Representations of Native and Foreign Talkers in Brain and Behaviour

David Fleming

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

School of Psychology
College of Science and Engineering
University of Glasgow

September 2015

Abstract

Human listeners possess good speaker recognition abilities, and are capable of discriminating and identifying speakers from a range of spoken utterances. However, voice recognition can be enhanced when a listener is capable of understanding the speech produced by a talker. A well-established demonstration of this is known as the "Language-Familiarity" Effect (LFE) for voice recognition. This effect manifests as an impairment for voice recognition in foreign language speech conditions, as contrasted with recognition of talkers who are speaking in a listener's mother tongue, and has been repeatedly demonstrated across a range of different tasks and languages. The LFE has previously been conceptualized as an analogue to the even better-known "Other-Race" Effect (ORE) for face recognition, where own-race faces are better remembered than other-race faces. An influential theoretical model of the ORE posits that faces are represented in a multidimensional "face-space", whose dimensions are shaped by perceptual experience and code for features which are diagnostic for face individuation (Valentine, 1991). Over the course of an individual's perceptual experience, these dimensions might become attuned for own-race face recognition; as a consequence, the dimensions will be sub-optimal for other-race recognition, leading to the illusion of increased similarity among different other-race faces, relative to own-race faces – what has been termed the "they-all-look-alike" effect. The idea of a complementary "voice-space" has already been posited in the auditory domain, and might serve as a useful model for the LFE. Speakers might be individuated on the basis of diagnostic dimensions which might code for important voice-acoustical attributes. However, these dimensions might also be shaped according to linguistic experience, and voice individuation (and recognition) might be optimised when listeners can take advantage of both general voice acoustics and stored representations of their native language to tell speakers apart. The

face-space hypothesis represents a plausible model for the ORE, and evidence for it has accrued through computational modelling and neuroimaging work. Conversely, however, at present it merely serves as a descriptive model for the LFE. In this thesis, I combine behavioural testing, and neuroimaging studies using functional Magnetic Resonance Imaging (fMRI) to probe the nature of the representations of native and foreign speakers.

Chapter 1 provides a general overview of voice processing with an emphasis on voice recognition. Subsequently, I provide a review of relevant literature pertaining to the LFE, and introduce a brief comparison to the ORE for faces in the context of the Valentine (1991) similarity model, ending with a description of the aims of the thesis. In **Chapter 2,** I present the results of a behavioural experiment where native English and Mandarin speaking listeners rated all pairwise combinations of a series of English- and Mandarin-speaking voices. Crucially, the LFE does not appear to be dependent on full comprehension of the linguistic message, as young infants can better tell apart speakers in their native language than in a foreign language before their speech comprehension abilities are fully mature. This suggests that exposure to the sound-structure characteristic of infants' nascent mother tongue might be sufficient to enhance native language speaker discrimination, in the absence of full comprehension. Therefore, to examine a counterpart in adults, speech stimuli were subjected to time-reversal, a process which precludes lexical and semantic access but which leaves intact certain phonemic properties of the original speech signal. Both the English and Mandarin listeners rated pairs of native-language voices as sounding more dissimilar than foreign voices, suggesting that the language-specific sound-structure elements remaining in the reversed speech enabled an enhanced individuation of native voices. Next, in **Chapter 3,** I aimed to probe the neural basis of this enhanced individuation in an fMRI experiment

which was intended to capture dissimilarities among paired cerebral responses to unintelligible native and foreign speakers. Here, I did not find a direct correlate of the behavioural effect, but did find that local patterns of response estimates in the bilateral superior temporal cortex (STC) appear to "discriminate" the different language categories in both English and Mandarin listeners. Specifically, when the pairwise dissimilarity in brain responses to different speakers was collected, relatively high dissimilarity was observed for pairs consisting of a response to an English speaker and a Mandarin speaker, whereas relatively low dissimilarity was observed for pairs consisting of two English or two Mandarin speakers. In **Chapter 4**, I report what is, to my knowledge, the first explicit examination of the neural basis for the LFE in intelligible speech. A monolingual sample of English speakers participated in an fMRI experiment where they listened to the voices of English and Mandarin speakers. Importantly, speech stimuli in both language conditions were matched in inter-speaker acoustical variability. Combined response patterns from bilateral voice-sensitive temporal lobe regions enabled a learning algorithm to decode the identities of the voices who elicited the responses, but, crucially, only in the native speech (English) condition. Interestingly, native-language speaker decoding was also achieved from a left-hemisphere voice-sensitive region alone, but not a right-hemisphere region. This putative leftward bias might reflect a higher discriminability of native-language talkers in the brain, via an enhanced ability to individuate voices on the basis of indexical variation around stored speech-sound representations. Finally, in **Chapter 5,** I conclude with a general discussion of the foregoing results, their implications for an analogous conception of the LFE and ORE, and some strands of thought for future investigation.

Table of Contents

Abstract 2

List of Tables 8

List of Figures 9

Acknowledgements 10

Author's Declaration 12

1. General Introduction 14

1.1. A specialized system for a very special sound: cerebral processing of the human voice. 16

1.2. Why is the voice treated specially – is it like an auditory 'face'? 18

1.3. "Who said that?" The nature of human voice recognition 19

1.3.1. What sounds are important 19

1.3.2. Where does voice identity processing "happen" in the brain? 21

1.3.3. Interactions of "speech" and "voice" systems: Voice recognition is not all about "how" you sound, but also "what" you say. 24

1.4. The Language Familiarity Effect (LFE) in voice recognition 26

1.4.1. Is the LFE attenuated by the degree of phonological 'overlap' between native and foreign speech? 28

1.4.2. Can the LFE be ameliorated with foreign language experience? 30

1.4.3. At what developmental stage does the LFE appear? 33

1.4.4. The LFE: Interim summary 35

1.4.5. The LFE as an auditory analogue to the ORE: the Valentine model 36

1.5. Aims of this thesis 41

2. A language-familiarity effect for speaker discrimination without comprehension. 45

2.1. Introduction 46

2.2. Methods 49

2.3. Results 52

2.4. Discussion 53

2.5 Supplementary Information 59

3. “Discrimination” of unintelligible language categories in superior temporal cortices 62

3.1. Introduction 63

3.2. Methods 69

3.3. Results 76

3.4. Discussion 81

3.5. Conclusion 89

3.6. Supplementary Information 90

4. Brain-based speaker identity decoding in a native and a foreign language 95

4.1. Introduction 96

4.2. Methods 99

4.3. Results 113

4.4. Discussion 118

4.5. Conclusion 126

4.6. Supplementary Information 127

5. General Discussion 129

5.1. The LFE as an analogue to the ORE? 131

5.2. Limitations 137

5.3. Future Directions 141

5.4. General Conclusion 144

6. List of References 145

List of Tables

Table 2.5.1: Table of Acoustics 59

Table 3.1: Cluster peaks and sub-peaks from the contrast of the general effect of reversed speech against scanner-noise baseline (regardless of language) collapsed across English and Mandarin-speaking participant groups 77

Table 3.2. Peaks of activation differences derived from the contrast of Mandarin speech > English speech, collapsed across Mandarin and English-speaking participants. 79

Table 3.3. Results of second level (RFX) analysis revealing areas where correlations between local multivariate patterns and the binary 'language-separation' model significantly differ from zero, across both participant groups. 81

Table 3.6.3. Peaks of activation differences derived from the contrast of Mandarin speech > English speech, (collapsed across Mandarin and English-speaking participants) following removal of the influence of stimulus f0SD values 93

Table 4.1: Values of acoustical attributes extracted from the final set of English and Mandarin speech stimuli. 102

Table 4.2: Results of F-tests comparing the level of within-group acoustical variance across the two stimulus sets, following length and amplitude normalization. 103

Table 4.3: Summary table of peak foci from group analysis of the independent voice-localizer scans 116

Table 4.4: Summary of classification results from voice-sensitive ROIs in participants' native image space. 118

List of Figures

Figure 2.1. Speaker dissimilarity ratings for pairs of Mandarin and English time-reversed sentences. 51

Figure 3.1.Illustration of predictor models used in the Representational Similarity Analysis (RSA). 68

Figure 3.2 Thresholded cerebral responses to all reversed speech sounds against scanner-noise baseline averaged over listener groups. 76

Figure 3.3. Results of the univariate contrast of Mandarin reversed speech versus English reversed speech. 78

Figure 3.4 Extent of significant correlations between cerebral dissimilarities and the "language-separation" predictor model (red). 80

Figure 4. 1: Schematic of training procedure used during behavioural testing and slow event-related fMRI protocol. 105

Figure 4.2: Behavioural identification performance from pre-scan voice training 114

Figure 4.3. Voice-sensitive temporal lobe regions and spherical ROIs around activation peaks. 117

Figure 4.4: Voice sensitive ROIs rendered on a template brain. 119

Figure 4.5: Three-way decoding results from voice-sensitive ROIs. 121

Figure 4.6.1: Spherical ROIs based on group-level voice-sensitive peaks. 127

Acknowledgements

Many people have made important contributions to the work contained in this thesis. Firstly, I owe an enormous debt of gratitude to my supervisory team, Pascal Belin and Bruno Giordano. Pascal, your mentorship over the years has been invaluable to me. You have encouraged me to follow my own ideas, and nurtured my curiosity while lending a critical eye at crucial times. I also owe you thanks for welcoming me into your lab as an undergraduate student, back in the days when I was still harbouring designs on moving into research on face perception. On this note too, I must thank Roberto Caldara for first introducing me to Pascal and for his continued interest in my work, despite my abandonment of faces!

Bruno, I will never forget the time you have invested in assisting me. Without any obligation to do so, you have spent countless hours analysing data with me, discussing experimental designs and statistics, and, generally, providing excellent input and friendship.

On the technical side, I must thank Frances Crabbe, for her patient assistance in preparing scanning sequences and in training me to operate the MRI system in Glasgow. Without her, I am certain that my research would have progressed much more slowly. My research would certainly have further stalled were it not for Bo Yao and Jingting Zhang, to whom I am indebted for their assistance in preparing Mandarin stimulus materials.

I have been very fortunate to conduct my research in an institute which hosts so many talented, friendly individuals. I am even more fortunate to count a few of these

individuals among my good friends. I am particularly grateful for the friendships I have made with Chris Benwell, Oliver Garrod and Phil McAleer. I will look back fondly on the stimulating academic discussions we've had over the years; however, I think it's fair to say that I will treasure more the memories of the lengthy discussions we've had about sport, politics, cinema, and music (among other things) on those long nights in Brel and Oran Mor. You have all, at one time or another, provided valuable input to my work. More importantly, you've consistently provided sterling company, even when we've reached one of our frequent disagreements! The School of Psychology aside, I owe thanks to my close friends Steven Forrest and Stuart Ross. I am tremendously appreciative of the time we've spent together over the last few years, discussing everything and anything but work!

Lastly, and most importantly, I thank my parents, my grandparents, and my partner, Manon. I will be eternally grateful for the love and support with which you have all provided me over the course of my life and studies. In particular, Manon, you have been my rock during the turbulent writing-up period, tolerant beyond reason of my fits of anxiety and doubt. I love you so much, and hope someday to be able to pay you back for all that you have done for me.

Author's Declaration

I certify that this dissertation is my original work and that all references to the work of others have been clearly identified and fully attributed.

Chapter 2 is based on previously published work:

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). **A language-familiarity effect for speaker discrimination without comprehension.** *Proceedings of the National Academy of Sciences*, *111*(38), 13795–13798.

Chapter 1: General Introduction

Imagine you are walking down a street in London, on your way to meet a friend at a cafe. You enter the cafe, and, at first glance, you cannot see your friend. Then, suddenly, you hear a familiar cough. “Aha!” you think. “That sounds rather like John, he must have already arrived.” You are confident that you have heard John’s voice, but are still a little uncertain, until you hear him place his order for coffee. Both you and John are native English speakers from the UK, and you have just heard John uttering familiar speech in your common native language. Despite the murmurs of conversation coming from other patrons, you are now certain that John is present in the cafe, and you follow the sound of his voice to his table.

Next, imagine a similar situation in different surroundings. On this occasion, you are on vacation in Beijing, and you have arranged to meet a Chinese friend, Dongmei, in a cafe. You are a monolingual native English speaker and, while you have heard Mandarin spoken, you have no functional knowledge of the language. In other words you are largely ‘deaf’ to Mandarin speech – you can’t really tell where one word ends and where another begins. As before, you hear a familiar cough upon entering the cafe. Your ears prick up, and, in the absence of a view of Dongmei, you surmise that she has already arrived. Now, you hear some Mandarin speech, and you are reasonably confident that this speech was produced by the coughing voice from before. However, you are confused; Mandarin conversations are being held at other tables and, to you, many of the other female voices sound rather similar to Dongmei’s. Was it really her voice that you heard coughing and ordering? If you heard her speaking in English, as in all of your previous interactions, you

feel that you would doubtlessly have confirmed the voice you heard as hers by now. Alas, on this occasion, you may need to hear more speech from the voice, or confirm Dongmei's presence visually.

The toy examples above describe a phenomenon known as the "Language Familiarity Effect". We are quite capable of recognizing the voices of those familiar to us, or even telling apart different unfamiliar voices; indeed, sometimes a familiar laugh, cry or cough can be enough to evoke a memory trace of a well-known person. Our ability to recognize voices is enhanced when we hear those voices uttering familiar sounds, such as speech in our maternal language. When we cannot understand the speech, as in the case of foreign language speakers, then we may be less successful in later identifying a voice than if it had spoken to us in our first language.

In this thesis, I will attempt to address some outstanding questions relating to the "Language Familiarity Effect" (or LFE). For example, despite sustained interest from the cognitive neuroimaging community in the neural substrates of speech and voice perception, relatively little consideration has been granted to the brain basis of the LFE. Additionally, it is not clear whether comprehension of the spoken message is necessary *per se* for a listener to show a native-language advantage in voice discrimination. For example, it may be that exposure to the sound elements of one's native language, embedded within an unintelligible speech signal may be sufficient to engender such a bias. Over the course of later chapters, these issues will be studied. In this opening chapter, I will begin with a brief overview of general voice processing. Thereafter, I will briefly discuss general voice recognition, before providing a review of the existing literature on the LFE.

1.1. A specialized system for a very special sound: cerebral processing of the human voice

Voices are arguably the most important sounds in our environment, and seem to be treated as such by the auditory system. Just as dedicated patches of the extended visual cortex appear to be more responsive to human faces or bodies than other classes of visual objects (Downing, Jiang, Shuman, & Kanwisher, 2001; Kanwisher, McDermott, & Chun, 1997), areas within the temporal lobe seem to be more sensitive to human vocal sounds than to other types of environmental sounds, including the vocal sounds of other species. These *Temporal Voice Areas* (TVA) have been consistently identified in neuroimaging studies, using functional Magnetic Resonance Imaging (fMRI) in particular (Ahrens, Awwad Shiekh Hasan, Giordano, & Belin, 2014; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Belin, Zatorre, & Ahad, 2002; Bethmann & Brechmann, 2014; Bethmann, Scheich, & Brechmann, 2012; Pernet et al., 2015). The TVA are located within the bilateral superior temporal sulcus (STS) and superior temporal gyrus (STG) of the temporal lobe, although generally with a right hemisphere preponderance. A recent analysis of voice-sensitive regions in a very large testing cohort concluded that the TVA appear to be organized in three spatially distinct clusters along the antero-posterior axis of the superior temporal cortex (STC; Pernet et al., 2015). Additionally, while fMRI investigations have well-characterized the spatial extent of these voice-sensitive areas, electrophysiological studies using electroencephalography (EEG) have also revealed the temporal characteristics of neural voice processing. For example, a Voice-Sensitive Response (VSR) has been found to occur at around 300-400 milliseconds following the presentation of a human voice stimulus (Levy, Granot, & Bentin, 2001; Levy, Granot, & Bentin, 2003). More recently, a fronto-temporal positivity to voice (FTPV) has been

reported at latencies of 100-200ms, over fronto-temporal electrodes (Charest et al., 2009). A complementary finding from Magnetoencephalography (MEG) research has revealed a magnetic counterpart to this response, which has been referred to as the FTPVm (Capilla, Belin, & Gross, 2013). The source of this event-related field was localized to the middle portion of the bilateral STG and STS, overlapping with the classical spatial location of the TVA. The temporal characteristics of these electro- and magnetoencephalographic responses suggest that voices may be processed in a similar time-window to faces, given the similarity in FTPV/m latency to that of the famous N170 event-related negativity elicited by face stimuli (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Rossion & Jacques, 2008).

Sensitivity to voices is present from a very early point in development; indeed, foetal heart rate appears to increase in response to the mother's voice as compared to a stranger's voice (Kisilevsky et al., 2003) and an electrophysiological index of voice-familiarity is present in new-born infants who are only a few hours old (Beauchemin et al., 2011). Later, 7-month-olds (but not 4-month-olds) show an increased response in bilateral superior temporal cortex (STC) to human vocal sounds as measured with Near-InfraRed-Spectroscopy (NIRS), compared to other sounds (Grossmann, Oberecker, Koch, & Friederici, 2010). A further report of early cerebral sensitivity to voice in a cohort of 3-7-month-old infants found a spatial activation profile in the temporal lobe which is similar to that found in the adult TVA (Blasi et al., 2011). Taken together, these findings indicate that voice-sensitivity appears very early along the developmental trajectory, as soon as the immediate ante-natal period (Beauchemin et al., 2011) or perhaps even prior to birth, during foetal gestation (Kisilevsky et al., 2003). Cerebral voice-sensitivity also has a rich evolutionary history: voice-preferential areas have been identified in the brains of

macaques (Perrodin, Kayser, Logothetis, & Petkov, 2011; Petkov et al., 2008)and even domestic dogs (Andics, Gácsi, Faragó, Kis, & Miklósi, 2014).

1.2. Why is the voice treated specially – is it like an auditory 'face'?

Why should voices be accorded such a privileged status by our nervous system, and the nervous systems of our evolutionary relatives? In addition to carrying speech, the main human medium for communication, voices bear a rich trove of information which the listener can utilize to form an impression of a speaker. We are capable of approximately determining a speaker's gender, age, weight, and emotional state even on the basis of very short vocal utterances (Belin, Fecteau, & Bédard, 2004). Given more information, such as running speech or single words, we can tell which nation or city a speaker comes from through their accent, and may even perceive clues as to their social background. Remarkably, listeners even form consistent personality judgements (Klofstad, Anderson, & Peters, 2012; McAleer, Todorov, & Belin, 2014), voting preferences (Tigue, Borak, O'Connor, Schandl, & Feinberg, 2012) and impressions of personal attractiveness (Bruckert et al., 2010), all based on short utterances. Thus, the voice has come to be regarded by some as an "auditory face", for, as with the face, a specialized cerebral architecture seems to exist for the processing of voice, and both social stimuli convey a broad tapestry of information about the bearer (Belin et al., 2004; Yovel & Belin, 2013). As such, recent theories and empirical works emphasize similar steps in the encoding and recognition of both.

Belin and colleagues (Belin, Bestelmeyer, Latinus, & Watson, 2011; Belin et al., 2004) adapted the influential face-processing model of Bruce and Young (Bruce & Young, 1986)

for voice, emphasizing a similar workflow: as vocal sounds enter the brain, they are first interrogated in a low-level structural analysis phase, which may be executed in sub-cortical structures, followed by primary auditory cortex. Thereafter, interacting streams in secondary auditory cortices and beyond may exist for the processing of the main three types of vocal information: affect, identity and speech. We shall remain with the second of these: voice identity information.

1.3. "Who said that?" The nature of human voice recognition

1.3.1. What sounds are important?

Vocal sounds are generated within an extended apparatus consisting of the lungs, the larynx (specifically, the glottis, consisting of the vocal folds and the opening between them), the pharynx, and the oral and nasal cavities. According to the "source-filter" theory, the production of a vocal sound involves two independent processes: the sound energy produced by air passing through the vocal cords as they open and close is referred to as the glottal source, and this sound energy is subsequently filtered by the supra-laryngeal vocal tract which introduces resonances referred to as formants. Filter characteristics are varied by changing the size and shape of the vocal tract (e.g., through movements of the lips, tongue and jaw).

Insights as to how the information carried by this apparatus may be used for the purposes of identifying or discriminating different speakers are presented in experiments investigating listeners' subjective judgements of the similarity of different talkers. Results from these studies suggest that the fundamental frequency (f0) or pitch of a voice

(corresponding to the laryngeal source), and its formant frequencies (the resonances generated in the supra-laryngeal vocal tract, which characterise vowel quality) play an important role in voice differentiation (Baumann & Belin, 2010; Creel & Bregman, 2011; Matsumoto, Hiki, Sone, & Nimura, 1973), a finding borne out by inter-speaker variability in acoustical measurements (Hanson, 1997). More recently, a three-dimensional space for the representation of voices was proposed, which included f0, the average distance between successive formants (known as formant dispersion) and Harmonics-to-Noise ratio (HNR), a measure of the "smoothness" of a voice (Latinus, McAleer, Bestelmeyer, & Belin, 2013). Variability in these vocal acoustic properties could, therefore, be important for individuating different voices.

Listeners are also capable of generalizing recognition to unlearned utterances from familiar speakers (e.g., Perrachione & Wong, 2007; Sheffert, Pisoni, Fellowes, & Remez, 2002). Such generalization performance suggests that listeners are capable of extracting information to underpin perceptual representations of speakers which are robust against variations in the acoustical input. Voice identity representations might also be encoded in relation to a voice "prototype" which represents the average (or "norm") of all voices encountered by the listener (Belin et al., 2011; Latinus & Belin, 2011). Experiments which report perceptual after-effects (that is, changes in the perceived quality of a stimulus after repeated exposure to an adaptor stimulus) provide support for this notion. For example, when listeners are adapted to an "anti-voice" (i.e., a "caricatured" version of an average of many voices relative to a given single speaker's voice, lying on an identity dimension opposite to that of the original voice) through repeated presentations, they subsequently identify a morphed average of many different voices as the identity opposite to the adapting anti-voice (Latinus & Belin, 2011). In other words, if a listener is

repeatedly presented with an anti-version of the voice of speaker A, then that listener will identity a subsequently presented ambiguous average probe stimulus as speaker A, even though the ambiguous average probe corresponds to no specific identity. These perceptual adaptation effects support the idea of a prototype- or "norm"-based framework which has proven plausible for both faces (Leopold, O'Toole, Vetter, & Blanz, 2001; Leopold, Bondar, & Giese, 2006; Rhodes & Jeffery, 2006) and voices (Andics, McQueen, & Petersson, 2013; Bruckert et al., 2010; Latinus & Belin, 2011; Latinus et al., 2013; Zäske, Schweinberger, & Kawahara, 2010).

