the OFA, for faces presented from different viewpoints (Anzellotti et al., 2014), and two studies shown viewpoint-invariant representations of face identity in the pSTS (Anzellotti & Caramazza, 2017; Visconti Di Oleggio Castello et al., 2017). It is possible that representations in the OFA and pSTS are easier to detect when using familiar faces that are also naturalistically varying. Lastly, one study found evidence of face identity representations in the precuneus (Visconti Di Oleggio Castello et al., 2017), but discrimination of face identities in the TVAs has not been shown before, to the best of my knowledge. Given that the TVAs cover a broad area of the temporal lobe, including regions of the anterior, mid, and posterior STS, it is likely that the observed face identity representations were due to the overlap with one or more face-selective regions, such as the rpSTS, or the mid and anterior STS (Fox et al., 2009; Pitcher et al., 2011), which were not localised in this study. In sum, the current thesis extends previous findings of face identity representations to show that these representations generalise to novel, naturalistically varying videos of the face.

Representations of voice identity were identified in the voice-selective TVAs and STS/STG and in the multimodal rpSTS, OFC, FP, and rTP-aIT. A previous study showed some evidence of representations of voice identity that generalised across different vowel sounds in the TVAs, but did not test whether these representations generalise to novel tokens of the same voice (Bonte et al., 2014). An earlier study from the same group showed voice identity representations in the STS/STG that generalise to novel voice tokens, but did not explicitly localise voice-selective regions (Formisano et al., 2008). The present study demonstrated that voice identity representations within voice-selective regions generalise to novel and naturalistically varying sentences spoken by the same identity. The ROIs for the rpSTS and rTP-aIT showed a large degree of overlap with the voice-selective ROIs in the right hemisphere, and this may account for the observed voice identity representations in these regions. Finally, to the best of my knowledge, invariant representations of voice identity have not been found previously in OFC and FP. This may be due previous studies presenting initially unfamiliar voice identities (Bonte et al., 2014; Formisano et al., 2008), in contrast to the present work.

To conclude, Chapter 3 aimed to identify brain regions that were able to both ‘tell people together’, i.e. categorise different, variable images of the same face, or recordings of the same voice, as being of the same identity, as well as ‘tell people apart’, i.e. distinguish between faces or voices belonging to different identities (Burton, 2013; Lavan, Burton, et al., 2018). Stimulus-invariant representations were found for face identity in the face-selective rFFA and rOFA, and for voice identity in the voice-selective bilateral TVAs and STS/STG, and in the multimodal OFC, FP, and rTP-aIT. Invariant representations of both face and voice identity were found in the rpSTS and in the precuneus/posterior cingulate. This work was able to demonstrate for the first time that identity representations in these regions generalise across multiple, naturalistically varying videos of the face/recordings of the voice of the same identity, simulating the recognition of familiar people in everyday life (Burton, 2013; Lavan, Burton, et al., 2018.

6.5 The informational content of face and voice identity representations

After identifying representations of face and voice identity in face-selective and voice-selective regions, the next question was what type of information is used by the different regions to distinguish between identities in their preferred modality. Previous work has associated face-responsive regions with the processing of visual information (Carlin & Kriegeskorte, 2017; Loffler et al., 2005; Weibert et al., 2018; Xu & Biederman, 2010; Xu et al., 2009), face gender (Contrereas et al., 2013; Freeman et al., 2010; Kaul et al., 2011; Mattavelli et al., 2012), and social traits (Engell et al., 2007; Freeman et al., 2014; Mattavelli et al., 2012; Said et al., 2009, 2011; Todorov et al., 2008; Todorov & Oosterhof, 2011; Winston et al., 2002). Voice-responsive regions have been associated with the processing of auditory information (Formisano et al., 2008; Latinus et al., 2013; von Kriegstein et al., 2007, 2010), voice gender (Charest et al., 2013; Lattner et al., 2005; Sokhi et al., 2005; Weston et al., 2015), and social traits (Bestelmeyer et al., 2012). However, as discussed in the Introduction chapter, the interpretation of these previous results is constrained by five main limitations. First, the majority of studies focused on identifying the neural correlates of one type of face or voice information, and there is a lack of studies using the same paradigm and stimuli to investigate multiple types of information. Therefore, with the exception of Mattavelli et al. (2012) and Said et al. (2011), these studies were not able to directly compare different types of information. Second, although many findings concern regions that overlap with the estimated location of face- or voice-selective regions, many studies did not explicitly investigate properties of independently defined face- and voice-selective regions. This is particularly the case for the TVAs, and as a result very little is known regarding the information content of voice representations in these regions (Charest et al., 2013; Latinus et al., 2013). A third limitation concerns the use of stimuli that were controlled in terms of their low-level visual and auditory features, primarily in studies investigating the processing of physical face or voice properties, and which appeared artificial compared with the natural variability in the faces and voices encountered in everyday life. As a consequence, findings from these studies may not apply to more naturalistically varying faces and voices. Fourth, very few studies used multivariate fMRI methods to investigate the information used by face-selective and voice-selective regions to distinguish between individual face identities (Carlin & Kriegeskorte, 2017; Contreras et al., 2013; Kaul et al., 2011; Weibert et al., 2018) or voice identities (Formisano et al., 2008). Finally, there is a lack of research on the informational content of face and voice representations in multimodal brain regions that respond selectively to both faces and voices.

Chapter 4 aimed to overcome the limitations of previous studies by comparing face and voice representations in face-selective, voice-selective, and multimodal (face-selective and voice-selective) regions to multiple models of both perceived and objective face and voice characteristics. Specifically, brain representations of individual face and voice stimuli were compared with models that captured both perceived and objective physical properties of the stimuli, face and voice gender, and perceived social traits. Models of perceived visual/auditory properties were computed based on ratings on visual/auditory pairwise similarity tasks, and models of objective visual/auditory properties were computed based on measures of stimulus similarity obtained using the OpenFace and Gabor-Jet programs for faces, and using f0 and AVTL for voices. A model of gender predicted that response patterns would be more similar between same-gender faces/voices than between different-gender faces/voices. Models of perceived social traits were computed based on ratings of the stimuli on trustworthiness, dominance, attractiveness, and positive-negative valence. In addition to face-selective, voice-selective, and multimodal brain regions, the informational content of the amygdala, which was anatomically defined, was also investigated. The amygdala was included as a ROI in order to test its proposed involvement in the processing of both social and non-social information in faces (Engell et al., 2007; Freeman et al., 2014; Mattavelli et al., 2012; Said et al., 2009, 2011; Todorov et al., 2008; Todorov & Oosterhof, 2011; Winston et al., 2002).

A review of the literature investigating the information processed in face-selective regions, presented in the Introduction chapter, suggested that these regions were most likely to process the visual properties of faces, including those used to categorise face gender (Carlin & Kriegeskorte, 2017; Loffler et al., 2005; Mattavelli et al., 2012; Said et al., 2011; Weibert et al., 2018; Xu & Biederman, 2010; Xu et al., 2009). This was confirmed by the results, which showed that brain representations of faces in the rFFA and rOFA are associated with information about the objective similarity between faces and face gender, and that brain representations of faces in the rFFA are also associated with information about the perceived visual similarity between faces. Relating back to Chapter 3, which showed that the rFFA and the rOFA could discriminate between different face identities, these findings suggest that these regions may use information relating to the visual similarity between faces and face gender to distinguish between different identities. However, the models for perceived and objective visual similarity and gender were correlated with each other, suggesting that they describe similar information, and their independent contributions could be disentangled in the present analysis. Brain representations in the rOFA, but not the rFFA, correlated with the Gabor-Jet similarity model, and therefore computations in this region are likely to rely more on low-level visual information to distinguish between different identities. In contrast, the rFFA, but not the rOFA, was associated with the perceived similarity between faces, and may therefore rely more on higher-level visual information, such as gender and age, to distinguish between different identities. Correlations between the Gabor-Jet model and other models, including perceived similarity, were low, suggesting that this model captures distinct information. Lastly, an exploratory analysis comparing face representations across different regions showed that representations in the rFFA and rOFA were correlated, suggesting that information in these regions is at least partially shared. It should be noted that the results also showed a correlation between representations in the rOFA and perceived dominance. Although this finding suggests that the rOFA may also process high level social information from faces, the correlation with dominance was very low and well below the noise-ceiling, and thus should be interpreted with caution.

Previous findings showed that voice-responsive regions process information about acoustic properties of voices (Formisano et al., 2008; Latinus et al., 2013; von Kriegstein et al., 2007, 2010), and that sub-regions of the TVAs may be sensitive to gender information in voices (Charest et al., 2013), These findings were supported by the results of Chapter 4 showing that brain representations of voices in the bilateral TVAs and STS/STG were associated with perceived auditory similarity (with the exception of the lTVA), with objective stimulus similarity as defined by f0 and AVTL, and with voice gender. Chapter 3 demonstrated that all four of these regions could distinguish between individual voice identities, and it is likely that this discrimination is based on auditory similarity and voice gender. The models for auditory similarity and gender were correlated with each other, suggesting that they describe similar information. However, the independent contribution of each model could not be defined in the present analysis, and in future it would be interesting to test the contributions of the different models to explaining the variance in the brain RDMs using multiple regression analysis. Lastly, an exploratory analysis presented in Chapter 4 showed similar voice representations across the bilateral TVAs and STS/STG, supporting the findings that these regions process similar information. Taken together, these findings substantially extend the current knowledge of the informational content of voice representations in the TVA by suggesting that they use acoustic information, and information associated with voice gender, to distinguish between individual voice identities.