1.3.2. Where does voice identity processing "happen" in the brain?

Generally, the available evidence from functional imaging and neuropsychological investigations indicates a strong involvement of the right temporal lobe in the processing of voice identity (Andics et al., 2010; Belin & Zatorre, 2003; Bethmann et al., 2012; Bonte, Hausfeld, Scharke, Valente, & Formisano, 2014; Formisano, De Martino, Bonte, & Goebel, 2008; Gainotti, 2013a, 2013b; Hailstone et al., 2011; Imaizumi et al., 1997; Latinus, Crabbe, & Belin, 2011; Nakamura, Kawashima, Sugiura, & Kato, 2001; Van Lancker & Canter, 1982; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Von Kriegstein & Giraud, 2004; von Kriegstein & Giraud, 2006). Early attempts to determine the cerebral locus of voice recognition predate the initial reports of the TVA: Imaizumi and colleagues (Imaizumi et al., 1997) used Positron Emission Tomography (PET) to measure changes in cerebral blood flow in listeners who performed speaker and emotion identification tasks. Their analysis revealed a network of areas, including the bilateral temporal poles, which produced a stronger signal while listeners performed the identification task. Later, members of the same group (Nakamura et al., 2001), again using PET, showed increased

responses during discrimination of familiar voices in the right temporal pole, among other areas. A subsequent influential fMRI study by Belin and Zatorre (2003) showed attenuated blood-oxygen-level-dependent (BOLD) responses in right anterior STS to repetitions of the same speaker (i.e., neural *adaptation* effects; Grill-Spector, Henson, & Martin, 2006), even when the syllable uttered by the speaker differed across presentations. Taken together, these latter two results implied that the right anterior temporal lobe plays a role in content-invariant representations of voices, which might aid listeners in matching incoming information to identity traces. Indeed, this is a notion supported by evidence suggesting a functional segregation of right STC: where the posterior portions might be involved in a more "bottom-up" processing of the complex spectro-temporal properties of voices, anterior regions appear to be under "top-down" control, engaging with the specific process of matching stored identity representations to presented voices in explicit recognition tasks (von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004; Schall et al., 2015). Further compelling evidence for a right temporal involvement in the "invariant" representation of voice identity comes from a study which used multi-voxel pattern analysis (MVPA) to decode speaker identity from distributed pattern activity elicited by voices in auditory cortices (Formisano et al., 2008). Distributed patterns spread throughout the bilateral auditory cortices – but with a strong right hemisphere preponderance – enabled a machine-learning algorithm to recover the identity of the speaker whose voice elicited the response pattern; in other words, to "decode" identity. These distributed response patterns not only supported identity decoding, but also enabled the classification algorithm to generalize from learned utterances to responses elicited by untrained utterances from learned speakers.

While many of the right temporal lobe areas mentioned in the above studies overlap with the 'classical' TVA (Belin et al., 2000; Pernet et al., 2015), results from studies which independently localize generally voice-sensitive areas reveal that these also evince sensitivity to voice identity. TVA voxels are, for example, sensitive to changes in voice acoustics along a continuum morphed between different speakers (Andics et al., 2013; Latinus, Crabbe, & Belin, 2011). Furthermore, the TVA appear to code voices in relation to an acoustical "prototype", in a mean-based fashion; voices which are closer to the mean voice, elicit reduced BOLD response in the TVA, and are perceived as less distinctive (Andics et al., 2013; Latinus, McAleer, Bestelmeyer, & Belin, 2013). Furthermore, distributed pattern content elicited by different speakers' voices within the TVA support voice-identity decoding, using MVPA methods (Bonte et al., 2014). Finally, the TVA appear to be functionally and structurally coupled with the face-sensitive "Fusiform Face Area" (FFA), indicating that they may play a role in the modality-general representation of person identity (Blank, Anwander, & von Kriegstein, 2011; von Kriegstein & Giraud, 2006).

It should be acknowledged that, despite their consistent appearance in the relevant literature, identity effects are not limited to the temporal lobes. Indeed, areas in the right inferior frontal cortices (IFC) show sensitivity to acoustical changes between voices, and in learned perceptual difference in voice samples which have been morphed between different speakers along a continuum (Andics et al., 2010, 2013; Latinus et al., 2011). Beyond these areas, the left middle temporal gyrus (MTG), left STC (Stevens, 2004; von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004), and left cingulate gyrus (Arnott, Heywood, Kentridge, & Goodale, 2008; Latinus et al., 2011) are responsive to identity information, particularly in explicit tasks. Furthermore, the decoding results of Formisano and colleagues described above (Formisano et al., 2008) suggest that voice identity

representations may be carried by auditory cortical voxels in both hemispheres - their maps of the maximally discriminative voxels included very sparse areas in the left Heschl's Gyrus/Sulcus, and left STS/STG. Likewise, Andics et al., (2010) find an overall leftward lateralization for trained perceptual identity effects in the temporal lobe (as opposed to a right lateralization for sensitivity to stimulus acoustics) and Bonte and colleagues (2014) observe that voice identity decoding performance from left-hemisphere sound-sensitive STC voxels predicts performance on a speaker matching-to-sample task, where right hemisphere STC voxels do not. Therefore, despite an apparent strong reliance on right superior temporal lobe, identity processing is not restricted to this area.

1.3.3. Interactions of "speech" and "voice" systems: Voice recognition is not all about "how" you sound, but also "what" you say.

While natural vocal acoustics are important for telling speakers apart, discrimination and recognition can be achieved when indicators of voice quality are removed or disrupted. For example, inter-talker similarity ratings are comparable whether estimated on pairs of talkers uttering clear speech or on sine-wave analogues of their voices, which eliminate natural voice quality (Remez, Fellowes, & Nagel, 2007). Others have shown that explicit talker recognition (as opposed to subjective similarity rating) can also be achieved from sine-wave voices (Sheffert et al., 2002), and from voices filtered through an electrolarynx (Perrachione, Stepp, Hillman, & Wong, 2014), which fixes fundamental frequency and other acoustical properties across talkers. Such results show that the characteristic manner in which an individual talker articulates speech sounds (i.e., their "idiolect"), can enable a listener to tell them apart from other voices, when natural vocal source and filter information is absent (Remez, Fellowes, & Rubin, 1997). Conversely, when non-

linguistic (or “indexical”) information is preserved, but speech intelligibility is disrupted such as by time-reversing a speech signal (Sheffert et al., 2002; Van Lancker, Kreiman, & Emmorey, 1985), voice recognition can still be achieved. However, note that in all of these cases, talker identification performance is maximised when the linguistic information in an utterance can be apprehended by the listener, *and* when natural voice acoustics (the “para-linguistic” or “indexical” information characteristic of a given speaker) remain present in the signal. Interestingly, the reverse appears also to be true, where performance in speech perception tasks (which rely on linguistic information) is influenced by the quality of non-linguistic voice information. For example, the familiarity of a voice can affect the perceived intelligibility of an utterance (Nygaard & Pisoni, 1998). Likewise, increasing the number of different voices which are heard by a listener during training can affect their ability to perceive speech in noise (Mullennix, Pisoni, & Martin, 1989) and the phonemic properties of speech are not perceived independently of non-linguistic voice information (Mullennix & Pisoni, 1990).

By way of further illustration, a very recent study showed that increasing the amount of relevant linguistic information in a voice can positively impact upon voice recognition. Zarate and colleagues (Zarate, Tian, Woods, & Poeppel, 2015) tested monolingual English-speaking listeners on their ability to recognize different talkers across a range of different utterances, including non-speech vocal sounds, foreign speech (German and Mandarin), pseudo-English (i.e., “nonsense” speech which uses the natural phonology of English), and intact English. Listeners performed best on natural English, followed by “native” pseudo-speech, German, Mandarin, and, lastly, non-speech vocalizations. Notably this study demonstrates a well-known psychological effect which serves as the focal point of this thesis: the so-called “Language-Familiarity Effect” (LFE) for talker identification.

1.4. The Language Familiarity Effect (LFE) in voice recognition

The previous section described instances where performance in talker identification tasks may be improved as listeners are granted more access to relevant linguistic information (Zarate et al., 2015). Notably, while the behavioural and neuroimaging research described above involves a range of different speech sounds (from vowels, through longer syllables, to connected sentences), these speech sounds are typically drawn from listeners' native languages. Listeners are known to show improved performance in talker recognition tasks when they are acquainted with the spoken language used by a talker. This "Language-Familiarity Effect" (or LFE) has been robustly demonstrated over a range of different talker identification tasks and language conditions, first being identified nearly 30 years ago.

To my knowledge, the first attempt to study the influence of language knowledge on voice recognition was conducted by McGehee in 1937 (McGehee, 1937), involving a sample of native English speakers from the United States. An unknown talker read a 56-word paragraph whilst hidden from view from the participants. The listeners were invited to return for a secondary phase after a variable retention interval to participate in a line-up identification phase, where 5 speakers were heard and the listener had to indicate to which speaker's voice they had been previously exposed. The listeners evinced comparable performance on the line-up task for an English speaker (83% accuracy) as they did for a German speaker (81% accuracy), after a 48-hour delay. However, the line-up in the foreign condition consisted not of five speakers who were homogeneous with respect to spoken language, but a mixture of speakers, including only one other German speaker. It is therefore unsurprising that the previously encountered speaker was reliably

identified, given the high-level of phonological dissimilarity among speakers presented as part of the testing battery. Later, Goldstein and colleagues (Goldstein, Knight, Bailis, & Conover, 1981) found comparable results: in a group of American English-speaking listeners, no difference was found between recognition scores for English or Spanish speaking voices. However, while they did find that reducing the phonemic content of the stimulus negatively impacted upon overall recognition abilities, this performance reduction was most pronounced when listeners attempted to recognize English-speaking voices with a strong Taiwanese accent.

Following these early studies, a series of identification experiments provided the first evidence for an influence of spoken language upon voice recognition. Thompson (1987) showed that English-speaking listeners evinced a recognition advantage for English-speaking voices, as contrasted with Spanish voices, and voices speaking English with a Spanish accent. This appears to be the first demonstration of the LFE in the voice recognition literature, and it has spawned many replications since. Historically, studies of the LFE have taken the form of forensic voice “line-ups”, where participants are initially presented with one voice during a familiarization phase. Later, after some retention interval, they are presented with multiple voices and asked to indicate whether the voice they heard during familiarization was present. Since the Thompson (1987) studies many such line-up studies have been conducted with convergent results (Goggin, Thompson, Strube, & Simental, 1991; Koster & Schillert, 1997; Philippon, Cherryman, Bull, & Vrij, 2007; N. Schiller & Koster, 1996; Schiller, Koster, & Duckworth, 1997; Sullivan & Kügler, 2001; Sullivan & Schlichting, 2000). More recently, investigations of the LFE have been conducted with tighter experimental controls, and with more consideration given to stimuli, presentation conditions, participant training and data analysis (Bregman & Creel,

2014; Levi & Schwarz, 2013; Neuhoff, Schott, Kropf, & Neuhoff, 2014; Perrachione, Del Tufo, & Gabrieli, 2011; Perrachione, Pierrehumbert, & Wong, 2009; Perrachione & Wong, 2007; Wester, 2012; Zarate et al., 2015). Nonetheless, seemingly regardless of the nature of the task, or the languages used to produce testing items, some form of the LFE has emerged in each of the cited works. I will now discuss some of the "highlights" which have emerged since the first reports of the LFE.

1.4.1. Is the LFE attenuated by the degree of phonological 'overlap' between native and foreign speech?

Since the initial reports of the LFE, its nature has captured the attention of psycholinguists and cognitive psychologists alike. One issue concerns the degree to which two languages must differ before the LFE becomes manifest. In other words, does the effect become attenuated as a foreign language becomes phonologically more similar to a listener's native language? Studies which have experimentally manipulated accent have found that voices speaking in participants' native language without heavy foreign accents are better remembered than voices with heavy accents (Goggin et al., 1991; Goldstein et al., 1981; Thompson, 1987). However, two of these studies also found that voices speaking English with a heavy Spanish accent were better remembered by English speaking listeners than Spanish voices (Goggin et al., 1991; Thompson, 1987), although not as well as "unaccented" American speakers of English. As Spanish-accented performance was intermediate, it seems that the benefit of native speech was somewhat reduced by the Spanish accents, which perhaps made the phonemic content of the experimental stimuli sound somewhat 'less English' to the participants.

More formal studies of this “linguistic similarity” hypothesis have yielded mixed results. Koester and Schiller (1997), had employed English and German language conditions in their previous works (Schiller and Koester, 1996; Schiller et al., 1997) and noted that they both belong to the West Germanic language family. Consequently, they examined whether varying the degree of phonological overlap between native and foreign language conditions would have any bearing upon the observed LFE. On the one hand, they find that Mandarin and Spanish speakers with knowledge of German perform more poorly in German voice recognition than English speakers with German experience but better than their compatriots with no German knowledge; on the other, they find that Mandarin speakers with no German knowledge outperform both English and Spanish speakers with no German knowledge. As German is more closely related to English than either Spanish or Mandarin, this appears to contradict the notion that linguistic overlap might ameliorate the LFE. Furthermore, in a voice-pair discrimination study, Wester (2012) found no benefit of phonological similarity; listeners showed comparable within-language speaker discrimination performance in Finnish, German and Mandarin.

In contrast, elegant recent results from Zarate and colleagues (Zarate et al., 2015) provide support for the “linguistic overlap” hypothesis. English-speaking participants were tasked with learning and recognizing voices in 5 conditions: producing non-speech stimuli (laughs, cries, etc.) or uttering single words in English, Pseudo-English, German and Mandarin. Critically, stimuli were uttered by the same talkers in all language conditions, in order to limit non-linguistic differences between speakers which could have been introduced by the use of multiple talker groups. Voice recognition improved as more information became available to the listeners, who performed worst on non-speech items, followed by Mandarin, German, pseudo-English and English. Interestingly, the

differences between the “phonologically-familiar” German, pseudo-English and English conditions were not statistically significant, indicating that the listeners performed relatively similarly on this subset of items. This stands in contrast to the Mandarin condition, which featured phonology even less familiar to the listeners than the German and pseudo-speech conditions. Performance, therefore, appears to benefit when listeners can take advantage of stored native phonological representations, even when they appear in native pseudo-speech. Performance also appears to be graded, according to the phonological proximity of a foreign language to one’s native language.

1.4.2. Can the LFE be ameliorated with foreign language experience?

We have seen that the relatedness of a foreign language to one’s native language might bear on voice recognition abilities. A related question concerns whether a listener could improve their foreign language voice recognition performance by undertaking language training. If so, how much language training would be required – enough to master a few words, or enough to obtain bilingual, or native-like abilities?

Goggin and colleagues (Goggin et al., 1991) found that Spanish-English bilinguals show better identification of Spanish-speaking talkers than they do for English-speaking talkers, but this trend is not statistically significant, and not comparable in magnitude to the native-language identification bias shown by English monolinguals favouring the identification of English-speaking talkers. This suggests that the observed language familiarity effect is ameliorated as a function of a listener’s second-language experience. However, Schiller and Koester (1996) found that English learners of German recognized German voices at a comparable level to native German speakers. Here, the amount of

experience (“some” knowledge compared with native-like experience) appeared to have no impact on recognition ability.

Note that Schiller and Koester (1996) do not quantify the amount of German language experience that their English-speaking group had (merely remarking that they had “some” knowledge). A subsequent study (Sullivan and Schlichting, 2000) followed this up by recruiting four groups of listeners, all native English speakers with varying degrees of knowledge of Swedish: English-speaking high school students with no experience, and three groups of University students at various stages in undergraduate degree courses majoring in Swedish language (second semester, fourth semester and eighth semester). Two key findings were reported: Firstly, all Swedish-speaking participants performed better in a Swedish voice identification task than the group with no Swedish experience. Secondly, no performance differences were found between the Swedish-speaking groups. Taken together, these findings appear to vindicate the postulate of Schiller and Koester (1996): *some* language knowledge is important, but increasing one’s knowledge of a language will not necessarily result in foreign-language identification improvements, beyond the initial salutary effects of simply commencing study of that language. In a follow-up study by the same authors (Sullivan & Schlichting, 2001) results from the first sample were contrasted with new data from adult English-speaking students at the same University who had no knowledge of Swedish, or any other Scandinavian language. Again, the non-Swedish speaking groups displayed comparable performance on the line-up task, and both performed more poorly than those participants with experience of Swedish.

More recently, in contrast to the foregoing, a study using a different experimental approach found that the amount of foreign-language experience and individual has can

indeed affect their memory for voices. While previously cited works have implicitly examined differences between bilingual and monolingual listeners in the learning and recognition of native and foreign talkers (Goggin et al., 1991; Schiller and Koester, 1996; Koester and Schiller, 1997), Bregman and Creel (Bregman & Creel, 2014) explicitly measure the rate of talker learning in these groups. The influence of the degree of bilingualism in a listener was probed via their reported age of acquisition for a foreign language. In a sample of Korean-English bilinguals and English monolinguals, the classic interaction of stimulus and testing group - the hallmark of the LFE – was found. English participants reached an 85% training performance criterion for English-speaking voices more swiftly than they did for Korean voices and vice-versa for the Korean bilinguals. A relationship between age of English acquisition and time required to reach training criterion was also found, providing an indication that the amount of experience which one has with a language can enhance their ability to identify individuals speaking in that language; late bilinguals required more training repetitions to reach criterion on English items than early bilinguals.

A related study by Perrachione and Wong (2007), examined whether the LFE was truly a linguistic effect, or whether it could be ameliorated through general auditory exposure. In an elegant experimental paradigm, English and Mandarin native speakers were tested over 6 experimental sessions (conducted on separate days) on their ability to identify English and Mandarin voices. Both groups evinced the classic LFE at the outset of the testing program, and as training continued, their overall recognition performance continued to improve. Despite their general recognition improvement, monolingual English participants showed a reliable LFE at every stage of training; English recognition performance was always superior to Mandarin recognition performance. However, in the

Mandarin group, the performance discrepancy between English and Mandarin conditions had considerably narrowed by the end of training. Furthermore, the magnitude of the Mandarin group's LFE at the commencement of the training program was smaller than that of the English group. Both of these findings were likely due to the fact that the Mandarin testing group had considerable functional experience with the English language (as much as several years for some participants), meaning that they were by no means as "deaf" to the foreign speech condition as their English monolingual counterparts. Therefore, this study, in addition to suggesting that the LFE is indeed an effect related to linguistic processing rather than general auditory exposure, provides further evidence that the amount of experience which one has with a foreign language can impinge upon their LFE.

1.4.3. At what developmental stage does the LFE appear?

As we have seen, the age at which an individual acquires a second language may have a bearing on their individual LFE. We might next ask: how early on in life can the LFE be observed? Voice recognition abilities appear to develop very early in life (Kisilevsky et al., 2003; Mehler, Bertoncini, Barriere, & Jassik-gerschenfeld, 1978; Ockleford, Vince, Layton, & Reader, 1988) and sensitivity to the maternal language, likewise, is apparent at the age of 2 days, and even in-utero (Kisilevsky et al., 2009; Moon, Cooper, & Fifer, 1993). Furthermore, young infants are sensitive to speaker identity changes when a language changes across two utterances (Nazzi, Bertoncini, & Mehler, 1998; Nazzi, Jusczyk, & Johnson, 2000) and new-borns show an ability to discriminate unfamiliar native-language speakers uttering single words (Floccia, Nazzi, & Bertoncini, 2000). These findings indicate that young children show sensitivity to both language-structure elements and to indexical

variability in voices and it is pertinent, therefore, to ask to what extent these abilities interact.

To this end, Johnson and colleagues (Johnson, Westrek, Nazzi, & Cutler, 2011) tested voice discrimination abilities in young (7-month old) Dutch infants. These infants were habituated to the voices of different speakers during a training phase, and were then presented a voice during a test phase which could differ from the training voice in language or identity. Consistent with previous findings, the infants were highly sensitive to changes in language. However, they were also sensitive to changes in voice identity only when training and test voices spoke in their nascent maternal language. This result extends a previous finding which showed that 4-month-old infants who have been raised in an English-speaking environment notice a gender change only in their native language (Sundara and Kuhl, 2006). Taken together, these results suggest that early exposure to the sound of a language can confer a benefit in talker discrimination. Note that these early stages, speech comprehension abilities in infants are immature, even though the process of perceptual narrowing to the sounds of their 'native' language may have begun (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1991; Kuhl, 2004; Kuhl & Rivera-Gaxiola, 2008). Therefore, it seems that the LFE might emerge even before the listener is capable of fully understanding the linguistic message. It is of interest therefore to determine the extent to which this finding may apply to adult listeners, under circumstances where the sounds of their mother-tongue may still be present in an unintelligible speech signal.

Further along the developmental timeline, Levi and Schwarz (2013) find that typically-developing younger children (7-9 years old) show an LFE, but also poorer general voice recognition abilities than typically-developing older children (10-12 years old). These

older children who do not show a LFE are outperformed in native-language recognition by adults, but also perform better than those adults in foreign talker recognition. The development of voice recognition might therefore be a non-linear process (see also, Mann, Diamond, & Carey, 1979) where an initial native-language benefit is overridden by an upswing in general auditory abilities, followed by the attainment of language-specific adult talker processing. Indeed, this is consistent with the findings of Sundara and Kuhl (2006); there, 4-month-old children noticed a change in speaker gender only in their native language, where 12-month-olds noticed the change in both native and foreign conditions. These results are also interesting in the context of results reported in Perrachione et al (2011), which suggested that dyslexic adults show no benefit of native language in voice recognition, instead performing equally poorly on native and foreign speech relative to non-dyslexic controls. Specifically, while language-impaired adults show poorer native-language voice recognition performance than typical adults, the same pattern is not observed in a contrast of typical and impaired children in the results of Levi and Schwarz (2013).

1.4.4. The LFE: Interim summary

So far, we have covered nearly 30 years' worth of literature on the LFE, with cases ranging from relatively uncontrolled line-up studies, to well-controlled laboratory experiments involving adults, teenagers, children, and young infants. It is apparent that this phenomenon is robust and replicable, given its emergence in a range of different contexts, involving many different languages. What, then, might underpin this effect? We may look to another well-known phenomenon in the person perception literature for inspiration: the "Other-Race Effect" (ORE) for faces, where face recognition performance

is better for faces of one's own race groups, as compared to recognition performance for faces of another race group.