Little is known regarding the informational content of multimodal people-selective brain regions. However, it has been shown that regions such as the pSTS and the OFC, which were found to selectively respond to both faces and voices in Chapter 3, represent similar information from multiple modalities (Chikazoe et al., 2014; Peelen et al., 2010). Thus, it was speculated that multimodal regions may represent information that is available through both modalities, such as information regarding social traits (McAleer et al., 2014; Oosterhof & Todorov, 2008). Moreover, based on evidence from univariate fMRI studies that the amygdala processes both social and non-social information in faces (Bzdok et al., 2011; Mende-Siedlecki et al., 2013), it was predicted that the amygdala would represent both social and physical face information. However, Chapter 3 showed no evidence of correlations between face or voice representations in multimodal regions and models of perceived trustworthiness, dominance, attractiveness, and positive-negative valence, or any of the other models that were tested. Moreover, face representations in the amygdala did not correlate with any of the face models. One possible reason for this is that RDMs for multimodal regions and the amygdala showed very low inter-subject reliability, suggesting that a high level of noise was present in the activity patterns. For multimodal regions, a second possibility is that, given that these regions are not exclusively face-selective and voice-selective, they may code higher-level information that is highly abstracted from the input modalities. This may involve cognitive processes such as attention (Downar, Crawley, Mikulis, & Davis, 2000), reward processing (O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001), retrieval of general social knowledge about people (Olson, McCoy, Klobusicky, & Ross, 2013; Wang et al., 2017), social distance based on degrees of familiarity (Parkinson, Kleinbaum, & Wheatley, 2017; Parkinson et al., 2014), and episodic memory (Lundstrom et al., 2003). As mentioned previously in relation to the pSTS, future work may benefit from comparing brain representations of faces and voices in multimodal regions with more complex models based on the level of familiarity of each individual participant with each identity, including their level of exposure and engagement with each identity. For the amygdala, it is possible that its sensitivity to information in faces is confined to modulations in magnitude, and that this information is not used to distinguish between individual identities. Lastly, an exploratory analysis showed that face and voice representations in multimodal regions and the amygdala were correlated with representations in face-selective and voice-selective brain regions, suggesting that that they share some informational content; however, these correlations were very low, and should be interpreted with caution.

6.6 The relationship between perceived information in the face and voice

Although Chapter 4 found no evidence of multimodal brain regions that represent similar types of information from faces and voices, the study presented in Chapter 5 showed that, on a behavioural level, similar information is perceived from a familiar person's face and their voice in relation to both social traits and perceived physical similarity. Specifically, high correlations were observed between ratings of familiar faces and voices on trustworthiness, dominance, attractiveness, and positive-negative valence. These findings suggest that social and physical person-related information that is extracted from the faces and voices of familiar people is highly consistent across modalities. In addition to familiar people, Chapter 5 also showed correlations between ratings of the faces and voices of unfamiliar people on trustworthiness, dominance, attractiveness, valence, and perceived similarity. However, as predicted, these correlations were significantly lower than the correlations for familiar people. For familiar people, it is highly likely that the greater similarity between the judgements of the face and voice, compared with unfamiliar people, is due to prior knowledge of the person and multiple experiences of concurrent exposure to both their face and their voice. In contrast, for unfamiliar people there is no known correspondence between a person's face and their voice, and no prior knowledge of the person. There is, however, some evidence that the faces and voices of unfamiliar people convey concordant information regarding physical characteristics of a person, such as masculinity-femininity (Smith et al., 2016a). In regard to information on social traits, as discussed in the Introduction chapter, finding from studies that have compared ratings of unfamiliar faces and voices on attractiveness (Abend et al., 2015; Lander, 2008; Oguchi & Kikuchi, 1997;

Rezlescu et al., 2015; Saxton et al., 2009; Valentova et al., 2017; Wells et al., 2013) and dominance (Han et al., 2017; Rezlescu et al., 2015) have been inconsistent in regard to the extent to which ratings were similar across modalities, the direction of the correlations, and the influence of stimulus and/or participant gender on the relationship between face and voice ratings. Only one study compared ratings of trustworthiness across faces and voices, and found a correlation when male and female stimuli were analysed together, but not when they were analysed separately (Rezlescu et al., 2015).

The inconsistencies in the findings of previous studies that compared ratings of unfamiliar faces and voices on social judgements may have been due to two main factors. First, these studies collected ratings of a single face image or voice recording of each identity, and it has been shown that social judgements can be different for different face images of the same (unfamiliar) person (Sutherland et al., 2017; Todorov & Porter, 2014). Therefore, it is likely that the across-modality variability in judgements of different tokens of the same person's face and voice would be even greater, and would decrease chances of detecting potential similarities in social judgements across modalities. Second, in previous studies correlations between faces and voices were assessed at the stimulus level, and did not take into account individual differences in ratings of the faces and voices across participants. Specifically, correlations were computed between each individual face and voice across the ratings of all participants. While studies have shown that ratings of faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and voices (McAleer et al., 2014) are largely consistent across participants, there is also evidence that the personal attitudes and experiences of each individual can influence judgements of social traits in faces (Hehman et al., 2017; Kramer et al., 2018; Stanley et al., 2011; Watkins et al., 2010). Therefore, it is possible that not taking into account these potential individual differences in the ratings of faces and voices across participants results in less sensitivity to detect consistencies in ratings across the two modalities.

The study presented in Chapter 5 attempted to address the aforementioned issues arising from the variability in the ratings of different tokens of the same face and voice, and from individual differences across participants in their ratings of faces and voices. The first issue, concerning the within-person variability in ratings, was addressed by using a novel experimental paradigm in which ratings were based on the consecutive presentation of multiple, naturalistically varying face videos and voice recordings of each identity. The aim of this paradigm was to encourage participants to base their ratings on face and voice characteristics that remained stable across different presentations of the same person's face and voice. The second issue, concerning individual differences in ratings, was addressed by conducting all analyses at the individual participant level, in that correlations were computed between each participant's ratings of all faces and voices, as opposed to being computed between each individual face and voice based on the ratings of all participants. It is possible that both of these approaches contributed to the finding of similar ratings of faces and voices of unfamiliar people. However, as discussed in Chapter 5, it is also possible that the use of a diverse set of identities in terms of age, nationality, gender, and race in the present work may have enabled participants to correctly predict which face corresponded to which voice, thus encouraging similar ratings of the faces and voices of the same people. Future studies comparing ratings of faces and voices should therefore use a larger and more homogenous set of unfamiliar identities.

A possible reason for the existence of similarities in the ratings of the faces and voices of unfamiliar people is that people form multisensory representations of social traits that include information about what a person looks like and sounds like through their experience of others in everyday life. Specifically, the presence/absence of certain social traits may be associated through experience with certain face and voice characteristics that frequently co-occur across different people (Over & Cook, 2018), possibly as a result of physical and developmental changes that affect both the face and the voice. This is particularly likely to be the case for social traits that have been associated with physical person characteristics that are conveyed by both the face and the voice. For example, dominance has been associated with masculinity in both voices (McAleer et al., 2014) and faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013), and judgements of masculinity-femininity have been shown to be consistent across faces and voices (Smith et al., 2016a). These learnt associations between face and voice characteristics, on the one hand, and social traits, on the other hand, may influence judgements of the faces and voices of newly encountered unfamiliar people, leading to similar social judgements of the face and voice.

An exploratory analysis conducted in Chapter 5 showed that the relationships between different social judgements were similar for both faces and voices, with higher correlations between ratings of trustworthiness, attractiveness, and valence than between each of these three judgements and dominance. This was the case for both familiar and unfamiliar identities. Previous studies have investigated the relationship between multiple social judgements of unfamiliar faces and voices, independently for each modality, using dimensionality-reduction techniques, and have shown that valence/trustworthiness and dominance form orthogonal dimensions for faces (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and for voices (McAleer et al., 2014). The present findings suggest that individual traits are conceptualised in a similar way for faces and voices for both familiar and unfamiliar people. Moreover, given previous findings of similarities between the concepts of different emotions across faces and voices (Kuhn et al., 2017), as well as crossmodal representations of emotion from faces and voices in the left rpSTS (Peelen et al., 2010), it is seems increasingly likely that representations of social traits from both faces and voices may also co-exist exist in the same brain regions.

Finally, Chapter 5 also presented an analysis that compared ratings of perceived similarity with ratings of social judgements, separately for faces and voices, in order to test whether the differences between the ratings of different identities on social judgements was related to their degree of perceived similarity. The results of this analysis suggested that perceived similarity explained very little of the variance in social judgements. For example, if two faces are given similar trustworthiness ratings, this does not imply that they look similar. Therefore, it is possible that physical face and voice characteristics play a minor role in the formation of social impressions. However, the physical features used by participants to judge perceived similarity are unknown, and may differ from physical features that could influence social judgements.

Taken together, the findings regarding the relationship between face and voice ratings in familiar and unfamiliar people suggest that gaining familiarity with a person may increase the similarity between judgements of their face and voice. To explicitly test this possibility, future work should collect ratings of faces and voices before and after experimentally familiarising participants with a set of identities. The relationship between the ratings of the face and voice of each identity could then be compared before and after becoming familiar with the identity, to determine how, and to what extent, this relationship is influenced by familiarity.

6.7 General limitations and future directions

The current thesis used famous-familiar identities to investigate representations of face identity, voice identity, and person identity in the brain, and to investigate the relationship between information perceived from the face and information perceived from the voice. This work presents a significant advancement from previous neuroimaging studies that investigated face, voice, and person identity representations, the majority of which experimentally familiarised participants with the faces and voices of initially unfamiliar identities (e.g. Anzellotti et al., 2014; Formisano et al., 2008; Hölig et al., 2017; Joassin et al., 2011). Familiar faces have been shown previously to elicit a wider range of brain activation compared with recently learned faces, suggesting that they engage different processing systems (Leveroni et al., 2000). However, although participants reported being highly familiar with both the faces and the voices of the famous individuals, it is unlikely that they would have ever interacted with these people on a personal level. Instead, familiarity was most likely gained through the media from a third-person perspective. In contrast, familiarity with personally-familiar acquaintances is mostly acquired through direct social interactions. Personally-familiar faces have been shown to elicit stronger brain activation compared with famous-familiar faces in a number of brain regions, including the bilateral pSTS, posterior cingulate/precuneus, and fusiform gyrus (Gobbini et al., 2004; Sugiura et al., 2011). Therefore, it is possible that brain representations of individual face, voice, and person identities may differ between famous-familiar identities and personally-familiar identities. Moreover, the relationship between perceived information extracted from the face and perceived information extracted from the voice, and the way that different social traits are conceptualised in each modality, may differ depending on the level of familiarity. To investigate these possibilities, future work could attempt to investigate representations of the faces and voices of personally-familiar people. This approach is challenging because it would require either identifying people who are personally-familiar to all of the participants, or creating individual face and voice stimuli for each participant. The first approach would be easier to implement, and some researchers have used stimuli derived from university lecturers to test undergraduate student participant groups (e.g. Lavan et al., 2016), but given that it is unlikely that students are highly familiar with their lecturers, this approach also has its limitations. Alternatively, a group of participants who are personally familiar with each other could be recruited both to take part in the experiment themselves and to provide face and voice stimuli to be used for the testing of the other participants in the group. Such a group could potentially comprise undergraduate students attending the same course module.

A further potential limitation of the famous-familiar identities used in the current thesis is the high diversity between the different identities in terms of age, nationality, accent, gender, and race. These particular famous individuals were selected because they proved to be highly recognisable from both their faces and their voices, and therefore are likely to display particularly distinctive facial and vocal features. Although the different tokens of the face and the voice of each person were unconstrained and naturalistically varying, it is likely that the variability between the different identities was higher than the variability between the different tokens of the same person’s face and voice. Consequently, this variability in visual and auditory appearance may have facilitated the discrimination between the different face identities, voice identities, and person identities in the brain. In particular, there is a possibility that regions that were found to discriminate between pairs of identities could do so, at least in part, due to marked differences in facial features and vocal features between the different identities. To address this possibility, future research should attempt to replicate the findings of the current thesis using a more homogenous set of familiar identities. Specifically, identities could be matched in terms of gender, nationality, race, accent, and age group, but different tokens of the face and voice of the same person should vary naturalistically in terms of visual and auditory appearance. This approach would limit the between-person variability while at the same time maintaining a similar level of within-person variability across different tokens of the same person's face and voice as the present thesis.

A final overall limitation is the low intra-subject and inter-subject reliability of the representational dissimilarity matrices (RDMs) for face and voice identities that were computed in Chapter 3. Intra-subject reliability was measured for each ROI as the correlation between the face or voice RDM computed from data in scanning session 1 and the face or voice RDM computed from data in scanning session 2, which took place on a separate day. Chapter 3 showed very low correlations (mean r < .10) between RDMs across sessions in all ROIs for both faces and voices. These correlations were not significantly greater than zero across participants for any of the comparisons. Given than the brain activity patterns elicited by individual face and voice identities were found to be highly reliable across scanning sessions for faces in face-selective regions and for voices in voice-selective regions, it is unlikely that the low correlations between RDMs across the two sessions were due to inconsistent brain activity patterns. A possible explanation is that the individual representations of face or voice identities within a certain brain region, i.e. their elicited brain activity patterns, may have been equally dissimilar to each other. As a consequence, the distances between face or voice representations in the RDMs may have not been variable enough to detect similarities in representational geometry between two RDMs from different sessions.

Inter-subject reliability was measured for each ROI as the correlation between each participant's RDM (averaged across the two scanning sessions) and the average of the RDMs of all other participants. The aim of this analysis was to determine the maximum possible correlation between a brain RDM and a candidate model RDM given the noise in the data. Low mean correlations (< .25) were observed for all ROIs, for both faces and voices. On first thought, computing inter-subject reliability as a measure of noise ceiling for the correlations in Chapter 4 may not have been ideal, given that the main analysis compared each individual's brain RDM with models based on their own behavioural data. It is possible that representations of familiar faces and voices had a strong idiosyncratic component, which could be due to the degree of exposure that each participant had to the different identities, as well as their personal attitude and sentiments towards each identity. However, idiosyncratic face and voice representations are unlikely to have affected the results substantially, given that the intra-subject reliability values were so low. Instead, the reason for low inter-subject reliability may be the same as the potential reason for low intra-subject reliability: the brain activity patterns elicited by individual identities were equally dissimilar to each other, making it difficult to detect similarities between RDMs using correlation. In conclusion, future studies should focus on optimising experimental designs and protocols to obtain RDMs that are more reliable across different sub-sets of data, and across different brains. This practice may increase the potential to characterise the informational content of face and voice representations by comparing RDMs of faces and voices in the brain to candidate RDMs.

6.8 Conclusions

The present thesis attempted to answer three main questions: 1) how face and voice information is integrated in the brain to form representations of person identity, 2) what is the informational content of face and voice representations in face-selective, voice-selective, and multimodal brain regions, and 3) how information perceived from the face is related to information perceived from the voice in familiar and unfamiliar people.

In regard to the first question, it was demonstrated that face and voice information is integrated in the rpSTS, a multimodal brain region that responds selectively to both faces and voices. This finding provides support for the MP model of face and voice integration which proposes that person identity is represented in multimodal brain regions. No evidence was found to support the CFVP model, which proposes that face-responsive and voice-responsive regions exchange crossmodal information related to person identity.

The present thesis additionally showed stimulus-invariant representations of face identity in face-selective regions, and stimulus-invariant representations of voice identity in voice-selective regions, which generalised across multiple, naturalistically varying videos of the face and recordings of the voice of the same person. It was shown that brain representations of faces in the face-selective rFFA and rOFA were primarily associated with information relating to the perceived and objective similarity of faces and gender, and that brain representations of voices in the voice-selective bilateral STS/STG and TVAs were primarily associated with information relating to perceived and objective auditory similarity of voices and gender. These findings suggest that face-selective and voice-selective regions primarily use physical information that is directly related to their preferred modalities to distinguish between individual face and voice identities. However, the informational content of multimodal brain regions, including the rpSTS, could not be defined in the current thesis, and the informational content of face and voice representations in these regions remains an open question.

Lastly, in regard to the third question, it was demonstrated that highly consistent information regarding social traits and perceived similarity is extracted from the face and the voice of familiar people, whereas information extracted from the face and the voice of unfamiliar people is less consistent but still similar to an extent. These findings suggest that concordant information may be conveyed by the faces and voices of familiar people due to prior knowledge of the person, and that some concordant information may also be present in the faces and voices of unfamiliar people, possibly due to learned associations between social traits and certain face and voice features that frequently co-occur across different people.

In sum, this thesis addressed important gaps in the literature regarding representations of face, voice, and person identity and their informational content though novel applications of RSA, which enabled direct comparisons between brain representations across the faces and voices, and between brain representations and models of face and voice properties. Finally, this thesis demonstrated that representations of face, voice, and person identity in the brain can be detected when using naturalistically varying face and voice stimuli, and highlights the need for neuroimaging studies to use stimuli that better resemble the faces and voices encountered in everyday life.

References

Abend, P., Pflüger, L. S., Koppensteiner, M., Coquerelle, M., & Grammer, K. (2015). The sound of female shape: A redundant signal of vocal and facial attractiveness. *Evolution and Human Behavior*, *36*(3), 174–181. http://doi.org/10.1016/j.evolhumbehav.2014.10.004

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*.

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage*, *52*(4), 1528–1540. http://doi.org/10.1016/j.neuroimage.2010.05.048

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage*, *23*(3), 905–913. http://doi.org/10.1016/j.neuroimage.2004.07.060

Anzellotti, S., & Caramazza, A. (2014). The neural mechanisms for the recognition of face identity in humans. *Frontiers in Psychology*, *5*(JUN), 1–6. http://doi.org/10.3389/fpsyg.2014.00672

Anzellotti, S., & Caramazza, A. (2016). From Parts to Identity: Invariance and Sensitivity of Face Representations to Different Face Halves. *Cerebral Cortex*, *26*(5), 1900–1909. http://doi.org/10.1093/cercor/bhu337

Anzellotti, S., & Caramazza, A. (2017). Multimodal representations of person identity individuated with fMRI. *Cortex*, *89*, 85–97. http://doi.org/10.1016/j.cortex.2017.01.013

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, *24*(8), 1988–1995. http://doi.org/10.1093/cercor/bht046

Axelrod, V., & Yovel, G. (2013). The challenge of localizing the anterior temporal face area: A possible solution. *NeuroImage*, *81*, 371–380. http://doi.org/10.1016/j.neuroimage.2013.05.015

Axelrod, V., & Yovel, G. (2015). Successful decoding of famous faces in the fusiform face area. *PLoS ONE*, *10*(2), 19–25. http://doi.org/10.1371/journal.pone.0117126

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on* (pp. 59–66). IEEE.