1.4.5. The LFE as an auditory analogue to the ORE: the Valentine model

The ORE has a long and rich history, similar to the LFE in the sense that it has occupied thought in both forensic and cognitive psychology (for a meta-analytic review, see Meissner & Brigham, 2001). I will direct the reader to a particularly influential account of the ORE, for which empirical evidence has recently accumulated, and which might provide a theoretical basis for discussion of the mechanistic underpinnings of the LFE. The similarity-based "face-space" account, proposed by Valentine (Valentine & Endo, 1992; Valentine, 1991) holds that faces are represented as points in a hypothetical multidimensional space. When Valentine initially proposed this model, he considered it both in terms of a norm-based account and an exemplar-based account. In the norm-based version, faces are encoded by reference to the mean of the face space. The mean of the space represents an average of all faces encountered by the observer over the course of their perceptual experience. The exemplar-based version posits that individual faces are encoded with respect to their relative (dis)similarity to each other, so the centre of the space is irrelevant.

The dimensions of the face-space could represent features which are useful for identity diagnostics. For example, one dimension might encode the position of the eyes, or the shape of the nose, or the colour of hair. Importantly, regardless of what these dimensions might represent, they are proposed to be elaborated through perceptual experience; a consequence of this, is that the dimensions may be developed according to the faces to

which an individual is most exposed over the course of his or her life. Therefore, if a Western-Caucasian perceiver receives more exposure to other Western-Caucasian faces, then the dimensions of his hypothetical face encoding space will be attuned to those features which best aid in the differentiation of Western-Caucasian faces. For example, hair and eye colours span a relatively wide-range in Caucasians, as compared with, say, African or East Asian ethnicities. As such, an identity representation scheme which contains dimensions coding for these features will serve a perceiver well in recognition tasks. However, when a perceiver attempts to recognize faces from a group to which they have little exposure, then they may experience confusions between different individuals which could lead to recognition errors. As in our earlier example, consider the case of a Caucasian perceiver and a set of East Asian faces. If the observer has had little exposure to East Asian faces, then his encoding scheme will be asymmetrically optimised for the storage and retrieval of Caucasian faces. Consequently, when relying on diagnostic features which may help to individuate Caucasians – such as hair and eye colour – recognition of Asian faces will be impaired, as different Asian identities may not be as well differentiated as a function of variability in those diagnostic features.

By extension, we can now consider the LFE within this framework. This analogy has been drawn by others previously (e.g., Perrachione et al., 2007), focusing on representations of language phonology. Taking the many apparent similarities between face and voice processing into account (Yovel and Belin, 2013), we may posit a multidimensional voice space which resembles that which has been adduced for faces. Here, dimensions may include speaking fundamental frequency (f0) or the formant frequencies, or harmonicity ("smoothness") of a voice. Indeed, as was discussed earlier, this issue has already been examined in some depth (Baumann and Belin, 2010; Latinus et al., 2013). Importantly, the

manner in which a voice is encoded could be influenced by early linguistic experience. For example, as we have seen, infants display an early sensitivity to speaker changes in their native language only, which might follow on from perceptual narrowing to the sounds of their burgeoning mother-tongue (Kuhl, 2004; Kuhl and Rivera-Gaxiola, 2008). In the voice space, dimensions might be developed based upon the sounds an individual is most accustomed to hearing from other speakers in their environment; in other words, the native-language speech sounds most commonly heard. Under this conception, incoming voices will be assessed by reference to the listener's stored representations of the sounds of their mother tongue. The listener can bring their phonological knowledge to bear in computing the speaker's pitch, accent, or other features which might vary according to the speaker's anatomy or geographical background. After this evaluation, the voice will be stored in the space accordingly, with reference to an individual's internal voice "prototype". However, if they are tasked with evaluating voices which do not speak their native language, their voice-space dimensions (which have been shaped according to the phonology of that language) will be of lesser use for the encoding of those voices, which may result in a perception of increased inter-talker similarity, leading to a failure to tell-apart and recognize foreign voices.

While the framework proposed above describes the LFE with reference to stored representations of voices in a multidimensional space which has been shaped according to linguistic input, an alternative conception concerns the "abstract" or "prototypical" representation of familiar speech sounds and language-specific acoustics *themselves*, as has been alluded to above and by others (e.g., Perrachione et al., 2007). In spite of the interactions between "speech" and "voice" information described earlier, human listeners are still capable of discarding variability in paralinguistic input when performing

native-language speech perception tasks; a process referred to as talker "normalization" (e.g., Johnson, 2008; Nygaard & Pisoni, 1998). For example, the word "dog" is easily recognizable to a native English speaker, whether it is produced by a female or male speaker, despite what may be considerable differences in the acoustical profiles of the two signals (Johnson, 2008). The listener may take advantage of stored "prototypical" representations of familiar speech sounds (such as phoneme combinations and words) and language-specific acoustics (e.g., in Mandarin, a prototypical representation of the characteristic f0 pattern of the rising tone) during speech processing in order to allow them to disregard pronunciation differences between speakers, which may result from such properties as accent or gender (Belin et al., 2004). Conversely, in the case of the LFE, just as this variability can be "tuned out" in speech perception tasks, it may in fact be brought to bear in a manner which benefits the listener when performing a voice recognition task in their native language. Rather than discarding variability, the listener could take advantage of inter-speaker differences in the pronunciation of familiar words and use these as cues for identity differentiation. In other words, the ability to process non-linguistic variability around robust, stored prototypical representations of native language speech sounds may facilitate voice recognition in a listener's native language. On the other hand, as in the case of an unknown language, impoverished (or non-existent) representations of the speech sounds which are used by a talker will leave the listener reliant purely upon para-linguistic cues to voice identity, which, while useful, do not enable successful recognition to the same degree as cases where both linguistic and para-linguistic information is available (Perrachione et al., 2011; Zarate et al 2015). As efficient voice individuation is impeded, different foreign voices might therefore sound highly similar to a listener, leading to recognition failure.

With regards to the ORE, the "multidimensional space" model has garnered empirical support from computational modelling studies (Caldara and Abdi, 2006) and in recent electro-encephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) studies showing enhanced individuation of own-race faces in putatively face-preferential neurophysiological responses (Vizioli, Rousselet, & Caldara, 2010; Vizioli, 2012). Given that a norm-based version of this account may be plausibly extended to voices (Bruckert et al., 2010; Latinus and Belin, 2011; Andics et al., 2013; Latinus et al., 2013; Yovel and Belin, 2013), it might also represent a useful basis for interpreting the LFE. Behavioural studies provide some support for voice recognition failure based on degraded phonological representations (Perrachione et al., 2011), but do not explicitly evaluate listeners' perceptions of similarity among different native and foreign talkers. Furthermore, while the ORE has enjoyed sustained interest from the neuroimaging community (for a review, see Natu, Raboy, & O'Toole, 2011), no explicit examination of the neural basis of the LFE has been conducted, to my knowledge. The neural bases of speech perception (Hickok & Poeppel, 2007; McGettigan & Scott, 2012; Price, 2010, 2012; Scott & McGettigan, 2013) and voice recognition (Belin et al., 2004; Belin et al., 2011; Yovel and Belin, 2013; Schweinberger et al., 2014) have been, respectively, relatively well-studied, but most investigations evaluate these phenomena separately. However, neuro-cognitive models (Belin et al., 2004; 2011) and behavioural findings – such as the LFE – strongly suggest that the two systems are coupled in realizing successful voice recognition. Indeed, a handful of recent neuroimaging works have begun to investigate this interaction. For example, it was recently shown that speech perception areas in the posterior left superior temporal cortex (STC) responded to speaker-related changes in vocal tract parameters (von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010). Furthermore, right posterior STC processed speaker-related information during a speech

recognition task as contrasted with a voice recognition task, and these left and right posterior superior-temporal regions were functionally connected. More recently, Chandrasekaran and colleagues (Chandrasekaran, Chan, & Wong, 2011) showed that the left posterior middle temporal gyrus (MTG) exhibited reduced BOLD-signal adaptation to variations in both indexical (speaker-related) and lexical (word) information, as contrasted with a condition where all stimulus content was repeated (i.e., same speaker identity and same word repeated). The authors interpreted this as evidence of neural integration of "what" and "who" information in this region. Taking these findings into account, therefore, it is important to determine the extent to which such putative integrative regions are perturbed under foreign speech conditions, as contrasted with the familiar speech conditions previously used.

1.5 Aims of this thesis

Heretofore, I have provided a summary of the relevant literature on the LFE. The emergent picture is of a robust effect which is manifest across a broad range of testing conditions and listener groups. The effect may share a common mechanism with the ORE for faces for the representation of person identity; it seems that as different own-race faces appear more dissimilar than different other-race faces to observers, different native voices may also *sound* more dissimilar than foreign voices to listeners. However, as acknowledged above, while this representational mechanism has been explored for the ORE, both in modelling and neuroimaging work, it remains untested explicitly with regard to the behavioural and neural basis of the LFE. Furthermore, it is unclear whether speech comprehension is necessary for the perception of increased foreign language similarity to take hold, as opposed to, say, a familiarity with certain phonological elements

characteristic of a language which might remain present in an incomprehensible speech signal. Results from infant participants indicate that a full understanding of the linguistic message is not a pre-requisite for the LFE, as they can sense a talker change only in their native tongue, in the absence of fully developed speech comprehension abilities (Sundara and Kuhl, 2006; Johnson et al., 2011). However, it is unclear whether the same native-language discrimination advantage will be present in adults, when the intelligibility of a native speech signal is compromised.

Therefore, **Chapters 2 and 3** focus on the role that speech comprehension plays in the LFE. As described above, young infants with immature language comprehension abilities evince an LFE for discrimination (Sundara and Kuhl, 2006; Johnson et al., 2011). **Chapter 2** examines whether the same LFE for discrimination is present in groups of adult English and Mandarin speakers. Crucially, speech intelligibility was disrupted by time-reversing the stimulus material, which consisted of English and Mandarin speech spoken by native speakers. Listeners recorded similarity ratings of pairs of spoken voices, providing an explicit test of whether different foreign voices sound more similar to listeners than native voices. If speech comprehension is a pre-requisite for the LFE, then we should see no interaction between listener and speaker language; however, if listeners can take advantage of language specific information present in the reversed speech signal, then, under the proposed representational scheme, they should show a discrimination advantage for pairs of native voices, judging them as more dissimilar-sounding than foreign pairs.

In **Chapter 3,** I present an fMRI study which examined the similarities among neural responses elicited by time-reversed native and foreign speech, in the groups of English-

and Mandarin-speaking participants from **Chapter 2**. Within a Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008) framework, we compare models which capture the patterns of dissimilarities among cerebral responses to different pairs of voices (hypothesized under the Valentine encoding scheme), to the actual, observed patterns. As in **Chapter 2**, if comprehension is a key component of the LFE, then there should be no differences observed between the dissimilarity arrangement of responses to native and foreign speakers.

In **Chapter 4**, I present an fMRI experiment involving monolingual English speaking participants and intelligible speech stimuli. Participants were scanned whilst listening to clips of English and Mandarin speech and performed a 1-back repetition task. The intention of this experiment was to examine whether brain representations of different talkers were more "differentiated" during native speech listening, as contrasted with foreign speech. Previous multivariate neuroimaging studies of the basis of voice identity have shown that multivariate patterns of BOLD-signal activation enable 'decoding' of speaker identity (e.g., Formisano et al., 2008; Bonte et al., 2014). These studies utilized speech material drawn from participants' native language as stimuli; therefore, in this chapter, I examine whether neural identity representations vary as a function of spoken language, by testing whether multivariate identity decoding accuracy is reduced for foreign-language voices relative to native-language voices.

Finally, in **Chapter 5**, I present a general discussion of the outcomes of all of the foregoing work, and present strands of thought for future investigations of the LFE.

Chapter 2: A language-familiarity effect for speaker discrimination without comprehension*

Abstract

The influence of language familiarity upon speaker identification is well established, to such an extent that it has been argued that "Human voice recognition depends on language ability" [Perrachione TK, Del Tufo SN, Gabrieli JDE (2011) Science 333(6042):595]. However, 7-mo-old infants discriminate speakers of their mother tongue better than they do foreign speakers [Johnson EK, Westrek E, Nazzi T, Cutler A (2011) Dev Sci 14(5):1002–1011] despite their limited speech comprehension abilities, suggesting that speaker discrimination may rely on familiarity with the sound structure of one's native language rather than the ability to comprehend speech. To test this hypothesis, we asked Chinese and English adult participants to rate speaker dissimilarity in pairs of sentences in English or Mandarin that were first time-reversed to render them unintelligible. Even in these conditions a language-familiarity effect was observed: Both Chinese and English listeners rated pairs of native-language speakers as more dissimilar than foreign-language speakers, despite their inability to understand the material. Our data indicate that the language familiarity effect is not based on comprehension but rather on familiarity with the phonology of one's native language. This effect may stem from a mechanism analogous to the "other-race" effect in face recognition.

***Corresponding publication:** Fleming, D., Giordano, B. L., Caldara, R. & Belin, P. (2014) **A language-familiarity effect for speaker discrimination without comprehension**.

Proceedings of the National Academy of Sciences, 111**(38)**, 13795-13798.

2.1 Introduction

The human voice carries linguistic information as well as paralinguistic information about a speaker's identity, and normal listeners possess abilities to extract both types of information. The neuro-cognitive mechanisms underlying speech comprehension and speaker recognition are dissociable, as evidenced by cases of both patients with receptive aphasia (impaired speech comprehension but preserved speaker recognition) and patients with phonagnosia (impaired speaker recognition but preserved speech comprehension) (Assal, G., Aubert, C., & Buttet, 1981; Garrido et al., 2009; Van Lancker, Cummings, Kreiman, & Dobkin, 1988; Van Lancker, Kreiman, & Cummings, 1989; Van Lancker & Kempler, 1987), as well as by differences in the cortical networks engaged by the two abilities (Belin et al., 2000; Belin, Zatorre, & Ahad, 2002; McGettigan & Scott, 2012; Poeppel, Idsardi, & van Wassenhove, 2008; Poeppel, 2003; von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004). However, speech and voice identity processing also interact to a considerable degree. Speech recognition is influenced by speaker variability and familiarity: listeners better understand and remember speech spoken by familiar speakers (Martin, Mullennix, Pisoni, & Summers, 1989; Mullennix et al., 1989; Nygaard & Pisoni, 1998; Pisoni, 1993). Conversely, speaker identification is influenced by language familiarity: listeners are typically poorer at identifying speakers of a foreign language. This so-called "Language-Familiarity Effect" (LFE) has been demonstrated across a diverse range of languages (Thompson, 1987; Goggin et al., 1991; Koester and Schiller, 1997; Winters et al., 2008; Perrachione et al., 2009) and is behaviourally robust, persisting even after several days of training (Perrachione and Wong, 2007).

A crucial, unresolved point of debate is whether the LFE depends upon linguistic mechanisms involved in speech comprehension, or rather reflects the greater familiarity with the phonological structure of one's own language without necessarily requiring an understanding of the linguistic message. On the one hand, evidence from dyslexic participants, whose phonological processing abilities are impaired (Gabrieli, 2009), supports the importance of linguistic processing for general speaker identification abilities: English-speaking dyslexic participants do not show the LFE, (i.e., better memory for English-speaking than Chinese-speaking voices) shown by normal participants (Perrachione et al., 2011). On the other hand, a LFE is already apparent in infants before they can fully comprehend speech: 7-mo-olds notice a speaker change in their native language but not in an unfamiliar language (Johnston et al., 2011). Although results from dyslexic participants suggest a specific link between the LFE and "language ability" (Perrachione et al., 2011), results from infants (Johnson et al., 2011) suggest that experience with the phonology of the maternal language, rather than comprehension, may underpin the LFE. If this is the case, then the enhanced individuation of own-language speakers observed in 7-mo-olds should be observed in normal adult participants, even for unintelligible speech.

Here we tested this hypothesis by comparing dissimilarity ratings of own- and different-language speakers with time-reversed speech stimuli. Note that reversing speech disrupts intelligibility, but preserves "considerable phonetic information" (Binder et al., 2000) as well as sufficient indexical information to enable listeners to recognize voices (Bricker & Pruzansky, 1966; Sheffert et al., 2002; Van Lancker et al., 1985). We collected speaker dissimilarity ratings from Chinese and English listeners for all pairwise combinations of a set of Mandarin (n = 20) and English (n = 20) time-reversed speech clips, and compared

these dissimilarity ratings between groups of speakers and listeners. If the LFE is based primarily on language comprehension, then we should observe no inter-language difference in discrimination performance, as time-reversal rendered all stimuli unintelligible. Conversely, familiarity with a language's characteristic phonological structure may suffice to engender a LFE for speaker discrimination, even without comprehension. Mandarin Chinese is a tonal language, whereas English is stress-based; as such, a Mandarin speaker and an English speaker may differ in terms of the language structure elements that they use to differentiate speaking voices. For example, Mandarin and English differ in speaking fundamental frequency (Eady, 1982; Keating & Kuo, 2012; Mang, 2001) and phonemic inventories: Mandarin features around 1,300 syllables, whereas English uses around 15,000 (Shu and Anderson, 1999); the languages have very little consonant overlap; and English features a high frequency and variety of consonant clusters, whereas Mandarin has no consonant clusters (Yeong & Rickard Liow, 2012); Duanmu, 2000). Time-reversal preserves the formant structure (in a "mirrored" form; Binder et al., 2000) of many phonemes and their mean fundamental frequency, and, given that these features may differ across the two languages in natural speech, and that they are relatively well-preserved upon reversal, then native speakers of both languages may still be sensitive to these differences even where intelligibility is disrupted. If sensitivity to such differences drives a LFE then each group should show higher dissimilarity ratings for pairs of voices speaking their native language than for pairs speaking the other language

2.2. Methods

2.2.1 Participants

Twenty Mandarin-speaking Chinese (8 female, mean age = 23.7, SD = 2.58) and 20 native English-speaking UK participants (10 female; mean age = 24.25, SD = 3.01) were recruited. Chinese participants' average duration of UK residency was 9.35 mo (SD = 7.34) and all had attained a minimum score of 6.5 on the IELTS test of English as a foreign language, or a comparable score on an equivalent test. English-speaking participants reported no experience with Mandarin Chinese. All participants were right-handed and reported no history of hearing difficulties or pathology. Participants gave written, informed consent for their involvement and received a monetary reward. The experiment was approved by the Ethical Committee of the University of Glasgow's College of Science and Engineering.

2.2.2. Stimuli

Testing stimuli were drawn from a pool of 400 clips of 40 female speakers (20 native English-speaking and 20 native Mandarin-speaking) reading 10 sentences (Open Speech Repository, 2005). Recordings were digitized at a 16-bit/44.1-kHz sampling rate and cut into individual sentences.

Full, sentence-length stimuli were imported into Adobe Audition 2 (Adobe Systems), where they were time reversed and trimmed to 1,250 ms. Where applicable, the automated noise reduction tools in Audition were used to remove unwanted line hum, or clicks/pops from the clips. Finally, stimuli were imported into MATLAB 7.10 (R2010a)

where they were normalized for RMS amplitude. The use of time-reversed clips of English and Mandarin speech ensured that stimuli are of equal intelligibility to both participant groups; time-reversed speech precludes access to lexical and semantic information, and it is therefore unlikely that participants extracted any meaning from the stimuli (although we did not explicitly measure this).

2.2.3. Procedure

Testing took place within an anechoic cabin, where participants were seated at a desktop PC. The experiment was programmed in MATLAB 7.5 (R2007a). On every trial, participants heard a pair of voices and were instructed to rate the likelihood that both voice clips had been produced by the same speaker, using a visual analogue scale where 0/far-left corresponded to “Same” and 1/far-right corresponded to “Different.” Participants were advised to use the full extent of the scale to record their responses and were permitted to replay a trial as many times as they felt necessary before responding. This procedure was repeated for all possible paired combinations of voice identity, yielding a total of 820 pairs (40 × 39/2 + 40 same-identity pairs). The assignment of sentence to speaker was randomized across identities, ensuring that no two voices in a pair ever produced the same sentence clip and that each participant received a unique series of sentence-to-speaker pairings, in addition to a unique identity pairing order. The self-paced experiment took participants ~2 h to complete, including an optional break when they had reached trial 411 (i.e., halfway through the experiment). Participants had received previous exposure to the voice stimuli in this experiment through their participation in an earlier functional Magnetic Resonance Imaging (fMRI) experiment, the results of which are not discussed here (see **Chapter 3** of this thesis).