Baumann, O., & Belin, P. (2009). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. http://doi.org/10.1007/s00426-008-0185-z

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190–1192. http://doi.org/10.1038/nn1333

Beauchamp, M. S., Lee, K., Argall, B., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. http://doi.org/10.1016/S0896-6273(04)00070-4

Belin, P. (2017). Similarities in face and voice cerebral processing. *Visual Cognition*, *25*(4-6), 658–665. http://doi.org/10.1080/13506285.2017.1339156

Belin, P., Fecteau, S., Bédard, C., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. http://doi.org/10.1016/j.tics.2004.01.008

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker ' s voice in right anterior temporal lobe. *Control*, *14*(16), 12–16. http://doi.org/10.1097/01.wnr.0000091689.94870.85

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. http://doi.org/10.1038/35002078

Benetti, S., Novello, L., Maffei, C., Rabini, G., Jovicich, J., & Collignon, O. (2018). White matter connectivity between occipital and temporal regions involved in face and voice processing in hearing and early deaf individuals. *NeuroImage*, *179*(May), 263–274. http://doi.org/10.1016/j.neuroimage.2018.06.044

Bernstein, M., Erez, Y., Blank, I., & Yovel, G. (2018). An Integrated Neural Framework for Dynamic and Static Face Processing. *Scientific Reports*, *8*(1), 2–11. http://doi.org/10.1038/s41598-018-25405-9

Bestelmeyer, P. E. G., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., & Belin, P. (2012). Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. *Cerebral Cortex*, *22*(6), 1263–1270. http://doi.org/10.1093/cercor/bhr204

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of objects and face recognition. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *352*, 1203–1219.

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct Structural Connections between Voice- and Face-Recognition Areas. *The Journal of Neuroscience*, *31*(36), 12906–12915. http://doi.org/10.1523/JNEUROSCI.2091-11.2011

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience and Biobehavioral Reviews*, *47*, 717–734. http://doi.org/10.1016/j.neubiorev.2014.10.022

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-Dependent Decoding of Speaker and Vowel Identity from Auditory Cortical Response Patterns. *Journal of Neuroscience*, *34*(13), 4548–4557. http://doi.org/10.1523/JNEUROSCI.4339-13.2014

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour*. http://doi.org/10.1038/s41562-018-0376-6

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105–116.

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., … Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female faces? *Perception*, *22*(2), 131–152.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. http://doi.org/10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(77), 305–327. http://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. http://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361–380. http://doi.org/10.1017/CBO9781107415324.004

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cognitive Science*, *40*(1), 202–223. http://doi.org/10.1111/cogs.12231

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video:Evidence from security surveillance. *Psychological Science*, *10*(3), 243–248. http://doi.org/10.1111/1467-9280.00144

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., … Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function*, *215*(3–4), 209–223. http://doi.org/10.1007/s00429-010-0287-4

Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, *11*(12), 1110–1123. http://doi.org/10.1093/cercor/11.12.1110

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., … David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. http://doi.org/10.1126/science.276.5312.593

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649–657. http://doi.org/10.1016/S0960-9822(00)00513-3

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543. http://doi.org/10.1016/j.tics.2007.10.001

Carlin, J. D., Calder, A. J., Kriegeskorte, N., Nili, H., & Rowe, J. B. (2011). A head view-invariant representation of gaze direction in anterior superior temporal sulcus. *Current Biology*, *21*(21), 1817–1821. http://doi.org/10.1016/j.cub.2011.09.025

Carlin, J. D., & Kriegeskorte, N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Computational Biology*, *13*(7), 1–28.

http://doi.org/10.1371/journal.pcbi.1005604

Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PLoS ONE*, *7*(2). http://doi.org/10.1371/journal.pone.0031353

Chan, A. W., & Downing, P. E. (2011). Faces and eyes in human lateral prefrontal cortex. *Frontiers in Human Neuroscience*, *5*(June), 51. http://doi.org/10.3389/fnhum.2011.00051

Charest, I., & Kriegeskorte, N. (2015). The brain of the beholder: honouring individual representational idiosyncrasies. *Language, Cognition and Neuroscience*, *30*(4), 367–379. http://doi.org/10.1080/23273798.2014.1002505

Charest, I., Pernet, C. R., Latinus, M., Crabbe, F., & Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover FMRI design. *Cerebral Cortex*, *23*(4), 958–966. http://doi.org/10.1093/cercor/bhs090

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., … Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*, 127. http://doi.org/10.1186/1471-2202-10-127

Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, *17*(8), 1114–1122. http://doi.org/10.1038/nn.3749

Cleveland, T. F. (1977). Acoustic properties of voice timbre types and their influence on voice classification. *The Journal of the Acoustical Society of America*, *61*(6), 1622–1629. http://doi.org/10.1121/1.381438

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. *Frontiers in Human Neuroscience*, *10*(May), 1–11. http://doi.org/10.3389/fnhum.2016.00189

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … Haxby, J. V. (2012). The Representation of Biological Classes in the Human Brain. *Journal of Neuroscience*, *32*(8), 2608–2618. http://doi.org/10.1523/JNEUROSCI.5547-11.2012

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race. *PLoS ONE*, *8*(7).

http://doi.org/10.1371/journal.pone.0069684

Dabbs, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men. *Personality and Individual Differences*, *27*(4), 801–804. http://doi.org/10.1016/S0191-8869(98)00272-4

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition*, *35*(6), 1205–1210. http://doi.org/10.3758/BF03193594

Davies-Thompson, J., & Andrews, T. J. (2012). Intra- and interhemispheric connectivity between face-selective regions in the human brain. *Journal of Neurophysiology*, *108*(11), 3087–3095. http://doi.org/10.1152/jn.01171.2011

Davies-Thompson, J., Elli, G. V, Rezk, M., Benetti, S., Ackeren, M. Van, & Collignon, O. (2018). Hierarchical brain network for face and voice integration of emotion expression. bioRxiv doi: https://doi.org/10.1101/197426

Davies-Thompson, J., Newling, K., & Andrews, T. J. (2013). Image-invariant responses in face-selective regions do not explain the perceptual advantage for familiar face recognition. *Cerebral Cortex*, *23*(2), 370–377. http://doi.org/10.1093/cercor/bhs024

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. http://doi.org/10.1093/cercor/bhv111

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, *3*(3), 277–283. http://doi.org/10.1038/72991

Driver, J., & Noesselt, T. (2008). Multisensory Interplay Reveals Crossmodal Influences on “Sensory-Specific” Brain Regions, Neural Responses, and Judgments. *Neuron*, *57*(1), 11–23. http://doi.org/10.1016/j.neuron.2007.12.013

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science*, *1*(1), 393–416. http://doi.org/10.1146/annurev-vision-082114-035518

Eger, E. (2004). BOLD Repetition Decreases in Object-Responsive Ventral Visual Areas Depend on Spatial Attention. *Journal of Neurophysiology*, *92*(2), 1241–1247. http://doi.org/10.1152/jn.00206.2004

Ellis, A. W., Young, A. W., & Critchley, E. M. R. R. (1989). Loss of Memory for

People Following Temporal Lobe Damage. *Brain*, *112*(6), 1469–1483. http://doi.org/10.1093/brain/112.6.1469

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology (London, England : 1953)*, *88 (Pt 1)*, 143–56. http://doi.org/10.1111/j.2044-8295.1997.tb02625.x

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience*, *19:9*(2002). https://doi.org/10.1162/jocn.2007.19.9.1508

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*(12), 4772–4788. http://doi.org/10.1093/cercor/bhv136

Ewbank, M. P., & Andrews, T. J. (2008). Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. *NeuroImage*, *40*(4), 1857–1870. http://doi.org/10.1016/j.neuroimage.2008.01.049

Fairhall, S. L., & Ishai, A. (2007). Effective connectivity within the distributed cortical network for face perception. *Cerebral Cortex*, *17*(10), 2400–2406. http://doi.org/10.1093/cercor/bhl148

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences*, *4*(7), 258–267. http://doi.org/10.1016/S1364-6613(00)01494-7

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511–1522. http://doi.org/10.1121/1.427148

Föcker, J., Hölig, C., Best, A., & Röder, B. (2011). Crossmodal interaction of facial and vocal person identity information: An event-related potential study. *Brain Research*, *1385*, 229–245. http://doi.org/10.1016/j.brainres.2011.02.021

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). “Who” Is Saying “What”? Brain-Based Decoding of Human Voice and Speech. *Science*, *3225903*(5903), 970–973. http://doi.org/10.1017/CBO9781107415324.004

Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network:

Optimization of the functional localizer in fMRI. *Human Brain Mapping*, *30*(5), 1637–1651. http://doi.org/10.1002/hbm.20630

Freeman, J. B., & Johnson, K. L. (2016). More Than Meets the Eye: Split-Second Social Perception. *Trends in Cognitive Sciences*, *20*(5), 362–374. http://doi.org/10.1016/j.tics.2016.03.003

Freeman, J. B., Rule, N. O., Adams, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*(6), 1314–1322. http://doi.org/10.1093/cercor/bhp195

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala Responsivity to High-Level Social Information from Unseen Faces. *Journal of Neuroscience*, *34*(32), 10573–10581. http://doi.org/10.1523/JNEUROSCI.5063-13.2014

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*(9), 2273–2282. http://doi.org/10.1016/j.neuropsychologia.2011.04.027

Gainotti, G. (2014). Cognitive models of familiar people recognition and hemispheric asymmetries. *Frontiers in Bioscience (Elite Edition)*, *6*(1), 148–158.