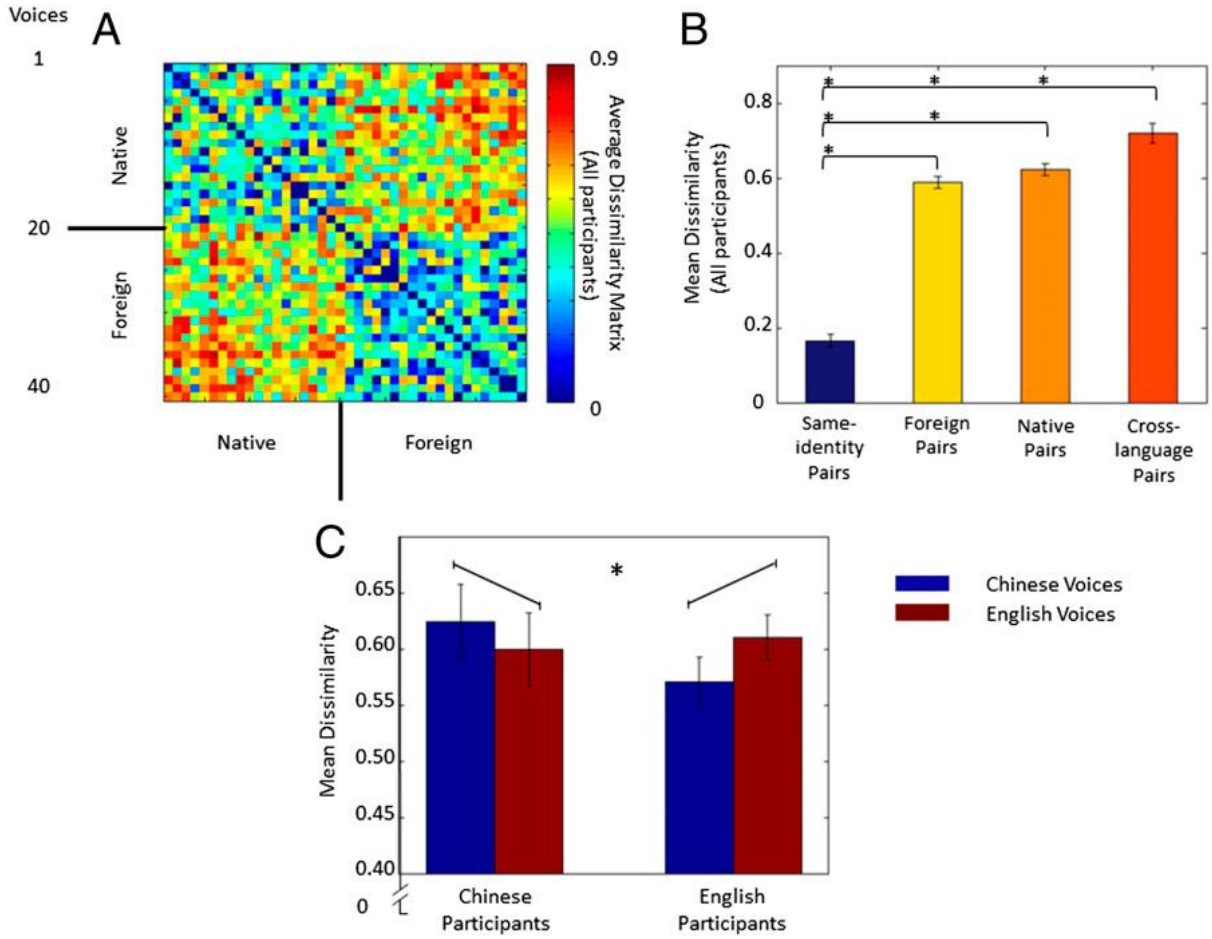


Fig. 2.1. Speaker dissimilarity ratings for pairs of Mandarin and English time-reversed sentences. (A) Matrix of dissimilarity ratings averaged across all participants in both listener groups (*N. Chinese Listeners* = 20; *N. English Listeners* = 20): individual participants' dissimilarity matrices are in a standardized arrangement, so that rows and columns 1–20 (top and left) represent native voices (Mandarin for Chinese listeners, English for English listeners), whereas rows and columns 21–40 represent foreign language voices, regardless of listener group. The colour scale indicates group-average dissimilarity ratings. (B) Average dissimilarity ratings for the four different types of pairs. Cross-language pairs were rated as most dissimilar. Within the same-language pairs, crucially, native-languagepairs were rated as more dissimilar than foreign-language pairs, even though all sentences were unintelligible. (C) Listener × speaker interaction: both participant groups record higher average dissimilarity ratings for native-language vs. foreign-language speaker pairs. Error bars indicate the SEM. All asterisks denote P < 0.05.

2.3 Results

All possible paired combinations of voices were presented to listeners who recorded their dissimilarity ratings via a computerized visual analogue scale, ranging from 0 to 1 (where a rating of 0 corresponded to maximum perceived similarity and 1 to maximum perceived dissimilarity). Fig. 2.1.A shows the dissimilarity matrix averaged across English and Chinese participants, where rows/columns 1–20 correspond to native voices and rows/columns 21–40 to foreign voices. Participants rated four types of pair: same-identity trials (where the same speaker was heard twice within a pairing), foreign–foreign trials, native–native trials, and native–foreign trials. No sentence clip was uttered twice within a pair. As shown in Fig. 2.1.A, inter-language pairs (where presentations consisted of one native and one foreign voice) were rated as more dissimilar than all other pairs, as reflected by the overall red color (high dissimilarity) of the upper right and lower left sub-matrices. Fig. 2.1.B illustrates the differences between each rating condition (same-identity mean = 0.16 ± 0.02 SE; foreign–foreign mean = 0.59 ± 0.02 SE; native–native mean = 0.62 ± 0.02 SE; native–foreign mean = 0.71 ± 0.02 SE). Each participant's mean ratings for each trial type were submitted to a repeated-measures ANOVA, which revealed a significant effect of pair type [$F(3, 39) = 314.2$, $P < 0.001$, $\eta 2$ partial = 0.89]. Post hoc tests also revealed significant differences for all pairwise comparisons of trial type (all P values < 0.02).

Crucially, when taking participant groupings into account, both Chinese- and English-speaking listeners produced higher average dissimilarity ratings for native-language voice pairs than for nonnative-language pairs (Chinese: native mean = 0.62 ± 0.03 SE, nonnative mean = 0.60 ± 0.03 SE; English: native mean = 0.61 ±0.02 SE, nonnative mean = 0.57 ±

0.02 SE) (Fig. 2.1.B and 2.1.C). We submitted these ratings to a 2 × 2 mixed-measures ANOVA, with listener language and speaker language as the between- and within-group (repeated) measures, respectively. A significant interaction between speaker and listener's language was observed, indicating that native-language dissimilarity ratings were higher, regardless of the language group of the listener [$F(1, 38) = 11.13$, $P = 0.002$, $\eta 2$ partial = 0.23]. The main effects of both listener and speaker language were not significant (P values > 0.2), suggesting that there were no statistical differences in rating behaviour between groups and that both sets of voices elicited similar rating behaviour. Paired t tests confirmed our prediction that both listener groups rated own-language pairs as more dissimilar than different language pairs (Chinese-speaking participants: Chinese > English [$t(19) = 2.57$, $P = 0.02$, Cohen's d = 0.17]; English-speaking participants: English > Chinese [$t(19) = 2.36$, $P = 0.03$, Cohen's d = 0.41]). To investigate the robustness of these results, we computed bootstrapped 95% confidence intervals of the native > foreign difference for each group, and for all participants taken together. We sampled participants' difference scores with replacement (10,000 iterations) and derived separate confidence intervals for each group (CI for Chinese participants: [0.007–0.04]; English: [0.005–0.07]; combined groups: [0.01–0.05]). As none of these confidence intervals contained zero, the observed effects may be considered reliable.

2.4. Discussion

We investigated whether the LFE in adults requires comprehension of the linguistic message. We found that listeners rated pairs of speakers of their own language as more dissimilar on average than pairs of speakers of a different language, even though all stimuli were rendered unintelligible by time-reversal. This result implies that the LFE is

not rooted in language comprehension *per se*, but rather is based on familiarity with the acoustical fingerprint of one's language, in a manner analogous to the "Other-Race Effect" (ORE) in face recognition.

Participants were presented with pairs of time-reversed sentences spoken by different speakers and were asked to judge how dissimilar the speakers were. Time-reversal was chosen because it is a simple procedure that compromises intelligibility while preserving some of the information present in the natural speech signal. For example, time-reversal disrupts the temporal attributes of speech segments, such as onsets and decays, and reverses pitch curves. Conversely, reversed speech is identical to natural speech in amplitude, duration, and mean fundamental frequency. Furthermore, the formant transition structure of many phonemes (e.g., fricatives and long vowels) is approximately mirrored in the reversed signal, and important indexical cues to speaker identity (for example, the harmonicity or "smoothness" of the voice, the speaker's speech rate, and the average pitch) are also retained. In sum, this remaining information can enable high inter-subject agreement in phoneme transcription tasks (Binder et al., 2000), and can be used by the listener to aid speaker recognition (Bricker and Pruzanksy, 1966; Van Lancker et al., 1985; Sheffert et al., 2002). Our participants were unable to extract any meaning from the stimuli, yet they showed reliable differences in their identity dissimilarity ratings. The most salient difference was between the different-language pairs (i.e., consisting of one sentence in English and one sentence in Mandarin) and the same-language pairs: The listeners reliably rated pairs of different-language speakers as more dissimilar than pairs where the language was consistent across identities (either both speakers in English or both in Mandarin), clearly visible in the dissimilarity matrix in Fig. 2.1.A as red and green sub-matrices. This result confirms that subjects were able to use

acoustical information in the time-reversed sentences and were sensitive to overall acoustical differences between the two languages (Supplementary Information 2.5.1 and 2.5.2).

Crucially for our hypothesis, listeners also rated pairs of speakers of their own language as more dissimilar than pairs of speakers of the other language. The effect is highly significant and apparent as an interaction when ratings are split by speaker and listener group in Fig. 2.1.C. This effect is not driven by one subject group, as there is no main effect of subject group on overall ratings and the own-language effect is significant for each subject group individually. Nor is it explained by one of the sets of stimuli as the effect of speaker language on the ratings was not significant either. However, despite the absence of a main effect of listener group, the native-language bias appears to be stronger in the English listener group compared with the Chinese, reflected in the differences in effect sizes. This result may be explained by the fact that our Chinese participants had been resident in the United Kingdom for 9 mo on average at the time of testing, and had considerable functional experience with the English language. It has been demonstrated, for example, that non-native speaker identification performance improves over several days' worth of training (Perrachione and Wong, 2007).

Our results provide the first evidence, to our knowledge, of a LFE in adult participants in the absence of speech comprehension. These findings extend the results of Johnson et al. (2011), who observed a similar effect in 7-mo-old infants: In both cases, subjects had a limited understanding of the stimuli, yet they were more sensitive to identity differences in native-language pairs compared with non-native pairs. Interestingly, however, the infants in Johnson et al.'s (2011) study did not show a discrimination bias for reversed

native speech compared with reversed foreign speech, as our adult listeners did. The infants' comparatively lower experience with the phonology of their native language may account for this; specifically, whereas 7 mo of exposure may be sufficient to enable differentiation of native speakers uttering normal speech, it may be insufficient for the kind of fine-grained differentiation required under alien processing conditions, as in the case of reversed speech. Indeed, even school-aged children may not display adult-like performance in speaker recognition tasks (Mann et al., 1979), suggesting that they cannot use the information available in an utterance as effectively as an adult listener despite their substantial experience of their native phonology and their greater exposure to different voices, compared with infants. Therefore, it may be that infants do not yet possess the ability to extract information from an unintelligible speech signal to aid speaker discrimination and recognition, in ways that adults can, as shown in our discrimination results and previous recognition results (Bricker and Pruzansky, 1966; Van Lancker et al., 1985; Sheffert et al., 2002).

Thus, our findings refine Perrachione et al.'s (2011) view that "human voice recognition depends on language ability" by supporting the notion that phonological processing is the key aspect of language ability which facilitates speaker individuation, but adding that comprehension of the spoken message is not necessary for such individuation. Their findings suggest that impaired voice recognition in dyslexics may be driven by their known deficits in phonological processing (Gabrieli, 2009), whereas our results show that the limited phonological information and indexical cues preserved after time-reversal are sufficient to allow listeners to differentiate speakers. Moreover, extended exposure with the particular distribution of acoustical features characteristic of their own language allowed our participants to perceptually "zoom-in" on speakers of that language,

resulting in higher native-language dissimilarity ratings, even when both native and non-native speech content was unintelligible.

These findings draw an interesting parallel with an analogous effect in another sensory modality: the ORE in face recognition. The ORE is the well-known phenomenon where observers are typically poorer at discriminating and recognizing faces from a different racial group compared with their own (for a review, see Meissner and Brigham, 2001). One influential account of the ORE suggests that individual faces are represented as points in a multidimensional space whose dimensions are shaped by perceptual experience and code for diagnostic features (Valentine, 1991; Valentine and Endo, 1992). Own-race faces, with which an observer has more experience, become distributed more diffusely about the origin of the space (i.e., the average or prototypical face). Other-race faces, as a result of a different statistical distribution of features, are encoded in a less efficient manner due to a reliance on diagnostic dimensions for individuation optimized for own-race faces. Other-race faces therefore mistakenly appear more similar to one-another to the observer and this confusability between faces underpins the impairment in other-race recognition performance. Indeed, this model has found support at the behavioural, computational (Caldara & Abdi, 2006) and neurophysiological levels (Vizioli et al., 2010; Vizioli, 2012; Brosch, Bar-David, & Phelps, 2013). An analogous model could be invoked to account for our results and those of Johnson et al. (2011). One could conceive of a similar "voice space" where voices are encoded as points based on experience with indexical and linguistic attributes. Indeed, the behavioural and physiological relevance of such a voice space model has already been demonstrated (Latinus et al., 2013). Speakers of one's native language would, in such a framework, be represented in a more distributed manner, resulting in higher inter-speaker

discriminability than for speakers of other languages to which the subject has had less exposure and are therefore represented in a less differentiated, more compact manner. Such a model, while acknowledging that comprehension can modulate speaker identification, would be consistent with the many noted similarities between cerebral face and voice processing (Yovel and Belin, 2013).

2.5. Supplementary Information

2.5.1. Table of Acoustics

	Chinese (Mandarin) Speakers				*U.K. (English) Speakers*			
	F0 (Hz)	***SD of F0***	***Formant Dispersion (kHZ)***	***HNR (dB)***	***F0***	***SD of F0***	***Formant Dispersion***	***HNR***
	224.57	33.92	1.06	13.18	251.83	42.42	1.06	14.76
	229.67	40.18	1.08	12.15	218.16	25.58	1.07	12.72
	230.27	45.49	1.01	12.19	174.31	41.78	1.01	9.99
	199.87	52.77	1.03	13.44	206.22	24.78	0.99	12.32
	197.43	28.92	1.07	12.49	213.58	37.60	1.07	12.57
	225.05	59.44	1.07	11.30	214.76	37.95	1.07	13.57
	221.89	43.33	1.09	11.83	169.44	47.94	1.02	10.29
	246.55	66.40	1.05	10.81	198.31	43.36	1.06	13.30
	188.68	43.93	1.03	9.65	184.95	30.16	1.05	10.95
	223.12	31.36	1.08	13.63	211.42	30.33	1.07	13.71
	176.75	28.68	1.09	11.16	177.98	38.40	1.02	11.89
	200.36	28.21	1.04	16.61	171.43	38.07	1.08	11.71
	208.59	44.81	1.07	9.99	193.04	28.75	1.03	11.99
	160.02	30.20	1.05	9.70	157.03	26.66	1.09	12.07
	238.02	35.29	1.06	13.97	195.79	18.77	1.07	14.71
	203.66	29.44	1.09	12.17	200.39	47.14	1.05	13.11
	265.72	61.79	1.01	13.22	216.28	21.57	0.98	12.79
	234.91	47.58	0.99	13.60	220.33	58.99	0.97	10.00
	214.22	48.55	1.09	11.97	186.75	28.14	1.04	10.67
	272.41	46.95	1.07	14.06	181.99	25.90	1.04	12.98
Mean	**218.09**	**42.36**	**1.06**	**12.36**	**197.20**	**34.72**	**1.04**	**12.31**
SD	**27.59**	**11.68**	**0.03**	**1.70**	**22.53**	**10.30**	**0.03**	**1.41**

2.5.2. Methods

We extracted fundamental frequency (F0), SD of fundamental frequency, formant dispersion and Harmonics-to-Noise (HNR) ratio values averaged across a given sentence clip. This yielded a total of 400 values per feature (10 sentences × 40 speakers) which were then reduced to 40 (one per speaker) by averaging each speaker's values for that feature over each of the recorded sentences. Data from each feature were subjected to two-sample t tests, which indicated that Chinese and English speakers produced significantly different values for F0 [$t(38) = 2.62$, $P = 0.01$] and SD of F0 [$t(38) = 2.20$, $P = 0.03$]. No significant differences were found between the groups for formant dispersion or HNR (P values > 0.05).

Chapter 3: "Discrimination" of unintelligible language categories in superior temporal cortices*.

Abstract

It is now well-known that human listeners perform better in recognition of voices which speak their mother tongue, as compared to voices speaking in a foreign language. We have recently shown that this "Language-Familiarity" Effect (LFE) appears to be present in a speaker discrimination task, even when native and foreign speech are time-reversed (**Chapter 2** of this thesis; Fleming et al., 2014). In order to extend our behavioural results, we attempted to examine the neural basis of this LFE for unintelligible speech with functional Magnetic Resonance Imaging (fMRI). Native English and Mandarin listeners were scanned whilst listening to time-reversed clips of English and Mandarin speech while performing a pure-tone detection task. Using Representational Similarity Analysis (RSA), we probed whether native-language unintelligible speakers were better differentiated in the brain than foreign speakers, reflecting better inter-speaker individuation when the unintelligible speech is "familiar". While we did not find any evidence for such a representational scheme, we did find that bilateral superior temporal cortex (STC) participated in a differentiation of the two unintelligible language types, where univariate analysis revealed only a right-lateralized main-effect of Mandarin speech in both listener groups. Specifically, dissimilarity among pairwise brain responses to speakers was highest when a pair consisted of a response to an English speaker and a Mandarin speaker, a finding somewhat reflected in our previous behavioural results. These results show that sufficient information remains in time-reversed speech for brain-based differentiation of responses to different language classes, and adds to previous results demonstrating the utility of multivariate methods in probing neural representations to different sound categories.

*Fleming, D., Giordano, B. L, McAleer, P., Caldara, R., & Belin, P. (***In preparation***).

Parts of this work were presented at the 2014 meeting of the Organization for Human Brain Mapping (OHBM), and at the 5th International Conference on Auditory Cortex (2014).

3.1 Introduction

Human listeners are capable of recognizing voices, even under circumstances where important acoustical clues are disrupted. This facility has been demonstrated in the case of reversed speech, where the speech signal is temporally 'flipped', rendering the linguistic message incomprehensible (Bricker and Pruzansky, 1966; Van Lancker et al., 1985; Sheffert et al., 2002). Time-reversal, while disrupting intelligibility, preserves some elements of the raw signal, albeit in a mirrored form. For example, reversed speech is identical to natural speech in amplitude, duration, and mean fundamental frequency and the formant structure of long vowels remains unaltered. Listeners can demonstrably take advantage of this preserved information in speech perception tasks, involving, for example, word transcription (Binder et al., 2000).

Notably, however, most reports of successful reversed-speech speaker recognition employ stimulus material which contains acoustical elements familiar to the listener, in the form of the phonology of their native language. It is well known that familiarity with a language confers a benefit in speaker identification tasks; indeed, the so-called language-familiarity effect (LFE) has been repeatedly demonstrated across a range of tasks and language conditions (Perrachione et al., 2007; 2009; 2011). The demonstration of this effect is consistent with the similarly well-established interaction of linguistic and paralinguistic information in both speech perception and speaker recognition tasks. For example, in speech perception tasks, familiarity with a speaker's voice or variability in the amount of presented voice-indexical information can influence performance (Martin et al., 1989; Mullenix et al., 1989; Pisoni, 1993; Nygaard and Pisoni, 1998). Likewise, as discussed above, while listeners are capable of successfully identifying talkers from

reversed speech-clips, their performance is appreciably improved when they can understand what is being said, as in the case of natural speech (Bricker and Pruzansky, 1966; Van Lancker et al., 1985).

Clearly, as shown in the case of the LFE and other scenarios where speech manipulation leads to impaired speaker recognition performance, speech comprehension is important for voice identification. However, an outstanding issue concerns the extent to which the emergence of the LFE reflects an inability to comprehend speech, or rather simply reflects a rich experience of the phonological structure of one's own language, acquired over the lifespan, without a necessity for understanding of the spoken content. Compelling recent evidence supports the notion that stored phonological representations are of vital importance for successful speaker identification. Perrachione and colleagues (Perrachione et al., 2011) report that adult dyslexic listeners do not show the classical LFE, as neurotypical controls do. Rather, their native and foreign identification performance scores are similarly impaired, relative to the native-language performance scores of the controls. This study shows the importance of intact phonological representations for successful speaker recognition (although see Perea et al., 2014). On the other hand, a LFE is already apparent in infants before they can fully comprehend speech: 7-month-olds notice a speaker change in their native language but not in an unfamiliar language (Johnson et al., 2011).

In **Chapter 2** (see also, Fleming, Giordano, Caldara, & Belin, 2014), furthermore, we found that listeners appear to show heightened inter-speaker discrimination sensitivity for time-reversed native-speech clips, as contrasted with foreign-speech clips. While supporting the position of Perrachione and colleagues (2011) that impoverished foreign-language

phonological representations are the underpinning of the LFE - we also show, consistent with the results of Johnson et al. (2011) – that the effect is manifest even when listeners cannot fully comprehend the linguistic message. That is to say, the robust phonological knowledge acquired over decades might result in a speaker discrimination advantage, when the spoken utterance contains even mirror-reversed language cues, specific to one's mother tongue.

Taking all of these results into account, we suggest that the LFE might be mechanistically similar to the other-race effect (ORE) for face-perception (where individuals show better face recognition performance for members of their own race group), in that different foreign-speaking voices might be perceived more similarly than different native-speaking voices, as listeners lack robust foreign language phonological representations to aid voice individuation. Indeed, this theoretical account has been proposed previously, most notably in the work of Perrachione and colleagues (2007; 2009) who invoke a popular theoretical account of the "other-race effect" (ORE) for faces originally proposed by Valentine (1991). In this account, Valentine proposes that encountered faces are represented within a multi-dimensional "face-space" whose dimensions represent features which serve as diagnostic identity cues (e.g., hair colour, eye-shape etc). The dimensions of the space are elaborated as a function of experience with encountered faces and are optimized for those types of faces most frequently encountered by the viewer. It is proposed, therefore, that when a viewer attempts to encode and subsequently retrieve an out-group face, their face-processing apparatus will be reliant upon encoding dimensions which are sub-optimal for that particular type of face, leading to increased inter-face confusion and a subsequent decrement in recognition performance. A similar framework may explain the LFE – impoverished or non-existent

representations of foreign-language phonology might render it difficult for the listener to compute inter-speaker variability and would therefore lead to the classical LFE for identification.

In the present study, we wished to build on the results discussed in **Chapter 2** (Fleming et al., 2014) by examining whether asymmetric phonological experience affects the neural representations of unintelligible native and foreign talkers, in the manner proposed by Valentine (1991) and Perrachione and colleagues (2007; 2009). Little attention has been granted to the neural basis of the LFE, unlike the ORE for faces. Many recent investigations have focused on the neural bases of speaker identity (for reviews, see Belin et al., 2011; Schweinberger et al., 2014), but none has explicitly examined how such bases may be influenced by language familiarity.

Therefore, using functional Magnetic Resonance Imaging (fMRI), we employed a whole-brain, locally multivariate "searchlight" approach in tandem with Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), which allowed us to examine the state of inter-speaker representations in the brain. English and Mandarin-speaking listeners listened to time-reversed clips of English and Mandarin-speaking voices during fMRI scanning. The RSA approach allows us to explore correlations between neural representations (across the entire measured cerebral volume) and models which capture a hypothetical representational structure. Initially, we generated two simple binary models: 1) A model which we might call the 'Valentine-Perrachione/compression-dilation' matrix, capturing the theorized increase in perceived dissimilarity for native-speaking voices, relative to foreign-speaking voices; 2) A model reflecting increased cross-language dissimilarity, as compared to within-language similarity. This second model was

generated to highlight brain regions which might be more sensitive to cross-language category effects, accounted for largely by acoustical differences in the English and Mandarin stimuli. As reflected in the behavioural results presented in **Chapter 2**, listeners produce the highest dissimilarity ratings for speaker pairs which consist of different languages, indicating that they are sensitive to such acoustical differences. To explore the relationship between the pattern of behavioural ratings in Chapter 2 and brain dissimilarity representations, we also created a ternary "behaviourally-informed" model, which captured the differences between the three pairwise rating conditions. Note that the listener sample from **Chapter 2** participated in this fMRI experiment prior to completing the behavioural experiment reported in that chapter.