Gangestad, S. W., Thornhill, R., & Yeo, R. A. (1994). Facial attractiveness, developmental stability, and fluctuating asymmetry. *Ethology and Sociobiology*, *15*(2), 73–85. http://doi.org/10.1016/0162-3095(94)90018-3

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The Fusiform “Face Area” is Part of a Network that Processes Faces at the Individual Level. *Journal of Cognitive Neuroscience*, *12*(3), 495–504. http://doi.org/10.1162/089892900562165

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage*, *22*(4), 1628–1635. http://doi.org/10.1016/j.neuroimage.2004.03.049

Goebel, R., & Van Atteveldt, N. (2009). Multisensory functional magnetic resonance imaging: A future perspective. *Experimental Brain Research*, *198*(2–3), 153–164. http://doi.org/10.1007/s00221-009-1881-7

Goesaert, E., & Op de Beeck, H. P. (2013). Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(19), 8549–58. http://doi.org/10.1523/JNEUROSCI.1829-12.2013

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, *107*(1–3), 293–321. http://doi.org/10.1016/S0001-6918(01)00019-1

Gschwind, M., Pourtois, G., Schwartz, S., Van De Ville, D., & Vuilleumier, P. (2012). White-matter connectivity between face-responsive regions in the human brain. *Cerebral Cortex*, *22*(7), 1564–1576. http://doi.org/10.1093/cercor/bhr226

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, *26*(6), 2919–2934. http://doi.org/10.1093/cercor/bhw068

Guntupalli, J. S., Wheeler, K. G., & Gobbini, M. I. (2017). Disentangling the Representation of Identity from Head View Along the Human Face Processing Pathway. *Cerebral Cortex*, *27*(1), 46–53. http://doi.org/10.1093/cercor/bhw344

Han, C., Kandrik, M., Hahn, A. C., Fisher, C. I., Feinberg, D. R., Holzleitner, I. J., … Jones, B. C. (2017). Interrelationships among men's threat potential, facial dominance, and vocal dominance. *Evolutionary Psychology*, *15*(1), 1–4. http://doi.org/10.1177/1474704917697332

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 (Hove, England)*, *17*(8), 830–9. http://doi.org/10.1080/09658210903264175

Hanley, J. R., Pearson, N. A., & Young, A. W. (1990). Impaired memory for new visual forms 62. *Brain*, *113 (4*), 1131–1148. https://doi.org/10.1093/brain/113.4.1131

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. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 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. *Scientific*

Reports, *6*, 37494. http://doi.org/10.1038/srep37494

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distrubuted and Overlapping Representations of Face and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. http://doi.org/10.1126/science.1063736

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*(1), 189–199. http://doi.org/10.1016/S0896-6273(00)80690-X

Haynes, J. D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, *15*(14), 1301–1307. http://doi.org/10.1016/j.cub.2005.06.026

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529. http://doi.org/10.1037/pspa0000090

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the Faces of Robbers Captured on Video. *Applied Cognitive Psychology*, *15*(4), 445–464. http://doi.org/10.1002/acp.718

Henson, R. N., & Mouchlianitis, E. (2007). Effect of spatial attention on stimulus-specific haemodynamic repetition effects. *NeuroImage*, *35*(3), 1317–1329. http://doi.org/10.1016/j.neuroimage.2007.01.019

Hill, H., & Bruce, V. (1996). Effects of Lighting on the Perception of Facial Surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986–1004. http://doi.org/10.1037/0096-1523.22.4.986

Hillenbrand, J., & Clark, M. (2009). The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception & Psychophysics*, *71*(5), 1150–1166. http://doi.org/10.3758/APP

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, *57*(2), 482–494. http://doi.org/10.1016/j.neuroimage.2011.04.056

Hoffman, E. a, & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature*

Neuroscience, *3*(1), 80–84. http://doi.org/10.1038/71152

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221–1240. http://doi.org/10.1068/p3252

Hölig, C., Föcker, J., Best, A., Röder, B., & Büchel, C. (2017). Activation in the angular gyrus and in the pSTS is modulated by face primes during voice recognition. *Human Brain Mapping*, *38*(5), 2553–2565. http://doi.org/10.1002/hbm.23540

Hughes, S. M., Harrison, M. a, & Gallup, G. G. (2002). The sound of symmetry: Voice as a marker of developmental instability. *Evolution and Human Behavior*, *23*, 173–180. http://doi.org/10.1016/S1090-5138(01)00099-X

Huk, A. C., Ress, D., & Heeger, D. J. (2001). Neuronal basis of the motion aftereffect reconsidered. *Neuron*, *32*(1), 161–172. http://doi.org/10.1016/S0896-6273(01)00452-4

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2018). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *115*(1), E113–E114. http://doi.org/10.1073/pnas.1721071115

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. http://doi.org/10.1016/j.cognition.2011.08.001

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., & Campanella, S. (2004). When audition alters vision: An event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience Letters*, *369*(2), 132–137. http://doi.org/10.1016/j.neulet.2004.07.067

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*(3), 367–376. http://doi.org/10.1016/j.cortex.2010.03.003

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*(4), 2357–2364. http://doi.org/10.1016/j.neuroimage.2012.02.055

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). "Putting the face to the voice": Matching identity across modality. *Current Biology*, *13*(19), 1709–1714. http://doi.org/10.1016/j

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–11. http://doi.org/10.1098/Rstb.2006.1934

Kaul, C., Rees, G., & Ishai, A. (2011). The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain. *Frontiers in Human Neuroscience*, *4*(January), 1–12. http://doi.org/10.3389/fnhum.2010.00238

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces, 1–17.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *NeuroImage*, *37*(4), 1445–1456. http://doi.org/10.1016/j.neuroimage.2007.06.020

Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia*, *47*(14), 3059–3066. http://doi.org/10.1016/j.neuropsychologia.2009.07.001

Krekelberg, B., Boynton, G. M., & van Wezel, R. J. A. (2006). Adaptation: from single cells to BOLD signals. *Trends in Neurosciences*, *29*(5), 250–256. http://doi.org/10.1016/j.tins.2006.02.008

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(51), 20600–20605. http://doi.org/10.1073/pnas.0705654104

Kriegeskorte, N., Goebel, R., & Bandettini, P. P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868.

http://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. http://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 1–28. http://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. http://doi.org/10.1016/j.neuron.2008.10.043

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar Representations of Emotions Across Faces and Voices. *Emotion*, *17*(6), 912–937. http://doi.org/10.1037/emo0000282

Lachs, Lorin, Pisoni, D. B. (2004). Crossmodal Source Identification in Speech Perception Lorin. *Ecological Psychology : A Publication of the International Society for Ecological Psychology*, *16*(3), 159–187. http://doi.org/10.1207/s15326969eco1603

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour*, *75*(3), 817–822. http://doi.org/10.1016/j.anbehav.2007.07.001

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It’s not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 905–914. http://doi.org/10.1037/0096-1523.33.4.905

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*(JUL), 1–12. http://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex*, *21*(12), 2820–2828. http://doi.org/10.1093/cercor/bhr077

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080. http://doi.org/10.1016/j.cub.2013.04.055

Lattner, S., Meyer, M. E., & Friederici, A. D. (2005). Voice perception: Sex, pitch, and the right hemisphere. *Human Brain Mapping*, *24*(1), 11–20. http://doi.org/10.1002/hbm.20065

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. *Experimental Brain Research*, *166*(3–4), 289–297. http://doi.org/10.1007/s00221-005-2370-2

Lavan, N. (2017). Commentary: “Hearing faces and seeing voices”: Amodal coding of person identity in the human brain. *Frontiers in Neuroscience*, *11*(11), 303. http://doi.org/10.1038/nrn1931

Lavan, N., Burston, L. F. K., & Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception in unfamiliar listeners. *British Journal of Psychology*.

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2018). Flexible voices : identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 1–13. http://doi.org/10.3837/tiis.0000.00.000

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(12), 1604.

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. http://doi.org/10.1023/A:1009656816383

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. http://doi.org/10.1016/S0047-259X(03)00096-4

Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R., & Rao, S. M. (2000). Neural Systems Underlying the Recognition of Familiar and Newly Learned Faces. *The Journal of Neuroscience*, *20*(2), 878–886.

Liu, C. H., Bhuiyan, M. A. A., Ward, J., & Sui, J. (2009). Transfer Between Pose and Illumination Training in Face Recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 939–947.

http://doi.org/10.1037/a0013710

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*(10), 1386–1390. http://doi.org/10.1038/nn1538

Lundstrom, B. N., Petersson, K. M., Andersson, J., Johansson, M., Fransson, P., & Ingvar, M. (2003). Isolating the retrieval of imagined pictures during episodic memory: Activation of the left precuneus and left prefrontal cortex. *NeuroImage*, *20*(4), 1934–1943. http://doi.org/10.1016/j.neuroimage.2003.07.017

Margalit, E., Biederman, I., Herald, S. B., Yue, X., & von der Malsburg, C. (2016). An applet for the Gabor similarity scaling of the differences between complex stimuli. *Attention, Perception, & Psychophysics*, *78*(8), 2298–2306.

Mattavelli, G., Andrews, T. J., Asghar, A. U. R., Towler, J. R., & Young, A. W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia*, *50*(9), 2205–2211. http://doi.org/10.1016/j.neuropsychologia.2012.05.024

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say “hello”? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), 1–9. http://doi.org/10.1371/journal.pone.0090779

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: A meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, *8*(3), 285–299. http://doi.org/10.1093/scan/nsr090

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. http://doi.org/10.3389/fpsyg.2013.00128

Mur, M., Ruff, D. A., Bodurka, J., Bandettini, P. A., & Kriegeskorte, N. (2010). Face-identity change activation outside the face system: “Release from adaptation” may not always indicate neuronal selectivity. *Cerebral Cortex*, *20*(9), 2027–2042. http://doi.org/10.1093/cercor/bhp272

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., … Kojima, S. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia*, *39*(10), 1047–1054. http://doi.org/10.1016/S0028-3932(01)00037-9

Natu, V. S., Jiang, F., Narvekar, A., Keshvari, S., Blanz, V., & O'Toole, A. J. (2010). Dissociable Neural Patterns of Facial Identity across Changes in Viewpoint. *Journal of Cognitive Neuroscience*, *22*(7), 1570–1582. http://doi.org/10.1162/jocn.2009.21312

Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9998–10003. http://doi.org/10.1073/pnas.1102433108

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition*, *44*(3), 342–366. http://doi.org/10.1006/brcg.1999.1196

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). http://doi.org/10.1371/journal.pcbi.1003553

Nolan, F., Mcdougall, K., & Hudson, T. (2011). Some Acoustic Correlates of Perceived (Dis) Similarity Between Same-Accent Voices. In *International Congress of Phonetic Sciences (ICPhS)* (pp. 1506-1509).

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Reward and punishment representations in the human orbitofrontal cortex during emotion-related learning. *Nature Neuroscience*, *4*(1), 1–8.

O'Mahony, C., & Newell, F. N. (2012). Integration of faces and voices, but not faces and names, in person recognition. *British Journal of Psychology*, *103*(1), 73–82. http://doi.org/10.1111/j.2044-8295.2011.02044.x

O'Neil, E. B., Hutchison, R. M., McLean, D. A., & Köhler, S. (2014). Resting-state fMRI reveals functional connectivity between face-selective perirhinal cortex and the fusiform face area related to face inversion. *NeuroImage*, *92*, 349–355. http://doi.org/10.1016/j.neuroimage.2014.02.005

Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological Research*, *39*(1), 56–61. http://doi.org/10.1111/1468-5884.00037

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene:a holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: A review and theoretical framework. *Social Cognitive and Affective Neuroscience*, *8*(2), 123–133. http://doi.org/10.1093/scan/nss119

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092. http://doi.org/10.1073/pnas.0805664105

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition*, *170*(October 2017), 190–200. http://doi.org/10.1016/j.cognition.2017.10.002

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5), 1–7. http://doi.org/10.1038/s41562-017-0072

Parkinson, C., Liu, S., & Wheatley, T. (2014). A Common Cortical Metric for Spatial, Temporal, and Social Distance. *Journal of Neuroscience*, *34*(5), 1979–1987. http://doi.org/10.1523/JNEUROSCI.2159-13.2014

Peelen, M. V, Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(30), 10127–10134. http://doi.org/10.1523/JNEUROSCI.2161-10.2010

Pegado, F., Hendriks, M. H. A., Amelynck, S., Daniels, N., Bulthe, J., Lee, M. H., … Op de, B. H. (2018). A Multitude of Neural Representations Behind Multisensory “Social Norm” Processing. *Front Hum.Neurosci.*, *12*(1662–5161 (Linking)), 153. http://doi.org/10.3389/fnhum.2018.00153

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior*, *25*(4), 229–241. http://doi.org/10.1016/j.evolhumbehav.2004.04.003

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization, *3*, 23. http://doi.org/10.3389/fpsyg.2012.00023

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G. G., … Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, *119*, 164–174. http://doi.org/10.1016/j.neuroimage.2015.06.050

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2008). Neural Representations of Faces and Body Parts in Macaque and Human Cortex: A Comparative fMRI Study. *Journal of Neurophysiology*, *101*(5), 2581–2600. http://doi.org/10.1152/jn.91198.2008

Pitcher, D., Dilks, D. D., Saxe, R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, *56*(4), 2356–2363. http://doi.org/10.1016/j.neuroimage.2011.03.067

Poon, S., & Ng, M. L. (2011). Contribution of Voice Fundamental Frequency and Formants To the Identification of Speaker ’ S Gender. Proceedings of the *ICPhsXVII*, 1630–1633.

Pourtois, G., Schwartz, S., Seigher, M I., Lazeyras, F., Vuilleumer, P, G. (2005). Portraits or people? Distinct representation of facer identity in the human visual cortex. *Journal of Cognitive Neuroscience*, *17*(7), 1043–1057.

Pyles, J. A., Verstynen, T. D., Schneider, W., & Tarr, M. J. (2013). Explicating the Face Perception Network with White Matter Connectivity. *PLoS ONE*, *8*(4), 1–12. http://doi.org/10.1371/journal.pone.0061611

Quiroga, R. Q., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain. *Current Biology*, *19*(15), 1308–1313. http://doi.org/10.1016/j.cub.2009.06.060

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. http://doi.org/10.1038/nature03687

Rajimehr, R., Young, J. C., & Tootell, R. B. H. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 1995–2000. http://doi.org/10.1073/pnas.0807304106

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant Voices and Attractive Faces: The Contribution of Visual and

Auditory Information to Integrated Person Impressions. *Journal of Nonverbal Behavior*, *39*(4), 355–370. http://doi.org/10.1007/s10919-015-0214-8

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, *19*(4), 473–497. http://doi.org/10.1016/0010-0285(87)90016-8

Ritchie, K. L., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports*, *7*(1), 1–8. http://doi.org/10.1038/s41598-017-00526-9

Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain and Cognition*, *69*(2), 269–278. http://doi.org/10.1016/j.bandc.2008.08.007

Robinson, K., Blais, C., Duncan, J., Forget, H., & Fiset, D. (2014). The dual nature of the human face: There is a little Jekyll and a little Hyde in all of us. *Frontiers in Psychology*, *5*(MAR), 1–9. http://doi.org/10.3389/fpsyg.2014.00139

Roswandowitz, C., Kappes, C., Obrig, H., & von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. *Brain*, *141*(1), 234–247. http://doi.org/10.1093/brain/awx313

Rotshtein, P., Henson, R. N. R. N. A., Treves, A., Driver, J., & Dolan, R. J. R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci*, *8*(1), 107–113. http://doi.org/10.1038/nn1370

Saarimäki, H., Ejtehadian, L. F., Glerean, E., Jääskeläinen, I. P., Vuilleumier, P., Sams, M., & Nummenmaa, L. (2018). Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, *13*(5), 471–482. http://doi.org/10.1093/scan/nsy018

Saarimaki, H., Gotsopoulos, A., Jaaskelainen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., … Nummenmaa, L. (2015). Discrete Neural Signatures of Basic Emotions. *Cerebral Cortex*, *26*(6), 1–11. http://doi.org/10.1093/cercor/bhv086

Said, C. P., Baron, S. G., & Todorov, A. (2009). Nonlinear Amygdala Reponse to Face Trustworthiness: Contributions of High and Low Spatial Frequency Information. *Journal of Cognitive Neuroscience*, *21*(3), 519–528.