Firstly, we expected activation-dissimilarity profiles (particularly in the voice- and speech-sensitive superior temporal cortices) which correspond more to the cross-language model, reflecting sensitivity to the acoustical differences between the languages (for details, see section 2.1 and 2.5 of **Chapter 2**). Secondly, we hypothesized that native-language neural representations should be more dissimilar between speakers, particularly in brain regions sensitive to voice identity, reflecting an enhanced ability to individuate native-speaking voices. This hypothesis could be supported by a relationship between the cerebral dissimilarities and the compression/dilation model, or by a correspondence to the behaviourally informed model. While identity processing has classically been assumed to be right-hemisphere dominant (e.g., Belin and Zatorre, 2003; Formisano et al., 2008), we might also expect some involvement of left superior temporal cortices, which are implicated in various aspects of speech processing (Hickok and Poeppel, 2007; Price, 2009; Price, 2012), and which are also engaged in processing of time-reversed speech (Binder et al., 2000).

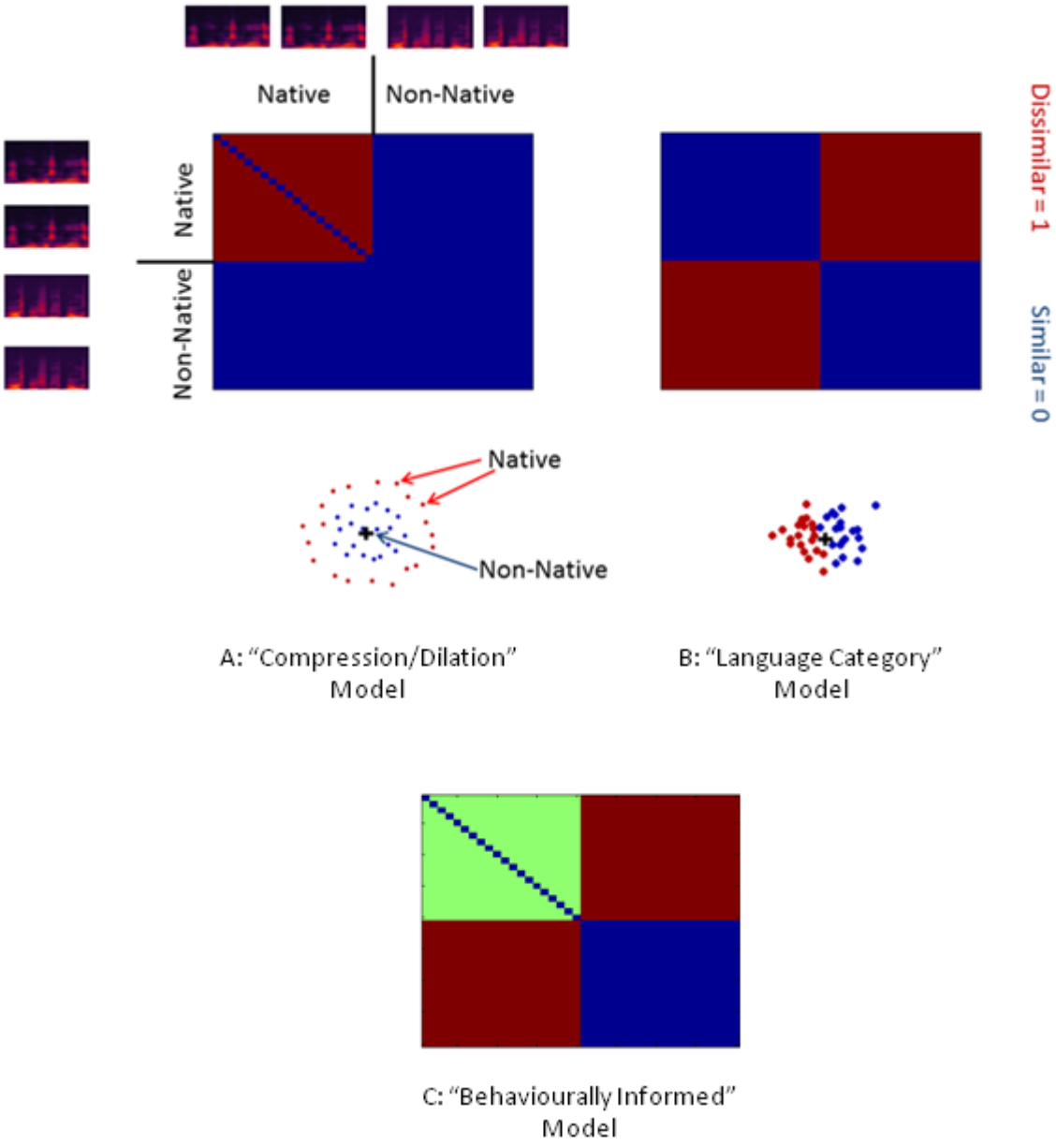


Figure 3.1.Illustration of predictor models used in the Representational Similarity Analysis (RSA). The simple binary "compression-dilation" (A; upper left) matrix shows the hypothesized dissimilar representation of different native-language voices and the relative similarity of foreign-language voices. The "language-separation" (B; upper right) matrix shows high cross-language dissimilarity and high within-language similarity. The "behaviourally informed" model (C; bottom) captures both the heightened perception of dissimilarity among native-native pairs, and the even higher perceived dissimilarity among inter-language pairs, as reflected in the behavioural ratings collected in Chapter 2

3.2 Materials and Methods

(N. B. Participants and stimuli from the behavioural experiment reported in Chapter 2 were used in the present study.)

3.2.1. Stimuli

Testing stimuli were drawn from a pool of 400 clips of 40 female speakers (20 native English-speaking and 20 native Mandarin-speaking) reading 10 sentences (Open Speech Repository, 2005). Recordings were digitized at a 16-bit/44.1 kHz sampling rate and cut into individual sentences. Sentence-length stimuli were subsequently time-reversed, standardized to duration of 1,250ms (from onset) and normalized for RMS amplitude. Stimuli were edited using Adobe Audition 2 (Adobe Systems, Inc.) and MATLAB 7.10 (R2010a).

3.2.2. Participants

20 Mandarin-speaking Chinese (8 female, mean age = 23.7, SD = 2.58) and 20 native English-speaking UK participants (10 female; mean age = 24.25, SD = 3.01) were recruited. Chinese participants' average duration of UK residency was 9.35 months (SD = 7.34) and all had attained a minimum score of 6.5 on the IELTS test of English as a foreign language, or a comparable score on an equivalent test. All participants were right-handed and reported no history of hearing difficulties or pathology, nor any familiarity with any of the recorded speakers' voices. Participants gave written, informed consent for their

involvement and received a monetary reward. The experiment was approved by the Ethical Committee of the University of Glasgow's College of Science and Engineering.

3.2.3. Model Dissimilarity Matrices

Based on our hypothesized within-category and between-category effects, we generated three model dissimilarity matrices. The first represented within-language dissimilarity (the "compression-dilation" matrix), and equalled 1 where two voices were drawn from a participant's native language and 0 elsewhere. The second matrix represented between-language dissimilarity (the "language-separation" matrix) and equalled 1 if two voices were drawn from different language categories (i.e., one Chinese voice paired with an English voice) and equalled 0 elsewhere. The third matrix was a "behaviourally informed" model which was based on the behavioural results presented in Chapter 2. Here inter-language cells equalled 1, native-native cells equalled 0.5, and foreign-foreign cells equalled 0. The assignment of these values was based on the average dissimilarity ratings obtained for each pair type in chapter 2 (i.e., cross-language pairs yielded the highest average behavioural dissimilarity ratings, followed by native-native and foreign-foreign pairs). As illustrated in Figure 3.1 (matrices and corresponding MDS representations), the compression-dilation matrix captures the hypothesized patterns of dissimilarity among responses to different native language voices (red matrix quadrant and diffuse MDS points -"dilation") and similarity among responses to different foreign voices (blue matrix sections and MDS points close together – "compression"). This model was intended to capture a representational scheme where one class of stimuli enjoys a more "differentiated" arrangement relative to another (Giordano et al., 2013). This is exactly the scheme which has previously been theorized for the ORE (own vs other race faces)

and the LFE (native vs foreign language voices) under the Valentine (1991) encoding scheme. However, this simple model ignores the high behavioural dissimilarity ratings for inter-language pairs in Chapter 3, as mentioned above, which is why the final behaviourally informed model was generated to account for this, in combination with the slightly enhanced differentiation of native-native pairs, relative to foreign-foreign pairs. The language-separation matrix, on the other hand, reflects sensitivity to language categories, where red panels illustrate dissimilarities among pairs of responses to speakers of different languages, and blue panels illustrate similarities among responses to pairs of speakers who share a language. These models were compared through the procedure detailed below to locally multivariate patterns of activity across the entire measured brain volume.

3.2.4. Procedure

Participants received nine blocks of auditory stimulation within one continuous scanning run. Each block consisted of all 40 speakers, one 1kHz pure tone, and lasted approximately 3 minutes. Following a rapid event-related protocol, the inter-stimulus interval (ISI) was randomly jittered between 1.5 and 2 seconds. Participants completed an anomaly detection task during scanning which required them to press a button on a response pad held in their right hand whenever they heard a pure tone (occurring once per block). In addition to these task instructions, they were advised to pay close attention to the presented voices and not to focus on what was being said, rather upon the sound of each individual's voice. No sentence clip was repeated within a block, and each participant received a unique speaker-to-sentence assignment. The sound stimulation blocks were separated by baseline (scanning noise only) periods lasting 20 seconds. The

main experimental run lasted approximately 25 minutes after which the participants underwent a voice-localizer scan and a T1-weighted anatomical scan, each lasting 10 minutes. Thus, the whole scanning session lasted approximately 45 minutes per participant.

3.2.5. fMRI Data Acquisition

MR data acquisition was performed with a Siemens 3-T Tim Trio scanner (Siemens, Erlangen, Germany), using a 32-channel head coil. Audio stimuli were presented through MR-compatible in-ear phones (Sensimetrics Corporation, USA) at a volume of ~90 dB SPL. The main experimental run consisted of a continuous, ascending, interleaved Echoplanar Imaging (EPI) sequence where the acquisition matrix was tilted to provide coverage of the entire temporal lobe and the inferior part of the frontal lobe (Voxel Size: 2mm x 2mm x 2mm; TR=2.5s; TE=39ms; FOV = 192mm x 192mm; Matrix Size = 96 x 96; Flip Angle = 82 degrees). Each volume consisted of 28 slices acquired in the transverse plane, with a 10% slice gap. Following the main functional run, participants completed a 'voice-localizer' scan, the results of which are not discussed here. This scan involved passive presentation of 20 blocks of human vocal sounds (continuous speech, syllables, laughs, cries and other physiological sounds) and 20 blocks of non-vocal sounds (environmental sounds, animal calls), each lasting 8s (Voxel Size: 2mm x 2mm x 2mm; 32 slices per volume with a 10% gap; TR=2s; TE=30ms; FOV = 192mm x 192mm; Matrix Size = 96 x 96; Flip Angle = 77 degrees). Finally, a high-resolution T1-weighted structural image covering the entire brain was collected (Voxel Size: 1mm x 1mm x 1mm; TR = 2,300ms; TE = 2.96ms; FOV = 256mm x 256mm; Flip Angle = 9 degrees).

3.2.6. fMRI Data Analysis

Data were preprocessed using SPM8 and further analysis was performed with custom MATLAB code. Functional and anatomical volumes were first re-oriented to the anterior-posterior commissure (AC-PC) plane, and functional images were subjected to slice-scan-time correction. Images were then spatially realigned using a 6-parameter affine transformation. Anatomical volumes were co-registered to the mean image of the functional time-series (generated during the spatial realignment phase) and segmented into grey-matter, white-matter and cerebrospinal fluid images. These segmentation parameters were then used to normalize functional volumes to the MNI (Montreal Neurological Institute) space, and normalized volumes were then smoothed with a Gaussian kernel (8mm Full-Width-at-Half-Maximum). The grey matter segmentation parameters were also used to generate participant-specific binary grey-matter masks, where voxels with a grey-matter probability greater than .5 were retained and converted to a value of 1, and all other voxels were converted to 0. Functional time series were high-pass filtered with a cut-off of 128 seconds (~0.0078Hz).

3.2.7. Univariate Analysis

Classical univariate analyses were conducted to examine whether listeners elicited enhanced cerebral responses to native unintelligible speech, as compared to foreign speech. Univariate analysis was conducted on functional volumes which had been spatially normalized and smoothed. First-level (fixed-effects) inference was performed by specifying participant-specific design matrices which consisted of separate condition regressors for onsets of Mandarin voices and English voices. Each participant's binarized

grey-matter mask (estimated following segmentation) was specified as an explicit mask for the analysis. Design matrices also included the six estimated realignment parameters as regressors of no interest. Stimulus onsets were convolved with a standard double-gamma model of the haemodynamic response function (HRF). Contrasts were estimated for the effects of all reversed speech conditions against baseline and for each individual speech condition against baseline. These contrasts were carried forward to second-level (random-effects) analysis which was performed using the GLM-Flex package (http://mrtools.mgh.harvard.edu/index.php/Main_Page#GLM_Flex). The second level analysis tested the main effects of speaker language (within-subjects effect) and listener language (between-subjects effect), and the speaker-listener interaction effect.

3.2.8. Representational Similarity Analysis

For the multivariate analysis, we used functional volumes which had been slice-time corrected and spatially realigned, but which remained unsmoothed and in participants' native image spaces. Participant-specific GLMs were run, involving separate condition regressors for each of the 40 individual speakers (collapsed across the 9 blocks) and including the six realignment parameters as nuisance regressors. Whole brain beta-maps were generated for each of the individual voices against baseline by convolving the stimulus onsets with the HRF. These beta maps were masked by the grey matter volume generated during the segmentation step. Next, we extracted representational dissimilarity matrices (RDMs; Kriegeskorte et al., 2008) which aim to quantify the dissimilarity between fine-grained patterns of BOLD activity elicited by different stimuli. Moving through a participant's entire brain volume, a spherical searchlight of 5.75mm radius was centred upon each voxel and the beta estimates for that voxel and

surrounding voxels were extracted. The radius of 5.75mm was chosen as it resulted in a consistent maximum voxel sphere count across participants, and because it was close to sphere radii previously recommended for searchlight mapping (Kriegeskorte, Goebel and Bandettini, 2006). The dissimilarity between responses to a given stimulus pair was calculated by subtracting the Pearson correlation coefficient between their associated activity patterns across voxels from 1, and this value was then mapped back to the searchlight centre, resulting in whole-brain maps of $1 - r$ values for each participant. Next, the association between brain RDMs and the predictor matrices was tested by computing – for each searchlight centre - the Spearman correlation between the $1 - r$ values for each stimulus pair and, independently, the 3 model matrices. This produced participant-specific rank-correlation maps which were then subjected to the variance-stabilizing Fisher-Z transformation, normalized to the MNI template and smoothed using a Gaussian kernel (8mm FWHM).

Finally, the significance of the association between brain and model dissimilarity matrices was tested with a random-effects approach (Carlin, Nili, Calder, Rowe, & Kriegeskorte, 2011; Carlin, Rowe, Kriegeskorte, Thompson, & Calder, 2012; Giordano, McAdams, Zatorre, Kriegeskorte, & Belin, 2013). The participant-specific maps of correlations between brain RDMs and the "compression-dilation", "language-separation" and "behaviourally-informed" model matrices were entered into separate random-effects t-tests intended to reveal where brain-model correlations significantly differed from zero. These initial tests were thresholded at $p<0.05$ (corrected for family-wise error) with an extent threshold of 20 voxels, and collapsed across groups, with subsequent between-group comparisons.

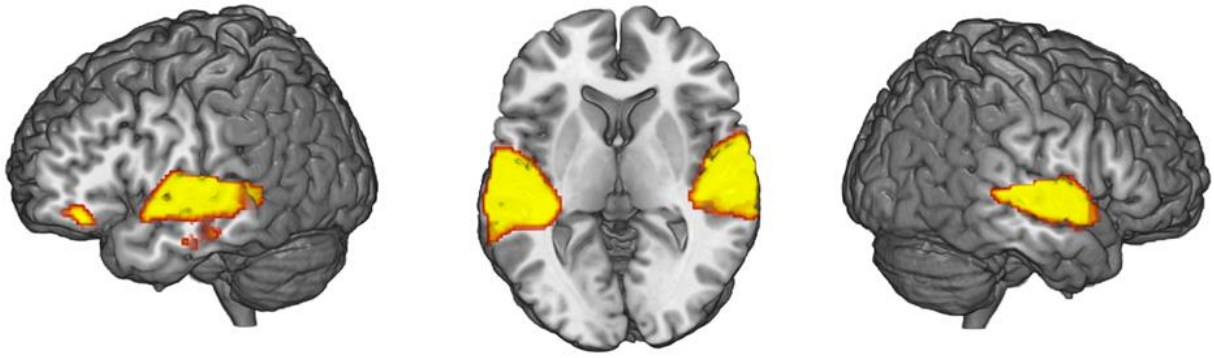


Figure 3.2 Thresholded cerebral responses to all reversed speech sounds against scanner-noise baseline averaged over listener groups. Reversed speech sounds elicited strong bilateral temporal lobe responses, indicating that speech stimuli were well-perceived by participants, despite the continuous scanner acquisition noise.

3.3. Results

Anatomical labelling of clusters emerging from all of the above analyses was performed by cross-checking the Harvard-Oxford cortical and sub-cortical atlases supplied in FSLView, and the AAL atlas supplied in the MRICron image viewer.

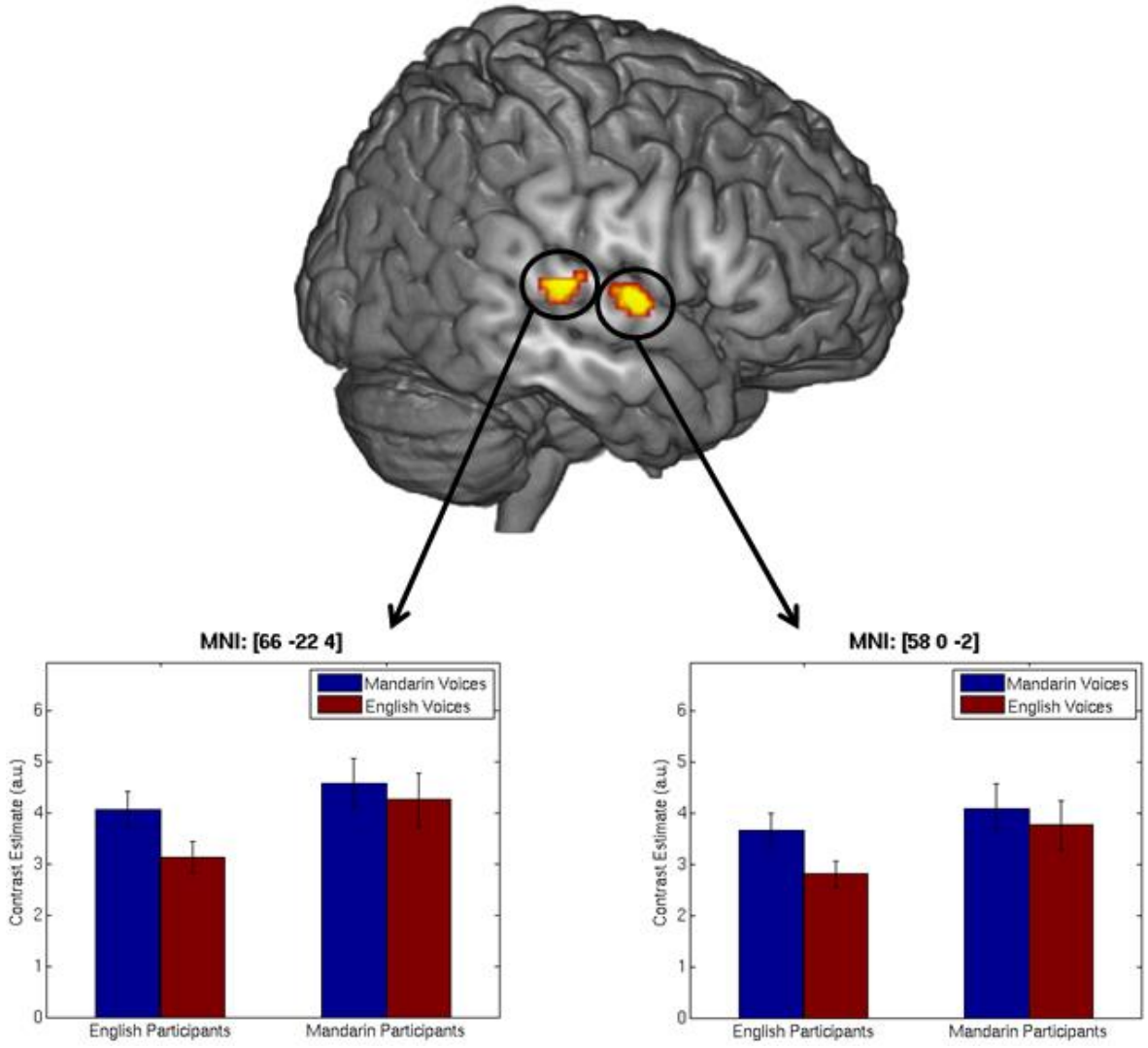
3.3.1. Univariate Analysis

The contrast of the effects of general reversed-speech stimulation against the scanner-noise baseline elicited broad bilateral activity throughout the temporal lobe (including Heschl's gyrus and large portions of superior temporal cortex), as summarized in table 3.1 and figure 3.2. The factorial analysis revealed no significant effects of listener group, nor a significant interaction effect. However, we did find a main-effect of speech condition: the contrast of Mandarin versus English reversed speech revealed two clusters in the right temporal lobe: one peaking in the right posterior STG (MNI coordinates of cluster peak: x = 66, y = -22, z = 4), and a more anterior STG cluster, peaking in the right Planum Polare

(cluster peak: x = 58, y = 0, z = 2), in which Mandarin speech elicited a stronger response than English speech, regardless of listener group. As no main effect of listener group was found, nor any significant interaction, this speech-condition effect appeared to be manifest regardless of the language provenance of the participants.