Said, C. P., Dotsch, R., & Todorov, A. (2011). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, *49*(4), 630–639.

http://doi.org/10.1016/j.neuropsychologia.2011.02.028

Said, C. P., Moore, C. D., Engell, A. D., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, *10*(2010), 1–12. http://doi.org/10.1167/10.5.11.Introduction

Santos, S., Almeida, I., Oliveiros, B., & Castelo-Branco, M. (2016). The role of the amygdala in facial trustworthiness processing: A systematic review and meta-analyses of fMRI studies. *PLoS ONE*, *11*(11), 1–28. http://doi.org/10.1371/journal.pone.0167276

Sapountzis, P., Schluppeck, D., Bowtell, R., & Peirce, J. W. (2010). A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. *NeuroImage*, *49*(2), 1632–1640. http://doi.org/10.1016/j.neuroimage.2009.09.066

Saxton, T. K., Burriss, R. P., Murray, A. K., Rowland, H. M., & Craig Roberts, S. (2009). Face, body and speech cues independently predict judgments of attractiveness. *Journal of Evolutionary Psychology*, *7*(1), 23–35. http://doi.org/10.1556/JEP.7.2009.1.4

Schall, S., Kiebel, S. J., Maess, B., & Von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage*, *77*, 237–245. http://doi.org/10.1016/j.neuroimage.2013.03.043

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing Famous Voices. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453–463. http://doi.org/10.1017/CBO9781107415324.004

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *50*(3), 498–517. http://doi.org/10.1080/027249897391991

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: A Journal of Neurology*, *124*, 804–815. http://doi.org/10.1093/brain/124.4.804

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Concordant cues in faces and voices: Testing the Backup Signal Hypothesis. *Evolutionary Psychology*, *14*(1), 1–10. http://doi.org/10.1177/1474704916630317

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, *78*(3), 868–879. http://doi.org/10.3758/s13414-015-1045-8

Sokhi, D. S., Hunter, M. D., Wilkinson, I. D., & Woodruff, P. W. R. (2005). Male and female voices activate distinct regions in the male brain. *NeuroImage*, *27*(3), 572–578. http://doi.org/10.1016/j.neuroimage.2005.04.023

Sormaz, M., Watson, D. M., Smith, W. A. P., Young, A. W., & Andrews, T. J. (2016). Modelling the perceptual similarity of facial expressions from image statistics and neural responses. *NeuroImage*, *129*, 64–71. http://doi.org/10.1016/j.neuroimage.2016.01.041

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. http://doi.org/10.1073/pnas.1014345108

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*(4), 409–419. http://doi.org/10.1080/20445911.2011.642859

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22*(3), 197–200. http://doi.org/10.1016/j.tics.2017.12.003

Sugiura, M., Mano, Y., Sasaki, A., & Sadato, N. (2011). Beyond the memory mechanism: Person-selective and nonselective processes in recognition of personally familiar faces. *Journal of Cognitive Neuroscience*, *23*(3), 699–715. http://doi.org/10.1162/jocn.2010.21469

Sugiura, M., Shah, N. J., Zilles, K., & Fink, G. R. (2005). Cortical Representations of Personally Familiar Objects and Places : Functional Organization of the Human Posterior Cingulate Cortex, 183–198.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. http://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397–415. http://doi.org/10.1111/bjop.12206

Thornhill, R., & Moller, A. P. (1997). Developmental stability, disease and medicine. *Biological Reviews*, *72*(4), 497–548. http://doi.org/10.1111/j.1469-185X.1997.tb00022.x

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707. http://doi.org/10.1121/1.397959

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness : a model based approach. http://doi.org/10.1093/scan/nsn009

Todorov, A., & Engell, A. D. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social Cognitive and Affective Neuroscience*, *3*(4), 303–312. http://doi.org/10.1093/scan/nsn033

Todorov, A., & Oosterhof, N. N. (2011). Modeling Social Perception of Faces. *IEEE Signal Processing Magazine*, (March), 117–122.

Todorov, A., & Porter, J. M. (2014). Misleading First Impressions: Different for Different Facial Images of the Same Person. *Psychological Science*, *25*(7), 1404–1417. http://doi.org/10.1177/0956797614532474

Todorov, A., Said, C. P., Oosterhof, N. N., & Engell, A. D. (2011). Task-invariant Brain Responses to the Social Value of Faces. *Journal of Cognitive Neuroscience*, *23*(10), 2766–2781. http://doi.org/10.1162/jocn.2011.21616

Troje, N. F., & Bulthoff, H. H. (1996). Face recognition under varying pose: The role of texture and shape. *Vision Research*, *36*, 1761–1771. Retrieved from Text/TrojeBuelthoff96.pdf

Turk-Browne, N. B., Norman-Haignere, S. V., & McCarthy, G. (2010). Face-Specific Resting Functional Connectivity between the Fusiform Gyrus and Posterior Superior Temporal Sulcus. *Frontiers in Human Neuroscience*, *4*(September), 1–15. http://doi.org/10.3389/fnhum.2010.00176

Valentine, T. (1991). A unified account of the effect of distinctivness, inversion and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161–204. http://doi.org/10.1080/14640749108400966

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*(5), 525–535.

Valentova, J. V., Varella, M. A. C., Havlícek, J., Kleisner, K., Havlíček, J., & Kleisner, K. (2017). Positive association between vocal and facial attractiveness in women but not in men: A cross-cultural study. *Behavioural Processes*, *135*, 95–100. http://doi.org/10.1016/j.beproc.2016.12.005

Van Lancker, D., Krieman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics*, *13*.

Van Valen, L. (1962). A study of fluctuating asymmetry. *Evolution*, *17*, 125–142.

Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, *51*(11), 2100–2108. http://doi.org/10.1016/j.neuropsychologia.2013.07.006

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. *Scientific Reports*, *7*(1), 1–14. http://doi.org/10.1038/s41598-017-12559-1

von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.-L., Kell, C. a, Grüter, T., … Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6747–6752. http://doi.org/10.1073/pnas.0710826105

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), 1809–1820. http://doi.org/10.1371/journal.pbio.0040326

Von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948–955. http://doi.org/10.1016/j.neuroimage.2004.02.020

von Kriegstein, K., Kleinschmidt, A., & Giraud, A. L. (2006). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, *16*(9), 1314–1322. http://doi.org/10.1093/cercor/bhj073

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive*

Neuroscience, *17*(3), 367–76. http://doi.org/10.1162/0898929053279577

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Ives, D. T., & Griffiths, T. D. (2007). Neural Representation of Auditory Size in the Human Voice and in Sounds from Other Resonant Sources. *Current Biology*, *17*(13), 1123–1128. http://doi.org/10.1016/j.cub.2007.05.061

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the Human Brain Recognizes Speech in the Context of Changing Speakers. *Journal of Neuroscience*, *30*(2), 629–638. http://doi.org/10.1523/JNEUROSCI.2742-09.2010

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*(February), 31–39. http://doi.org/10.1016/j.neuropsychologia.2018.02.023

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*(0), 188–200. http://doi.org/10.1016/j.neuroimage.2015.12.012

Wang, Y., Collins, J. A., Koski, J., Nugiel, T., Metoki, A., & Olson, I. R. (2017). Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences*, *114*(16), E3305–E3314. http://doi.org/10.1073/pnas.1621234114

Watkins, C. D., Jones, B. C., & DeBruine, L. M. (2010). Individual differences in dominance perception: Dominant men are less sensitive to facial cues of male dominance. *Personality and Individual Differences*, *49*(8), 967–971. http://doi.org/10.1016/j.paid.2010.08.006

Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, *50*, 125–136. http://doi.org/10.1016/j.cortex.2013.07.011

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. *The Journal of Neuroscience*, *34*(20), 6813–6821. http://doi.org/10.1523/JNEUROSCI.4478-13.2014

Weibert, K., Flack, T. R., Young, A. W., & Andrews, T. J. (2018). Patterns of neural response in face regions are predicted by low-level image properties. *Cortex*, *103*, 199–210. http://doi.org/10.1016/j.cortex.2018.03.009

Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior*, *42*(5), 805–811. http://doi.org/10.1007/s10508-012-0054-0

Weston, P. S. J., Hunter, M. D., Sokhi, D. S., Wilkinson, I. D., & Woodruff, P. W. R. (2015). Discrimination of voice gender in the human auditory cortex. *NeuroImage*, *105*, 208–214. http://doi.org/10.1016/j.neuroimage.2014.10.056

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.

Winston, J. S., Henson, R. N. A., & Dolan, R. J. (2004). fMRI-Adaptation Reveals Dissociable Neural Representations of Identity and Expression in Face Perception. *Journal of Neurophysiology*, *92*(3), 1830–1839. http://doi.org/10.1152/jn.00155.2004

Winston, J. S., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness. *Neuropsychologia*, *45*(1), 195–206. http://doi.org/10.1016/j.neuropsychologia.2006.05.009

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*, 277–283. http://doi.org/10.1038/nn816

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*(10), 1034–1043. http://doi.org/10.1093/cercor/13.10.1034

Xu, X., & Biederman, I. (2010). Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint. *Journal of Vision*, *10*(14), 36–36. http://doi.org/10.1167/10.14.36

Xu, X., Yue, X., Lescroart, M. D., Biederman, I., & Kim, J. G. (2009). Adaptation in the fusiform face area (FFA): Image or person? *Vision Research*, *49*(23), 2800–2807. http://doi.org/10.1016/j.visres.2009.08.021

Yang, H., Susilo, T., & Duchaine, B. (2016). The Anterior Temporal Face Area Contains Invariant Representations of Face Identity That Can Persist Despite the Loss of Right FFA and OFA. *Cerebral Cortex*, *26*(3), 1096–1107. http://doi.org/10.1093/cercor/bhu289