Anatomical Location	***x***	***y***	***z***	***T-value***	***Cluster Size (Voxels)***	***P-value***
L. PT	-58	-14	4	18.68	2250	<0.001
-	-56	-24	8	14.86	*Sub Peak*	<0.001
L. Heschl's Gyrus	-44	-22	8	13.86	*Sub Peak*	<0.001
R. PT	60	-14	0	16.84	1813	<0.001
R. PP	52	-6	-2	16.13	*Sub Peak*	<0.001
R. STG	64	-22	2	15.13	*Sub Peak*	<0.001
OFC/Frontal Pole	-44	38	-14	7.38	99	<0.001

Table 3.1 Cluster peaks and sub-peaks from the contrast of the general effect of reversed speech against scanner-noise baseline (regardless of language) collapsed across English and Mandarin-speaking participant groups. All reported clusters are associated with a p-value smaller than 0.05, corrected for multiple comparisons (FWE) with a cluster threshold of k = 20. The peaks of the Bilateral (l/r) cluster peaks included the Heschl's Gyrus (HG), Planum Temporale (PT), Superior Temporal Gyrus (STG) and Planum Polare (PP). OFC = Orbitofrontal Cortex. Millimeter coordinates (X, Y, Z) are in MNI space.



<u>Figure 3.3. Results of the univariate contrast of Mandarin reversed speech versus English reversed speech.</u> The contrast revealed activation in two clusters along the right superior temporal gyrus, which included the planum polare in the more anterior cluster ($p < 0.05$, FWE-corrected). Bar plots show the differences in response to Mandarin and English speech as compared to baseline (in arbitrary units) at the peak MNI co-ordinates of the Mandarin > English contrast.

Anatomical Location	*x*	*y*	*z*	*T-value*	*Cluster Size (Voxels)*	*P-value*
R. STG (posterior division)	66	-22	4	8.01	48	<0.001
R. STG/PP	58	0	-2	6.62	47	0.001

Table 3.2. Peaks of activation differences derived from the contrast of Mandarin speech > English speech, collapsed across Mandarin and English-speaking participants. A full-factorial analysis revealed no main effect of participant group, nor did it reveal any voxels significantly activated by the 'Participant x Speaker' interaction contrast. All results are significant at p<.05, FWE-corrected for multiple comparisons with a cluster threshold of k = 20. STG = superior temporal gyrus; PP = planum polare.

3.3.2. Multivariate RSA Analysis

Analysis of the brain-correlation maps revealed bilateral superior temporal clusters which were significantly positively correlated with the "language-separation" matrix (*p*s<0.05, FWE-corrected, 20-voxel extent threshold), reflecting higher dissimilarity between cross-language brain responses as compared to within-language responses. In the left hemisphere, we found two distinct clusters – one occupying the posterior portion of the STG, extending into posterior planum temporale (peak coordinates: x = -58, y = -26; z = 4), and the larger of the two occupying the mid-anterior STG, extending into planum polare, anterior planum temporale and overlapping with the antero-lateral edge of the Heschl's gyrus (peak coordinates: x = -60, y = -10, z = 0). The single cluster in the right hemisphere was situated upon the mid-posterior portion of the STG, extending into planum temporale and antero-lateral HG (peak coordinates: x = 64, y = -12, z = 2). We found no significant negative correlations with this model matrix; no significant positive or negative correlations with the compression-dilation model matrix; and no significant correlations

with the behaviourally-informed model matrix. Furthermore, we found no between-group differences in the brain- language-separation correlations.

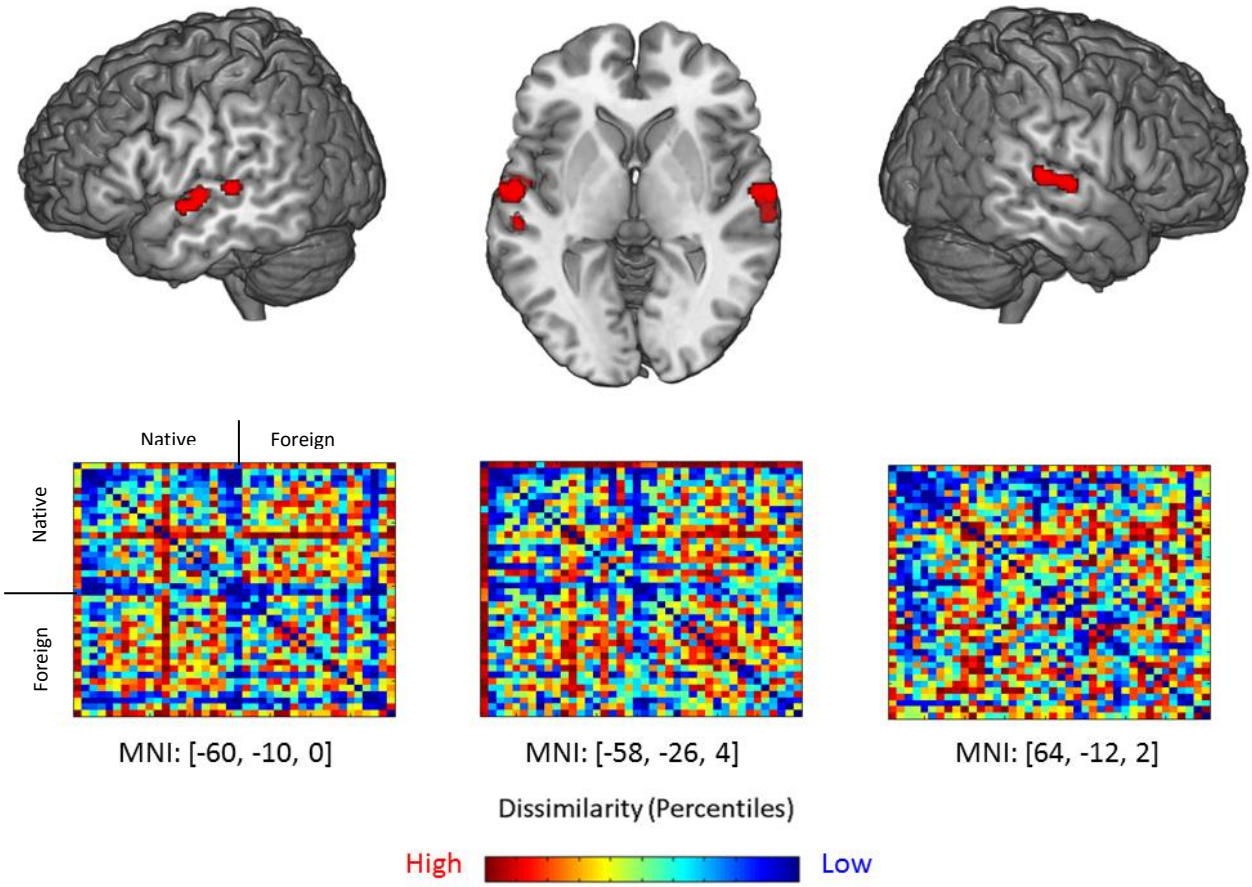


Figure 3.4 Extent of significant correlations between cerebral dissimilarities and the "language-separation" predictor model (red). All results are thresholded at p < 0.05, FWE-corrected at the voxel level. Dissimilarity matrices (collapsed across listener groups) extracted from the peaks of the random-effects analysis are shown in the bottom row. Raw 1-*r* scores were converted to percentiles. Brain dissimilarities at these locations were significantly positively correlated with the "language-separation" predictor model. No significant correlations of either sign were found between brain dissimilarities and the "compression/dilation" model or the "behaviourally-informed" model.

Antomical Location	*x*	*y*	*z*	*T-value*	*Cluster Size (Voxels)*	*P-value*
L. STG	-60	-10	0	6.40	165	0.004
-	-56	-4	-4	6.09	*Sub Peak*	0.008
R. STG/PT	64	-12	2	6.30	123	0.005
-	62	-24	6	5.86	*Sub Peak*	0.015
L. STG/PT	-58	-26	4	5.85	21	0.016

Table 3.3. Results of second level (RFX) analysis revealing areas where correlations between local multivariate patterns and the binary 'language-separation' model significantly differ from zero, across both participant groups. A two-sample t-test revealed no regions which differed in the magnitude of their relationship with this model as a function of participant group. No significant negative relationships emerged, nor any relationships with the 'compression-dilation' or 'behaviourally-informed' models. All results are significant at p<.05, FWE-corrected for multiple comparisons with a cluster threshold of k = 20. See Table 1 for key to abbreviations.

3.4. Discussion

We examined the cerebral responses to unintelligible English and Mandarin voices in a sample of native English-speaking and native Mandarin-speaking listeners. A classical mass-univariate random-effects analysis revealed a main-effect of speech condition: unintelligible Mandarin speech elicited stronger responses in right superior temporal cortical regions, regardless of the language provenance of the listeners. It is likely that this result is accounted for by inter-language differences in language acoustics, which remain present in the stimulus battery even after time-reversal (see table 2.5.1 in Chapter 2 and supplementary section 3.6 at the end of this chapter). For example, Mandarin speech typically has a more variable pitch contour than English speech (Eady, 1982; Mang, 2001), as is reflected in our time-reversed speech stimulus battery. Indeed, previous studies

have suggested that listeners with no experience of tonal languages show right-lateralized cerebral responses during lexical tone processing tasks in areas including the right posterior STG, supramarginal gyrus, and the inferior, orbital and ventrolateral frontal cortices. Conversely, tone-language speakers typically recruit a left hemisphere network in processing linguistically relevant tones (Hsieh, Gandour, Wong, & Hutchins, 2001; Klein, Zatorre, Milner, & Zhao, 2001; Pierce, Klein, Chen, Delcenserie, & Genesee, 2014; Zatorre & Gandour, 2008). As our time-reversed speech battery may have disrupted lexical access, but preserved the increased pitch variability which is characteristic of Mandarin speech, relative to English speech, it is unsurprising that we should see a right hemisphere involvement in the current results, regardless of the linguistic provenance of the participants. Notably, the more anterior of the two clusters showing the effect extended to the planum polare (PP), on the antero-lateral edge of Heschl's gyrus (HG), with the posterior cluster falling on dorsal STG. HG, PP and STG are responsive during pitch processing (Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002; Puschmann, Uppenkamp, Kollmeier, & Thiel, 2010; Zatorre, Belin, & Penhune, 2002; Zatorre & Gandour, 2008) which might explain the enhanced responses in these regions to a Mandarin stimulus containing more pitch variation than the English set. Interestingly, we found no interaction of speech condition and listener, which could indicate that univariate analyses may not be sufficiently sensitive to detect a correlate of the small behavioural effects reported in **Chapter 2**. Therefore, multivariate representational similarity techniques were enlisted to probe this issue.

3.4.1. Representational Similarity Analysis reveals discrimination of unintelligible speech categories in the bilateral superior temporal cortex

Using an RSA framework, we found evidence for a discrimination of different types of reversed speech in the bilateral superior temporal cortices. Our analysis revealed voxel clusters along the antero-posterior axis of the left STG and the right mid-posterior STG (overlapping with HG in both hemispheres) whose representational geometry correlated with our "language-separation" model, a predictor model coding for between-language dissimilarity. We found that these representations did not appear to differ as a function of participants' language provenance, as we observed no significant differences between the listener groups.

In contrast to traditional univariate analysis, our multivariate analysis revealed bilateral effects; in general, the spatial extent of effects derived from the multivariate analysis was greater than in the univariate analysis. The bilateral distribution of these multivariate results was similar to those reported in recent studies. For example, Okada et al. (Okada et al., 2010) used MVPA analysis to reveal above-chance classification of intelligible and unintelligible speech categories in core auditory regions within the dorsal plane of the STG (e.g., Heschl's Gyrus). These regions appeared sensitive not only to differences between intelligible and unintelligible speech, but also to differences *within* intelligible and unintelligible speech categories. For instance, these dorsal superior temporal regions could successfully differentiate responses to speech which had been subjected to spectral rotation and speech which had been both rotated and noise-vocoded. These manipulations produce speech categories which are largely unintelligible, but which differ in their respective acoustical profiles. Conversely, downstream (more ventral) auditory

regions in the bilateral STS appear more sensitive to speech intelligibility, as opposed to acoustical variations (what the authors refer to as “acoustical invariance”; although, these regions were, in fact, partially sensitive to acoustical differences within intelligibility categories). More recently, McGettigan and colleagues (Mcgettigan et al., 2012) presented a univariate analysis showing that areas in the bilateral STG responded to spectral and amplitude modulations in unintelligible (sine-wave speech) stimuli, while bilateral mid-posterior STS responded most strongly to the contrast of intelligible vs. unintelligible stimuli. In a subsequent MVPA analysis, they also showed that patterns within bilateral HG and STG/MTG could differentiate among responses to unintelligible stimuli with fixed amplitude and steady-state formant tracks, and unintelligible stimuli which dynamically varied in either amplitude or spectral content. Similar to Okada et al. (2010), classification of responses to intelligible vs. unintelligible stimuli were located (as opposed to the classifications based on acoustical categories mentioned above) in downstream regions, although here the authors report a leftward lateralization. Taken together, the above results from multivariate studies involving classification of responses to different types of acoustically-varying unintelligible speech are consistent with our observed results. Although we use a different multivariate technique (RSA vs MVPA), the interpretations are complementary: for RSA, the dissimilarity among brain responses in more dorsal bilateral STG (bordering on core auditory regions, such as the antero-lateral edge of HG in both hemispheres) is highest when comparing patterns elicited by two speakers speaking different (unintelligible) languages; and, for MVPA, the dissimilarity between responses patterns elicited by different types of acoustically-varying unintelligible speech allows a machine-learning algorithm to successfully tease apart the two categories, supporting good classification accuracy. With particular reference to the results of McGettigan et al. (2012), a bilateral preference for spectral modulations within

unintelligible stimuli, as compared to amplitude modulation was revealed within HG and STG/MTG. Our bilateral effects, as revealed with multivariate analyses appear, therefore, consistent with previous multivariate results showing discrimination of superior temporal cortex responses to acoustically varying categories of unintelligible speech (see supplementary section 3.6). Generally, in the above results, and in the present study, multivariate methods appear to yield greater sensitivity to subtle differences between response conditions which may not have been detected by univariate methods. With regard to our own results, however, we do not make strong conclusions regarding the differences between our univariate and multivariate analyses. For example, the differences may be entirely due to an insensitivity of multivariate methods to subject-level variability, which appears to be a crucial variance component in univariate analyses (Davis et al., 2014).

As has been described above, both our univariate and multivariate results were located within more dorsal parts of the STG, including planum temporale, planum polare and HG in some cases. The more dorsal parts of bilateral STG have been proposed as early cortical centres for the spectro-temporal analysis of speech (Hickok and Poeppel, 2007), and, as such, would likely be responsive to differences between the acoustical profiles of the English and Mandarin stimuli. For example, as has been stated in the previous chapter (**Chapter 2**; Fleming et al., 2014), English and Mandarin speech contain differing speech-sound inventories (Shu and Anderson, 1999; Yeong and Liow, 2012; Duanmu, 2000), and, while reversed speech stimuli are unintelligible, some of the information present in the original signal remains in a mirror-reversed form (Binder et al., 2000). The formant structure of vowels and pitch curves remain largely intact in time-reversed speech; for example, while reversed consonants become unintelligible, a reversed version of the

isolated vowel /a/ will still sound like /a/, albeit with a reversed pitch contour. Therefore, it is possible that inter-language differences in such information led to the observed patterns of inter-language response dissimilarity in dorsal STG.

The significant language-separation model correlation clusters also extended to the planum temporale in both hemispheres. The planum temporale is sensitive to acoustical dissimilarities among sounds, along dimensions such as pitch, loudness, and spectral centroid (or "brightness") and may be involved in the "abstract" representation of sound categories (Giordano et al., 2013; Staeren, Renvall, De Martino, Goebel, & Formisano, 2009). The PT has also been likened to a "computational hub" (Griffiths & Warren, 2002) which may match incoming sensory information to stored spectro-temporal representations. Interestingly, areas along the bilateral STG and PT also appear to respond more to degraded phonological information present in spectrally rotated speech, as contrasted with spectrally rotated speech which has also been noise-vocoded (Halai, Parkes, & Welbourne, 2015). In the present case, our category differentiation effects may again be explained by the differences in the speech-sound inventories of English and Mandarin, to which listeners remain sensitive even after time-reversal (**Chapter 2**; Fleming et al., 2014; supplementary section 3.6).

We did not, however, find the hypothesized correlation with the so-called "compression-dilation", nor the "behaviourally-informed" matrix in any brain location. As this appears at odds with the behavioural data from **Chapter 2**, we offer three possible explanations. One possibility is that, even with the increased sensitivity associated with multivariate group analyses (Davis et al., 2014), the behavioural effect is too small to be detected in our analysis; instead, the analysis picked up on the dissimilarity in listeners' representations

of the between-language acoustical differences, as reflected in the behavioural results in Chapter 2. Indeed, the highest dissimilarity ratings recorded by the participants in that study (who also participated in the present study) were gathered on "different-language" trials, where voice pairs consisted of one English and one Mandarin speaker.

An alternative explanation concerns the nature of the task presented to our participants. During fMRI acquisition, participants were instructed to listen to the sounds while performing a pure-tone detection task. While pure-tone detection tasks have been used in previous studies as a means of controlling attention, they may not be appropriate in studies where the main interest is in probing representations of voice identity. For example, as shown in previous studies of identity processing, the nature of the task (e.g., explicit voice identification) can modulate cerebral responses to voices (Von Kriegstein et al., 2003; Stevens, 2004; Von Kriegstein & Giraud, 2004; Bonte et al., 2009; Bonte et al., 2014). An explicit rating task similar to that used in Chapter 2, while difficult to operationalize in-scan, may have yielded results closer to those hypothesized. Nonetheless, in the sense that an apparent "separation" of the representations of language categories is evinced in the STC, we have observed at least a partial qualitative correspondence between the behavioural results presented in **Chapter 2** (where the highest collected pair dissimilarity ratings were recorded on trials involving two voices from different language groups) and the neuroimaging results presented here.

Finally, it is important to note that there are limitations associated with the scanning protocols used in the present experiment. Firstly, a continuous scanning protocol was employed and the presence of continuous acquisition noise may have impacted upon participants' ability to process subtle within-language inter-stimulus differences, leaving

them only capable of processing the more salient cross-language differences, as reflected in the results. Adoption of a clustered acquisition scheme ("sparse-sampling"/"sequence-with-gaps"), allowing sounds to be presented on a silent background, may have affected the outcome. Secondly, we employed a fast event-related design, which may be sub-optimal in terms of estimating responses for pattern analyses. With particular reference to investigations of speaker identity, previous multivariate studies have utilized slow event-related designs (e.g. Formisano et al., 2008; Bonte et al., 2014) in combination with acquisition gaps to allow presentation of stimuli on a silent background. Slow event-related designs allow for easier estimation of single trial responses, as the BOLD timecourses elicited by trials spaced apart in time do not overlap to the same extent as in rapid-ER designs, where stimuli are presented close in time (Mumford et al., 2012; Turner et al., 2012). While the multivariate pattern estimates we used in our RSA pipeline were based on presentation averages (i.e., 360 trials were reduced down to 40; one for each speaker, collapsed across sentences), adoption of either a slow-ER design with silent gaps, or a fast-ER design with silent gaps and an improved method for estimating trial responses (e.g., multi-parameter spatiotemporal response models as suggested by Turner et al., 2012) may have benefitted the present study.

A third limitation, related to the above consideration of a rapid-ER design, concerns the stimulus presentation schedule. In the present rapid-ER design, we did not include null events ("fixation" or "silent" trials) within the sequence of stimulation. The inclusion of null events within a rapid stimulation sequence improves design efficiency (Henson, 2006), and may have enhanced the potential of our protocol to detect subtle within-language representational differences, as compared to our present method of simply interspersing non-stimulation blocks between stimulation blocks.

As a final note, our data were acquired in one continuous run, lasting approximately 30 minutes. Recent evidence suggests that multivariate analyses are optimized by minimizing the length of presentation runs, for example, by using many short runs as opposed to fewer longer runs (Coutanche and Thompson-Schill, 2012). Therefore, dividing the protocol specified here into shorter segments may have been of benefit to our overall design efficiency.

3.5. Conclusion

Using multivariate analyses, we have shown that the dorsal parts of the bilateral superior temporal cortex (STC) differentiate between unintelligible speech from two different languages, regardless of the linguistic background of the listener. These findings partially correspond with behavioural results presented in **Chapter 2** (Fleming et al., 2014), which suggest that listeners are sensitive to differences in the speech-sound inventories of different language categories, even when speech stimuli are disrupted through time-reversal.

3.6. Supplementary Information

3.6.1. Supplementary Analysis: Accounting for the effects of stimulus acoustics

To investigate the role of stimulus acoustics in the observed neuroimaging results, we followed a pipeline described and developed by Giordano et al (2014) which aims to account for the impact of single-trial acoustical dissimilarities. Specifically, we re-estimated participant-specific beta images for the univariate and multivariate analysis pipelines based on fMRI data for which the variance explained by selected acoustical features (average f0, standard deviation of f0, formant dispersion, and harmonics-to-noise ratio; see supplementary section 2.5 in Chapter 2) had been removed. To remove variance explained by stimulus acoustics we fit new first-level models with a single non-baseline condition (i.e., stimulus onsets for all sounds were convolved with an HRF to form one "All Sounds" regressor), including head motion parameters, the intercept term and the same high-pass filter as in the original models. These first-level models also included 4 additional regressors (specified as parametric modulators aligned to the non-baseline condition), containing the trial-specific rank values of each of the measured acoustical features. These acoustical regressors had a value of zero for baseline trials and were convolved with the HRF. These participant-specific, first-level GLMs estimated the effects of acoustical variability on the differences in BOLD activation between different non-baseline trials without affecting the estimated baseline activity (Giordano et al., 2014). By extracting these acoustical regressors, we created images of the GLM prediction residuals (i.e., one 'residual' image for each image in the original unaltered fMRI time series) based only upon the trial-specific acoustical values (disregarding the baseline intercept term of the GLM), to which we then fit participant specific, first-level GLMs in

the manner described in the main methods section. For the univariate analysis, residual images were based on pre-processed fMRI data which had been normalized and smoothed. As with unaltered fMRI data, two regressors representing trial onsets for Chinese and English speech conditions were specified and the resultant contrast images were passed forward for second-level modelling (which in this case consisted of a t-contrast for the main effect of Mandarin vs. English speech, capturing the effects reported in the original analysis). For multivariate data, the prediction residuals were based on unsmoothed, native space data. In this analysis, 40 beta images were estimated, each representing the average of all activity for a given speaker identity, as for the unaltered fMRI data. These beta images were then used for an identical searchlight mapping procedure as described in the methods section.

3.*6.2 Results*

Regressing out the trial-specific acoustical variability (based on all 4 measured parameters) from the fMRI time series removed both the univariate effects (i.e., main effect of Mandarin speech vs. English speech) and the multivariate effects (i.e., correlation between neural dissimilarities and the "language-separation" model matrix) at the thresholds used in the original analysis (voxel-level FWE=0.05; cluster-extent = 20). Given that there were significant differences in only average f0 and standard deviation of f0 between the two speech stimulus classes (both significantly higher in Mandarin speech, compared with English speech; see supplementary section 2.5 in Chapter 2), we wished to determine whether either of these features alone might explain the observed results with unaltered data. Therefore, we created additional images of prediction residuals as specified above. However, rather than specifying sets of residual images

which removed the influence of all acoustical variability in an “omnibus” fashion, we independently focused only upon the effects of average f0 and standard deviation of f0. This resulted in one set of residual images for which variability explained by stimulus average f0 values had been removed; and another for which the variability explained by f0 standard deviation had been removed. As before, these images of prediction residuals were then used to generate first-level models identical to those used in the original analysis, the results of which were then considered in second-level analyses. For the multivariate data, the correlation between neural dissimilarities and the “language-separation” model was removed whether the analysis was based on average f0-residualized images, or standard deviation of f0-residualized images, indicating that these two parameters alone could explain the originally observed RSA results. The univariate effect of Chinese speech was removed when analysis was based on average f0-residualized images alone, but some significant voxels remained when analysis was based on residual images for which only the influence of standard deviation of f0 had been removed (see table 3.6.3). This suggests that the univariate effect of Chinese speech may have been more dependent on the higher average f0 values contained in the Chinese stimulus battery, as compared to the English battery. While f0 standard deviation (i.e., variability in f0/pitch across a clip) attenuated the effect, it did not fully remove it. We followed up this analysis by removing the variance explained by average f0 from the images which had been cleaned of variability associated with standard deviation of f0, and again fitting new first and second level models. Following removal of the variance explained by average f0 from the standard deviation of f0-residualized images, the univariate main effect of Chinese speech disappeared, confirming the role of average f0 values in the original effect. In sum, removing the variance explained by all of the considered acoustical features resulted in the disappearance of both univariate and

multivariate effects. However, both of these effects were removed when accounting only for the variance explained by average f0 and standard deviation of f0, both of which significantly differed, on average, between the two speech conditions (see supplementary section 2.5 in Chapter 2).

Anatomical Location	***x***	***y***	***z***	***T-value***	***Cluster Size (Voxels)***	***P-value***
R. STG (posterior division)	66	-22	2	7.19	20	<0.001
R. STG/PP	58	0	-2	6.60	38	0.001

Table 3.6.3. Peaks of activation differences derived from the contrast of Mandarin speech > English speech, (collapsed across Mandarin and English-speaking participants) following removal of the influence of stimulus f0SD values (see supplementary text 3.6.1-3.6.2 for further details). Note the reduction in the number of significant voxels, the reduction of peak t-values at both locations, and the – 2mm shift in the z-location of the primary peak, as compared with the original results shown in table 3.2. All results are significant at p<.05, FWE-corrected for multiple comparisons at the voxel level with a cluster threshold of k = 20. STG = superior temporal gyrus; PP = planum polare.

Chapter 4: Brain-based speaker identity decoding in a native and a foreign language*

Abstract

In everyday voice recognition, speech and voice information interact; an example of this interaction is the "Language-Familiarity" Effect (LFE) for voice recognition, where foreign language speaker recognition is impaired relative to the recognition of native language voices. Despite sustained interest in this effect, we are not presently aware of any studies of its neural underpinnings. Therefore, in the present study we sought to probe potential neural correlates of the LFE. Monolingual English speakers participated in an fMRI experiment while listening to English (native) and Mandarin (foreign) voices. We identified voice-sensitive regions in the temporal lobes and extracted local spherical volumes of pattern information from within these areas. Using multi-voxel pattern analysis (MVPA) we show that multivariate response patterns from voice-sensitive cortices enabled decoding of native – but not foreign – identities. Decoding in native-speech conditions was achieved from combined inter-hemispheric patterns and from patterns within a left voice-sensitive region which occupied part of the mid-posterior superior temporal cortex (STC) and the middle temporal gyrus (MTG) alone. Taken together, these results constitute what is, to our knowledge, the first evidence of a neural correlate of the LFE for voice recognition, in voice-sensitive regions which may overlap with speech-sensitive regions.

*Fleming. D., Giordano, B. L., Caldara, R and Belin, P. (***in preparation***)*

4.1 Introduction

Human listeners show a good facility for voice recognition, which can be realized across a range of different speech and non-speech vocalizations (for reviews, see Belin, Bestelmeyer, Latinus and Watson, 2011; Schweinberger, Kawahara, Simpson, Skuk and Zaeske, 2013). These voice recognition abilities have garnered much recent interest from the domain of cognitive neuroscience. Initially reported by Belin and colleagues (Belin et al., 2000), it is now well established that the superior aspect of the temporal lobe contains regions which respond preferentially to human vocal sounds as contrasted with other categories of non-human sounds, particularly in the superior temporal gyrus (STG) and superior temporal sulcus (STS), with a right hemisphere bias (Belin et al., 2000; Belin et al., 2002; Bonte et al., 2014; Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009; Grandjean et al., 2005; Pernet et al., 2015). Consequently, the temporal lobe has proven a popular focal region in investigations of the cerebral correlates of humans' voice recognition abilities. Early Positron Emission Tomography (PET) studies showed an increased engagement of the bilateral temporal poles when performing speaker identification tasks involving unfamiliar voices (Imaizumi et al., 1997) and right temporal pole during identification of familiar voices (Nakamura et al., 2001). More recently, Belin and Zatorre (2003) used Magnetic Resonance Imaging (fMRI) to reveal repetition-suppression (i.e., attenutated BOLD activity in response to stimulus repetition) in response to repeated presentations of the same speaker in the anterior portion of the right STS. Subsequent reports provide evidence of a broader cortical network sub-serving voice identity representation which may include portions of the bilateral superior temporal cortex (von Kriegstein et al., 2003; Stevens, 2004; von Kriegstein and Giraud,

2004; Andics et al., 2010; Latinus et al. 2011; Andics et al., 2013; Latinus et al., 2013; Bonte et al., 2014) and the right inferior frontal cortex (Andics et al., 2010; Latinus et al, 2011; Andics et al., 2013). Within this network several recent studies have localized superior-temporal voice sensitive cortices and have shown that these temporal regions play a role in a norm-based representation of voice identity (Andics et al., 2013; Latinus et al., 2013). Furthermore, distributed pattern information contained within these regions is sufficient to enable machine –learning algorithms to decode the identity of individual talkers (Bonte et al., 2014).

Despite this recent surge of interest in the processing of voice identity, fewer neuroimaging investigations take account of the known interactions between speech processing and voice recognition. For example, it is now well-established that speaker recognition performance is impaired when the listener has no knowledge of the language spoken by the talker. The Language Familiarity Effect (LFE) has been robustly demonstrated in a number of experiments involving various different language comparisons (Thompson, 1987, Goggin et al., 1991; Koester and Schiller, 1997; Perrachione and Wong, 2007; Perrachione et al., 2009) and even persists after intensive training in identifying foreign talkers (Perrachione and Wong, 2007). Previous research has focused on examining regions differentially involved in speech perception and voice recognition by contrasting cerebral responses elicited in response to these tasks (e.g., Belin and Zatorre, 2003; Stevens, 2004; von Kriegstein and Giraud, 2004). Furthermore, many of the aforementioned investigations of the brain bases of voice recognition use speech sounds – vowels, words or sentences - native to participants as stimulus materials. While native versus foreign language comparisons have been previously employed in functional neuroimaging to reveal those regions involved in aspects of native

language comprehension (e.g. Perani et al., 1996; Pierce et al., 2014), we are not presently aware of any investigation which has sought to investigate how language familiarity might interact with the neural representation of speaker identity.

Therefore, in the present study, we recruited monolingual native speakers of English in an fMRI study intended to probe the neural correlates of the LFE for voice recognition. During an initial behavioural session, participants were trained to identify sets of voices produced by three English and three Mandarin native speakers. Subsequently, they completed two sessions of functional imaging where they listened to the trained stimuli and performed a memory task which required that they attend to both speech and identity information present within the stimulus battery. Univariate methods are based on voxel-by-voxel differences in blood-oxygen-level-dependent (BOLD) contrast between conditions, and are insensitive to potential information carried by distributed response patterns which may be weak but consistent (Formisano et al., 2008; Staeren et al., 2009; Bonte et al., 2014). As such, we elected to use a multivariate decoding approach to examine whether pattern information contained in voice-sensitive brain regions enabled the decoding of native and foreign speaker identities. Mandarin speech was chosen as the foreign language contrast condition as it shares relatively little overlap with English - English belongs to the Indo-European language family, where Mandarin is a tonal language belonging to the Sino-Tibetan family. Furthermore, Mandarin has been used as a language condition in several key studies of the LFE (Perrachione and Wong, 2007; Perrachione et al., 2009; Perrachione et al., 2011).

Separate within-language classification schemes for both English and Mandarin speech were performed with voice-sensitive voxels as identified in an independent voice-localizer

scan. While voice sensitive areas in the temporal lobe appear to play a role in voice identity processing (Latinus et al., 2011; Andics et al., 2013; Bonte et al., 2014), they are also strongly responsive to speech sounds (Pernet et al., 2015), and, particularly in the left hemisphere, overlap with superior temporal regions which are implicated in the processing of intelligible speech (Evans et al., 2014; Mcgettigan et al., 2012; Narain, 2003; Scott, Blank, Rosen, & Wise, 2000). It is therefore possible that an algorithm's ability to decode speaker identity from activation patterns in these regions will be dependent upon the nature of the speech signal. As such, we hypothesize that speaker decoding accuracy will be higher in the English (native) as compared to the Mandarin (foreign) condition.

4.2 Methods

4.2.1 Stimuli

Speech material consisted of short subject-object sentences drawn from lists 1 and 2 of the Bamford-Kowal-Bench sentence battery (Bench, Kowal and Bamford, 1979). Mandarin translations of these sentences were prepared by a native speaker of Standard Mandarin Chinese. Sentences were initially recorded by 6 female native speakers of English (mean age: 25.8 years; SD: 2.49) and 4 native speakers of Standard Mandarin Chinese (mean age = 26 years; SD: 2). All speakers were non-smokers and reported no history of auditory or speech pathology. Recordings took place within a sound-attenuated chamber, using Adobe Audition software (Adobe Systems, San Jose, CA, USA) and a Microtech Geffel UMT 800 microphone (Microtech Geffel GmbH, Germany). Recordings were digitized at 44.1kHz sampling rate with 16-bit resolution. Stimuli were subsequently cut into individual sentences, and subjected to post-processing in Audition which involved

removal of transients (clicks and pops), automated noise reduction to remove obtrusive 50 Hz line hum, and the application of exponential ramps to the initial and final 100ms of the stimuli to remove sharp onsets and offsets.

Following initial recording and processing, we selected 3 speakers and 3 sentences from each of the two language conditions. Our selection was predicated upon minimizing the difference between the inter-speaker acoustical variability within each of the language conditions. Higher between-speaker acoustical variability in one language condition as contrasted with the other may render one set of stimuli more easily identifiable, so we sought to identify a set of stimuli and speakers which minimized this variability. To this end, we extracted summary acoustical statistics for each sentence clip using Praat software (Boersma and Weenink, 2015), including mean fundamental frequency (mean f0 in Hz), standard deviation of fundamental frequency (f0 SD in Hz); mean formant dispersion (i.e., the average distance between the first formant and the fourth formant, in kHz); Harmonics-to-Noise ratio (HNR in dB); jitter (i.e., local variation in fundamental frequency, defined as the average absolute difference between consecutive voiced intervals and measured in seconds); and shimmer (i.e., local variation in amplitude, measured in dB). Following extraction, we sampled for the stimulus set which minimized the difference in within-language variability in acoustical attributes between conditions. The difference in this variability was assessed by using the 'vartest2' function in MATLAB which implements F-tests of the difference of the variances between two samples. A separate F-test was performed for each selected acoustical feature, for a variety of different speaker and sentence combinations until a combination of stimuli was found where the individual F-ratios for all acoustical features were minimized. Small F-ratios indicate that the level of within-sample acoustical variance was comparable between the

final set of English and Mandarin stimuli. Following stimulus selection, all stimuli were equalized for duration to 1.09 seconds using the 'overlap-add' algorithm implemented in Praat. This particular duration reflected the mean of one English speaker's three sentence tokens and resampling to this duration resulted in no significant changes to the acoustical attributes of the speech tokens. This stimulus set was equalized for sound intensity, using root-mean square normalization. Finally, we ran another series of F-tests to ensure that length-equalization had no bearing upon within-sample acoustical variability and found no significant between-sample differences.

4.2.2. Participants

We initially recruited 10 native speakers of English (4 females; mean age: 20.1 years, SD: 1.96), all of whom were right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). All subjects reported no history of auditory pathology or hearing loss, no substantial musical experience, no experience with Mandarin Chinese, and no substantial experience with any other foreign language. One male participant did not complete the entire experimental program due to a technical error encountered during his first scanning session. Consequently, this participant was excluded and all reported analyses are based on data from 9 participants. All participants provided informed consent and received a monetary reward for their participation. The study was approved by the Ethical Committee of the University Of Glasgow's College of Science and Engineering.

Speaker	*Sentence*	*F0 (Hz)*	*SD of f0 (Hz)*	*Dispersion (kHz)*	*HNR (dB)*	*Jitter (µSecs)*	*Shimmer (dB)*
English 1	**1**	211.69	34.50	1.07	12.73	90.5	0.94
	2	206.43	18.25	1.06	16.85	57.4	0.72
	3	263.55	92.14	1.10	9.57	97.9	1.15
English 2	**1**	210.47	26.25	1.05	14.24	58	0.92
	2	205.25	27.99	1.00	16.64	47.4	0.88
	3	220.09	61.38	1.11	13.57	102.8	1.03
English 3	**1**	211.44	39.43	1.10	10.50	149.2	0.97
	2	204.11	22.94	1.10	15.57	66.9	0.78
	3	236.85	28.02	1.13	10.93	80.3	1.23
Mandarin 1	**1**	234.52	129.42	0.96	6.08	184.3	1.27
	2	258.93	106.26	1.04	5.59	124.6	1.20
	3	201.67	60.14	1.01	10.30	146.4	1.11
Mandarin 2	**1**	195.67	35.60	1.03	11.07	70.4	0.95
	2	225.65	54.03	1.04	7.26	127.1	1.19
	3	253.46	28.69	1.11	12.75	58.3	0.69
Mandarin 3	**1**	176.95	70.22	1.05	6.84	110.1	0.76
	2	249.46	68.81	1.03	9.24	92.4	1.09
	3	239.14	53.85	1.10	11.77	68.6	0.82

Table 4.1: Values of acoustical attributes extracted from the final set of English and Mandarin speech stimuli. Values were obtained from the stimuli following length equalization (to 1.09 seconds) with Praat's 'overlap-add' algorithm.

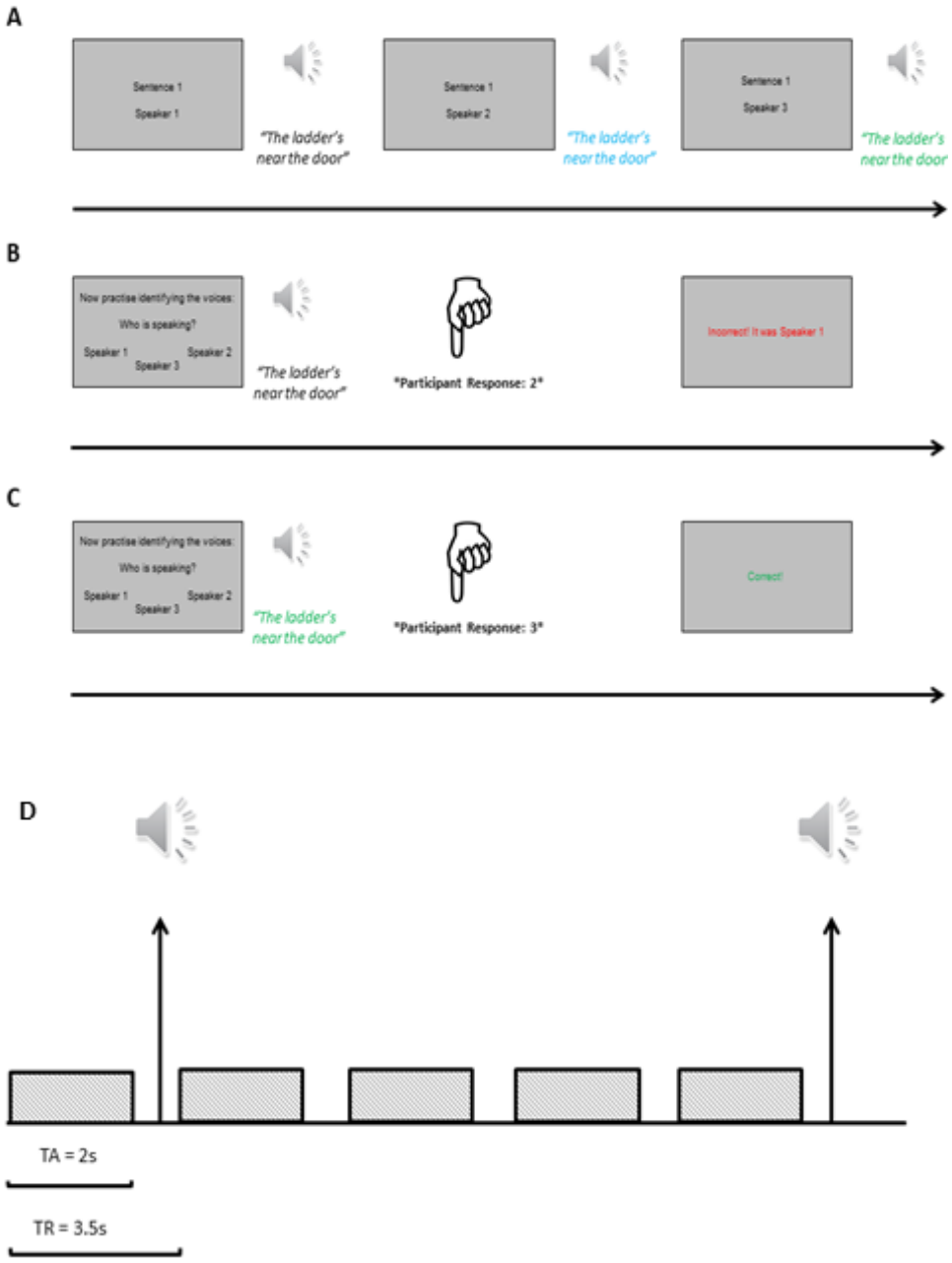
Acoustical variable	*F-ratio (English/Mandarin)*	*p-value*
Fundamental Frequency (f0)	0.47	0.30
Standard Deviation of f0	0.54	0.40
Formant Dispersion	0.80	0.77
Harmonics-to-Noise Ratio (HNR)	1.03	0.96
Jitter	0.58	0.46
Shimmer	0.59	0.48

Table 4.2: Results of F-tests comparing the level of within-group acoustical variance across the two stimulus sets, following length and amplitude normalization. English stimuli served as the numerator in these tests, and Mandarin stimuli as the denominator (degrees of freedom is ([n-1 = 8], [n-1 = 8]) for all tests, as nine stimuli were contained within each battery (3 speakers x 3 sentences). No F-test yielded a significant result, suggesting that the level of acoustical variance was comparable within between the sets of stimuli.

4.2.3. Procedure - Behavioural Testing

At the first session of the experiment, participants completed behavioural identification training. This training program was included in order to familiarize participants with the experimental stimuli and to examine whether they evinced a behavioural Language-Familiarity Effect. Testing took place within a sound-attenuated cabin and stimulus presentation was programmed in PsychToolBox (Brainard, 1997) in MATLAB (MathWorks, Natick, MA, USA). Stimuli were presented binaurally at a comfortable level.

The experimental program was very closely modelled on the identification paradigm presented in previous investigations of the LFE (Perrachione et al., 2007; 2009). Within a language condition, and for a given sentence, participants heard all 3 speakers reading the sentence in sequence. Following these initial readings, participants completed a short voice quiz, wherein a single voice was heard, and they were prompted to indicate which speaker had spoken. Speakers were labelled by number (e.g. “Speaker 1”) and participants recorded their responses using keyboard numbers 1-3. If they responded correctly, the phrase “Correct!” was displayed on-screen in green font. If they responded incorrectly, they saw the phrase “Incorrect! It was Speaker *n*”, in red font (where *n* referred to the number of the speaker who had actually spoken on that trial). This procedure was repeated for each speaker for each of the three sentences. Each sentence was practised 5 times in this manner. Subsequently, participants completed a final identification test consisting of all 9 trained items. They were prompted to identify the speakers as during the training phase, but on this occasion no feedback on their performance was provided to them. After this entire process was completed for one language condition, the procedure was immediately repeated for the second language condition. The order of language conditions was counterbalanced across participants and the entire training and testing program lasted for approximately 20 minutes. At all times (during the behavioural and scanning sessions), participants were instructed not to focus on the speech material, but, rather, to weight their attention towards the sound of the speakers’ voices.



<u>Fig 4. 1: Schematic of training procedure used during behavioural testing and slow event-related fMRI protocol.</u> *Panel A*: For each sentence, participants initially hear each of three speakers in a sequence. *Panel B*: Once all speakers have been heard reading the sentence, participants complete a short voice quiz where they hear each voice and practise identifying them. An "incorrect" trial is shown, where the participant has committed an error and corrective feedback has been issued. *Panel C*: Illustration of a correctly answered trial from the quiz phase. *Panel D*: Sound stimuli were played in a 1.5s silent gap occurring between volume acquisitions. Sounds were jittered according to scanning acquisitions, where an initial sound would be played and the subsequent sound would occur after 3, 4 or 5 scanning repetitions.