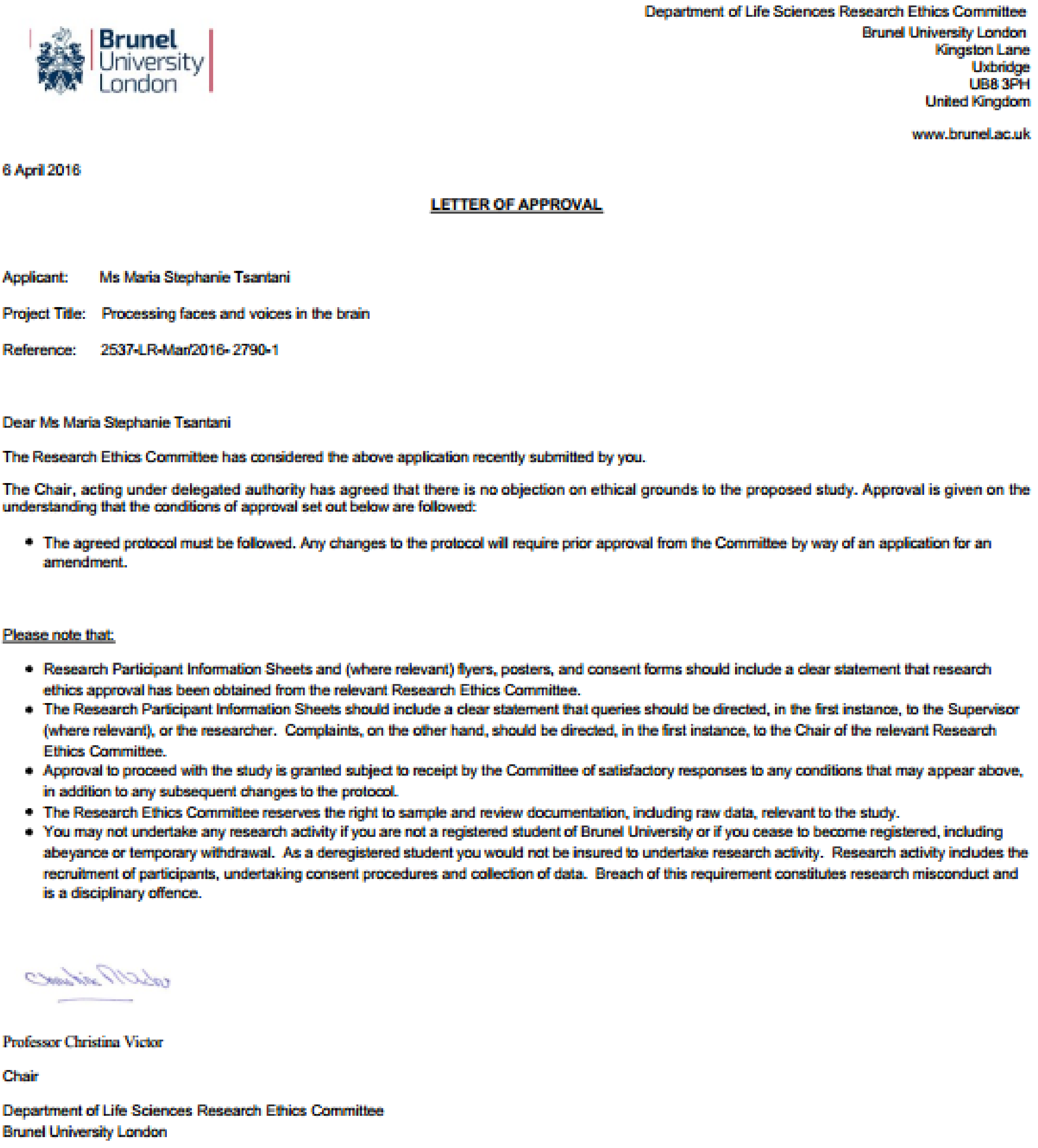
Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. http://doi.org/10.1016/j.tics.2013.04.004

Yovel, G., & O’Toole, A. J. (2016). Recognizing People in Motion. *Trends in Cognitive Sciences*, *20*(5), 383–395. http://doi.org/10.1016/j.tics.2016.02.005

Yue, X., Biederman, I., Mangini, M. C., Malsburg, C. von der, & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research*, *55*, 41–46. http://doi.org/10.1016/j.visres.2011.12.012

Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, *13*(2), 67–82. http://doi.org/10.1007/BF00990791

Appendix A – Ethical approval for the studies presented in Chapters 3 & 4



Department of Life Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

6 April 2016

LETTER OF APPROVAL

Applicant: Ms Maria Stephanie Tsantani

Project Title: Processing faces and voices in the brain

Reference: 2537-LR-Mar/2016- 2790-1

Dear Ms Maria Stephanie Tsantani

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed that there is no objection on ethical grounds to the proposed study. Approval is given on the understanding that the conditions of approval set out below are followed:

- The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.

Please note that:

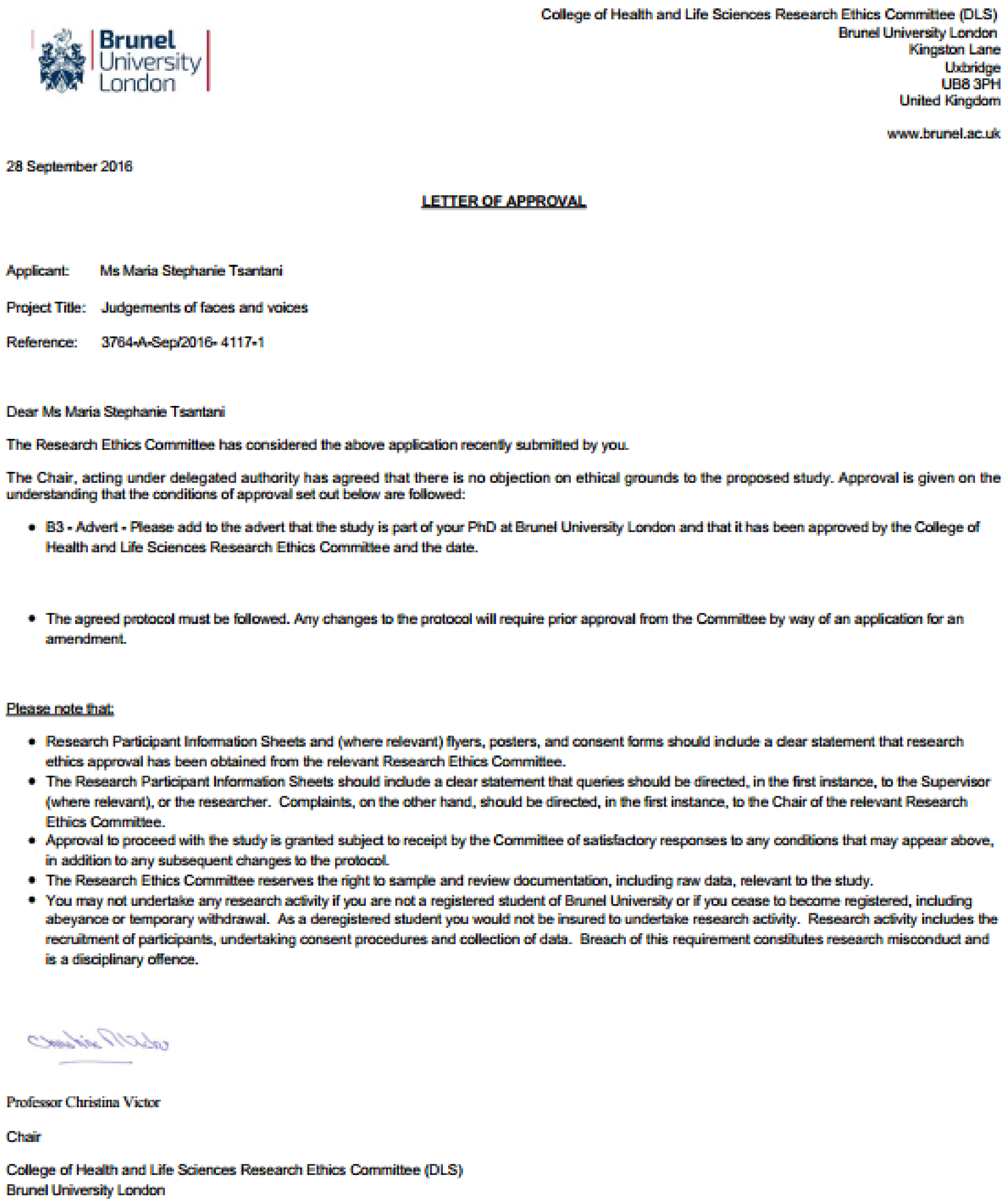
- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to receipt by the Committee of satisfactory responses to any conditions that may appear above, in addition to any subsequent changes to the protocol.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Professor Christina Victor

Chair

Department of Life Sciences Research Ethics Committee
Brunel University London

Appendix B – Ethical approval for the study presented in Chapter 5



College of Health and Life Sciences Research Ethics Committee (DLS)
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

28 September 2016

LETTER OF APPROVAL

Applicant: Ms Maria Stephanie Tsantani

Project Title: Judgements of faces and voices

Reference: 3764-A-Sep/2016- 4117-1

Dear Ms Maria Stephanie Tsantani

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed that there is no objection on ethical grounds to the proposed study. Approval is given on the understanding that the conditions of approval set out below are followed:

- B3 - Advert - Please add to the advert that the study is part of your PhD at Brunel University London and that it has been approved by the College of Health and Life Sciences Research Ethics Committee and the date.

- The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.

Please note that:

- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to receipt by the Committee of satisfactory responses to any conditions that may appear above, in addition to any subsequent changes to the protocol.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Professor Christina Victor

Chair

College of Health and Life Sciences Research Ethics Committee (DLS)
Brunel University London