4.2.4 fMRI Scanning

Experimental sessions 2 and 3 each involved an fMRI session consisting of 4 runs. Experimental session 2 was undertaken on the day following completion of session 1 and session 3 was completed within 7 days of session 2. Prior to the commencement of each MRI scanning session, participants completed three rounds of practice identification in order to reacquaint them with the stimuli. The procedure for these trials was identical to that specified for the initial training phase completed during session 1, but did not include the final testing phase without feedback. Following training, participants completed 4 runs of fMRI measurement. Within each run, participants heard one block of English speech and one block of Mandarin speech, each consisting of 2 presentations of each single speech token (3 speakers x 3 sentences x 2 repetitions x 2 languages = 36 stimulus presentations per run). During scanning, participants listened to the speech stimuli while they completed a 1-back repetition task, where they pressed down on a button upon a response pad placed in their right hand when they heard consecutive presentations of the same speech condition (e.g., speaker 1 reading sentence 1 on consecutive trials). Repetitions occurred 4 times per run (two per block). The language condition presented first within a run was arranged according to the following scheme – MRI Session 1: A-B-A-B; MRI Session 2: B-A-B-A. The assignment of English and Mandarin to the A or B positions within the fMRI presentation schedule depended upon the order of presentation during the behavioural testing phase. For example, if English was the language heard first during training, then it would be the language heard in the first half of runs 1, 3, 6 and 8 (and the first presented condition during the pre-scan training in Session 2/the first fMRI session). Over the course of the MRI program, participants heard each individual speech token at least (discounting repetitions) 16 times.

4.2.5. fMRI acquisition

fMRI data were acquired at the Centre for Cognitive Neuroimaging at the University of Glasgow with a Siemens 3-Tesla Tim-Trio MRI scanner (Erlangen, Germany) and a 32-channel head coil. During each session, between runs 2 and 3, field-map scans were collected to correct for geometric distortion in the EPI volumes caused by inhomogeneities in the scanner's magnetic field. Four functional runs each lasting approximately 12 minutes were collected in both imaging sessions, using a standard Echoplanar Imaging (EPI) sequence with a sequential-descending slice sampling scheme (Voxel Size: 2.5mm x 2.5mm x 2.5mm; TR = 3,500ms; TA = 2,000ms; TE = 30ms; FOV = 195mm x 195mm; Matrix Size = 78 x 78). Each volume consisted of 35 slices (with a 25% slice gap) acquired in the transverse plane and the acquisition matrix was oriented obliquely along the sylvian fissure which provided near whole-brain coverage in all participants. Sounds were presented in stereo with MRI-compatible electrostatic in-ear headphones (Sensimetrics Corporation, USA) at a level of 90dB SPL in the 1.5s silent gap following the TA period. Stimulus onset occurred 100ms following the end of the TA, to allow separation between scanner noise and stimulation. Following a slow event-related protocol, the inter-stimulus interval was jittered between 13.9 and 19.9 seconds (corresponding to 3, 4 or 5 scanning repetitions). Following the 4 functional runs collected during MRI session 2, participants completed 2 runs of a 'voice-localizer' scan which involved passive presentation of blocks of human vocal sounds (continuous speech, syllables, laughs, cries and other physiological sounds) and non-vocal sounds (environmental sounds, animal calls), each lasting 8s (20 blocks per condition). Identical scan parameters and matrix positioning were used for the localizer scan runs, with the exception that TR was increased to 10s to allow presentation of the 8s sound blocks upon

a silent background in a sparse-sampling scheme. A high-resolution structural scan was collected between runs 2 and 3 of the first imaging session with a T1-weighted 3D-ADNI sequence consisting of 192 sagittal slices (Voxel Size: 1mm x 1mm x 1mm; TR = 2,300ms; TE = 2.96ms; FOV = 256mm x 256mm; Flip Angle = 9 degrees).

4.2.6. fMRI Data Analysis - Pre-processing

All pre-processing of MRI data was performed in SPM8. All images were initially reoriented so that the origin of the image space sat on the anterior-posterior commissure (AC-PC) plane. Functional volumes (with the exception of those obtained from the block-design voice-localizer scans) were then subjected to scan slice-timing correction, and were spatially realigned to the first scan of the first session, using a 6-parameter affine transformation. Functional volumes were then unwarped (using the field map scans) to correct for spatial distortion (Andersson, Hutton, Ashburner, Turner, & Friston, 2001). Next, T1-weighted anatomical images were co-registered to the grand-average EPI volume (generated during spatial realignment) and segmentation parameters for grey and white matter, and cerebrospinal fluid were extracted. The participant-specific grey-matter images estimated during this stage were used to create binary grey-matter masks, where voxels with a grey-matter probability lower than .2 were discarded and all others were retained. Functional time series were temporally filtered using a high-pass filter with a 128-second cut-off (~0.0078Hz).

4.2.7. Univariate analysis of voice localizer data

First-level univariate analysis of data from the voice-localizer scans was performed in SPM8. Following the pre-processing steps described above, localizer data were spatially normalized to the MNI space, and smoothed with a Gaussian kernel (8mm Full-Width-at-Half-Maximum). For each participant, onset regressors were constructed for blocks of vocal sounds and blocks of non-vocal sounds. First-level design matrices also contained the 6 spatial realignment parameters as regressors of no interest. Stimulus onsets were convolved with a standard double-gamma haemodynamic response function (HRF), and the contrast of 'human vocal > non-vocal' sounds was estimated (contrast vector: [1 -1]). Contrast vectors were replicated across runs, resulting in one image per contrast, averaged across runs. The resulting contrast images were carried forward to second-level (Random-Effects) analysis which consisted of a 1-sample t-test to identify voice-sensitive regions at the group level.

4.2.8. Random-effects analyses: statistical significance

We used an established cluster-thresholding procedure to determine statistical significance in our random-effects analysis (Obleser, Meyer, & Friederici, 2011; Slotnick, Moo, Segal, & Hart, 2003). Further details are provided in the authors' original paper (Slotnick et al., 2003; code available at http://www2.bc.edu/~slotnics/scripts.htm), but, briefly, this procedure involved imposing an initial voxel-level threshold of $p = 0.001$, uncorrected for multiple comparisons. Subsequently, a cluster extent threshold was estimated by simulating whole-brain activation. The entire functional image matrix was modelled, assuming a voxel-level threshold of 0.001 and smoothing the resulting map

with a Gaussian kernel. 10,000 such simulations were conducted, where the probability of a given cluster size was determined and the cluster size that corresponded to a p-value which was lower than 0.05 (corrected for multiple comparisons) was selected as the extent threshold. In this case, that threshold corresponded to 19 contiguous voxels (296.88 mm^3). Anatomical labelling for all results was performed by cross-checking peak MNI co-ordinates against the Harvard-Oxford cortical and sub-cortical atlases, and the AAL atlas supplied in the MRICron and FSLView image viewers.

4.2.9. Multivariate Analysis

Multivariate analyses focused on slice-time corrected and realigned functional volumes which were unsmoothed and remained in participant native space. In order to estimate trial-specific patterns of activity, first level design matrices were specified. These matrices comprised a single regressor for each speech condition, resulting in 36 regressors of interest per run (18 speech conditions, each presented twice within-run). Design matrices also included participant-specific realignment parameters, and regressors for stimulus repetitions. Stimulus onsets were again convolved with a standard HRF and a beta-image (consisting of whole-brain regression weights) was generated for each condition. Beta-images were estimated within each participant's binarized grey-matter mask.

4.2.10. Multi-Voxel Pattern Analysis (MVPA)

Regions-of-Interest (ROIs) were created by constructing spheres of 7.5mm radius around the peaks of voice-sensitivity as defined by the group univariate contrast of vocal vs. non-vocal sounds. This relatively restricted ROI sphere size was chosen to ensure consistency

in the number of voxels extracted for classification in each participant, and also to avoid any possibility of overfitting. The MNI coordinates of the peak voxels in each hemisphere were converted into the corresponding co-ordinate in participants' respective native image spaces, and used as the centre voxel for a native-space sphere. Trial-specific beta-maps (based on unsmoothed, native-space volumes) were then extracted from within these spheres, and were used as inputs for multi-voxel pattern analysis (MVPA).

MVPA was performed with the libSVM software package (Chang and Lin, 2011), using a linear support vector machine (lSVM) classifier with the cost parameter *c* set to 1. To ensure that our classification results were not merely based on overall activation differences between conditions, we scaled the data for each trial-specific beta-estimate by z-scoring, where the mean across voxels was subtracted from each voxel, and the de-meaned values were divided by the original cross-voxel standard deviation (Misaki, Kim, Bandettini, & Kriegeskorte, 2010). This scaling preserves the within-trial spatial patterns, but removes any differences in mean activation between conditions. This scaling was performed independently for each of the two voice ROIs, but we also included a condition where patterns from both hemispheres were fed to the algorithm, in the knowledge that multivariate identity representations may be carried inter-hemispherically (Formisano et al., 2008; Bonte et al., 2014). In this case, the response estimates for both ROIs were first concatenated and then scaled.

The analysis involved 3-way identity classifications which were performed separately within language conditions. The multi-class classification problem was converted into a series of binary classifications in a one-versus-one scheme. Within this scheme, classification is based upon pairs of conditions and predictions for testing trials are made

according to the condition which is most frequently selected by the binary classifiers. Speaker classification was performed by grouping the trials according to identity, regardless of which sentence had been spoken. Cross-validation followed a leave-one-run out scheme consisting of 8 folds, where the classification algorithm was trained on data from 7 runs and tested on its ability to classify identity trials from the run which had been held out. Accuracy was determined by calculating the mean score from each of the cross-validation folds. Statistical inference was based upon a permutation scheme, where each within-language classification analysis was repeated 4999 times with separately shuffled training and testing labels (Zhang, Kriegeskorte, Carlin, & Rowe, 2013). Label shufflings were held constant across participants, and on each permutation fold the group accuracy was estimated by computing the mean across participants. This yielded a distribution of 4999 group-averaged classification accuracies derived from shuffled labels. The group-averaged classification performance based on the true stimulus labels was added to this distribution, and statistical significance was determined by computing the fraction of shuffled-label group averages which were greater than or equal to the average accuracy obtained from classification with the true stimulus labels. Ultimately, this produced a p-value for each ROI, in each language condition.

Finally, we examined whether any of the ROIs, or combinations thereof, yielded classification accuracies which were higher for one language condition than the other. Our central hypothesis here is that language familiarity should prove beneficial to speaker decoding, and thus we examined whether English classification accuracy was significantly better than Mandarin classification accuracy. These analyses were performed by comparing the observed mean differences within ROIs with the distribution of mean differences calculated by subtracting the permutation distribution for Mandarin

classification from the permutation distribution for English classification. Significance was determined as before, by comparing the observed difference in accuracy between English and Mandarin classification for a given ROI with its associated permutation distribution of differences.

4.3. Results

4.3.1. Behavioural performance

English-speaking participants showed a native-language advantage in speaker identification (English mean: 0.95, SE = 0.04; Mandarin mean: 0.73, SE = 0.08; t [8] = 3.20, p = 0.01). Although participants' identification performance in the Mandarin condition was poorer than in the English condition, their identification of Mandarin speakers still exceeded the theoretical 33% chance level of performance (t [8] = 4.88, p = 0.001), indicating that they were capable of recognizing the speakers, although more poorly than in the English speech condition.

Participants performed the in scanner 1-back task to a high level (mean = 81.25%). While there was a slight trend towards greater accuracy during blocks of English speech (English mean = 82.64%; Mandarin mean = 79.86%), the difference was not significant (p = 0.63).

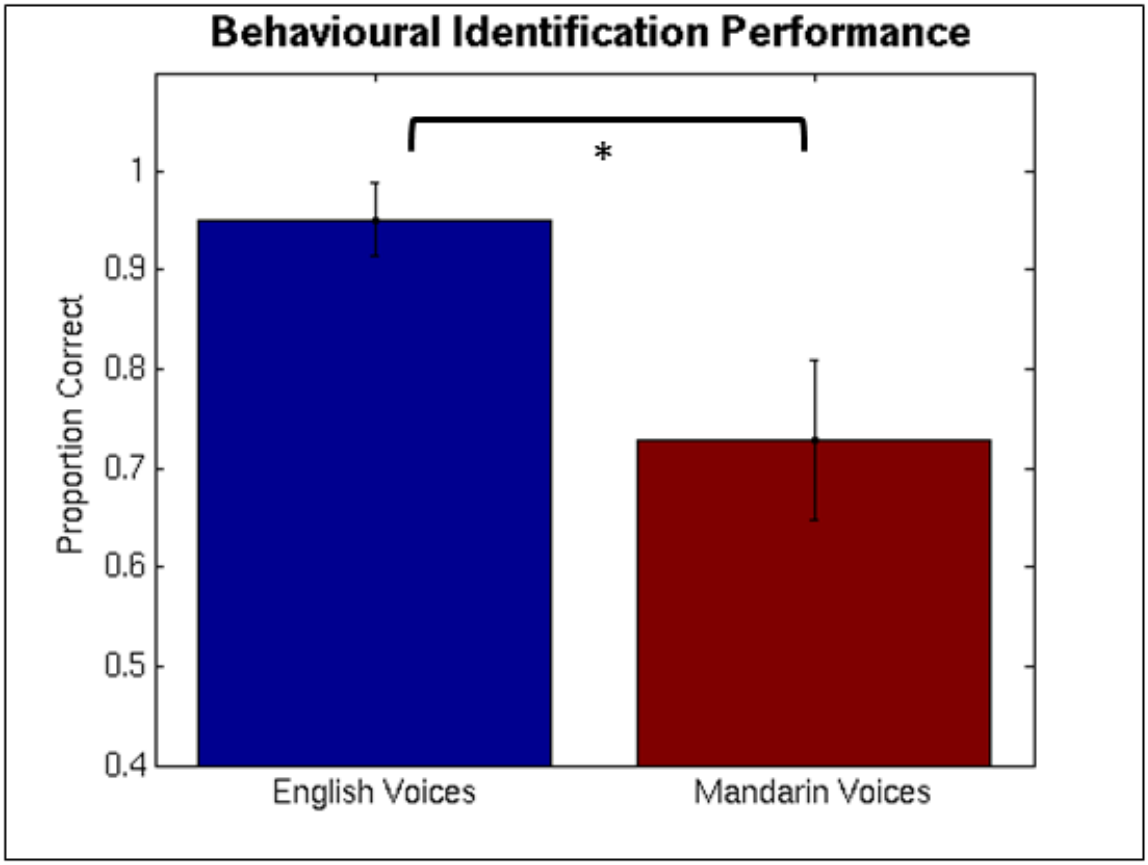


Fig 4.2: Behavioural identification performance from pre-scan voice training. English-speaking participants showed a significant recognition advantage for native-language voices (95% correct) as compared to foreign-language voices (73% correct).

4.3.2. Voice-sensitive ROI results: 3-way, within-language classification

A summary of voice-sensitive activation peaks is presented in table 4.3 below. Voice-sensitive regions were found in broad clusters throughout the bilateral superior temporal cortex (STC), with the largest left hemisphere cluster falling along the antero-posterior axis of the STG and STS (BA 21/22; MNI coordinate of peak: x = -60, y = -24, z = -2). A similar spatial distribution was observed in the right hemisphere (BA 21/22; peak: x = 70, y = -20, z = 2) with a peak in the lateral mid-posterior STG. Peak co-ordinates were transformed into individual participants' native image space and spheres of 7.5mm radius were constructed around the voice sensitive peaks to give two ROI masks of the Temporal Voice Areas (TVA). In the left hemisphere, this mask occupied portions of the mid-

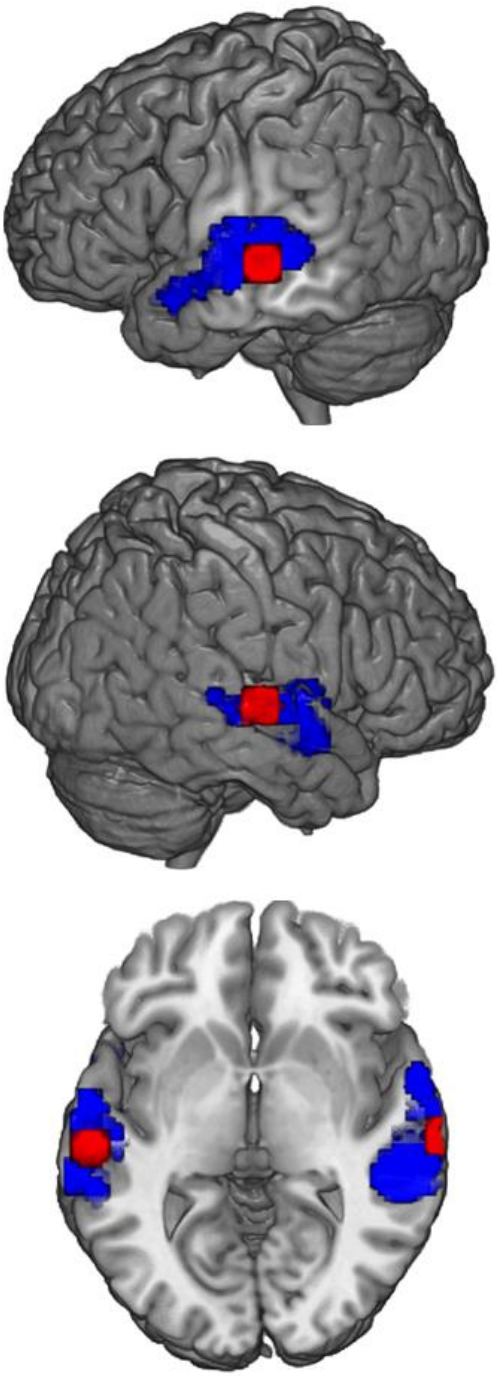
posterior superior temporal cortex (STC), including the STG, STS and middle temporal gyrus (MTG; BA 21/22); in the right hemisphere, the mask covered a portion of the more dorsolateral mid-posterior STG, with a ventral edge extending into STS and MTG (BA 21/22). The full extent of group activation which survived the cluster-thresholding procedure is also displayed below (figure 4.3), overlain with spherical ROIs drawn around peaks in both hemispheres for illustrative purposes (see also fig 4.4 and supplementary figure 4.6.1). Native-space beta weights were extracted from within the TVA masks, and within-language classification analysis proceeded as described above. Left and right voice areas were considered separately, and in combination.

English speaker identities were classified above chance-level (33.33%) in the left TVA ROI and in combined voice-areas, as shown in table 4.4 and figure 4.5 (left TVA = 35.96, p = 0.02; both TVA = 37.04, p = 0.003). Above-chance classification was not observed in the right voice ROI (34.10%, p = 0.28). Crucially, classification of Mandarin speakers did not exceed chance levels in any ROI, and was significantly lower than English classification in both the left and combined hemisphere analyses (both p-values = 0.01). Taken together, these results suggest that voice-sensitive voxels in the left temporal cortex carry more differentiated representations of different native-language voices as compared to foreign-language voices. Furthermore, the representation of identity in the native-language condition appears to be enhanced when pattern information from voice-sensitive regions in the right and left hemispheres is combined. A subsequent analysis involving voxels which were extracted from the full group TVA mask yielded similar results, where voxels in the left TVA and the combined TVA could classify only English voices above chance, albeit with no significant differences between English and Mandarin classification performance (English lTVA accuracy: 36.19%, p = 0.015; Both TVA: 35.49%, p

= 0.048; rTVA: 34.49%, p = 0.18).

Anatomical Location	*Cluster Size (mm^3)*	*x*	*y*	*z*	*T-value*
L. STS	9031.2	**-60**	**-24**	**-2**	**18.41**
L. STG	*Sub-peak*	-66	-14	0	14.67
-	*Sub-peak*	-68	-27	10	9.91
R. STG	9296.9	**70**	**-20**	**2**	**11.24**
-	*Sub-peak*	57	-4	-10	11.04
R. PT	*Sub-peak*	62	-12	2	11.01
L. PT	671.9	-43	-40	12	7.47

Table 4.3: Summary table of peak foci from group analysis of the independent voice-localizer scans Peaks of significant activation are shown (mm X, Y, and Z coordinates are in MNI space), along with t-scores and cluster extent. Bold row colouring in t-score and three-dimensional coordinate columns denotes an area which was used as a spherical ROI centre for classification analyses based upon voice-sensitive voxels. All p values < 0.001 uncorrected at the voxel level, with a cluster-threshold corresponding to p<0.05, FWE-corrected. Key to abbreviations: STG = Superior Temporal Gyrus; STS = Superior Temporal Sulcus; PT = Planum Temporale



<u>Figure 4.3. Voice-sensitive temporal lobe regions and spherical ROIs around activation peaks</u>. All regions where the contrast of human vocal > non-vocal sounds survived cluster correction are shown in blue. Voice sensitive areas are concentrated in the superior temporal cortex, including STG and STS bilaterally. Spherical ROIs from which classification data were extracted are shown in red around bilateral peaks of voice sensitivity in the mid-posterior superior temporal cortex.

4.4. Discussion

To our knowledge, we have conducted the first explicit investigation of the neural basis of the language-familiarity effect for speaker identification. In a sample of monolingual, native English-speaking listeners we provide evidence of a correlate of this effect, reflected in increased multivariate identity decoding accuracy of English speaking voices, relative to accuracy for Mandarin voices, in voice sensitive regions of the temporal lobe.

TVA ROI	*Voices*	*Mean Decoding Accuracy (%)*	*P-value*
lTVA	*English*	35.96 (1.56)	**0.02 (*)**
	Mandarin	31.95 (1.11)	0.87
	Difference (En – Ma)	4.01 (2.40)	**0.01 (*)**
rTVA	*English*	34.10 (1.22)	0.28
	Mandarin	33.18 (1.52)	0.57
	Difference (En – Ma)	0.92 (1.59)	0.30
Both TVA	*English*	37.04 (1.09)	**0.003 (*)**
	Mandarin	32.87 (1.20)	0.66
	Difference (En – Ma)	4.17 (1.90)	**0.01 (*)**

Table 4.4: Summary of classification results from voice-sensitive ROIs in participants' native image space. Classification accuracies and p-values for within-language analyses. English classification performance was greater than both the theoretical chance performance level (33%), and Mandarin classification performance when based on patterns extracted from the left hemisphere peak of voice-sensitivity, and when patterns from left and right hemisphere ROIs were combined. Asterisks denote a p-value smaller than 0.05, as determined by a group permutation test (see methods for details).

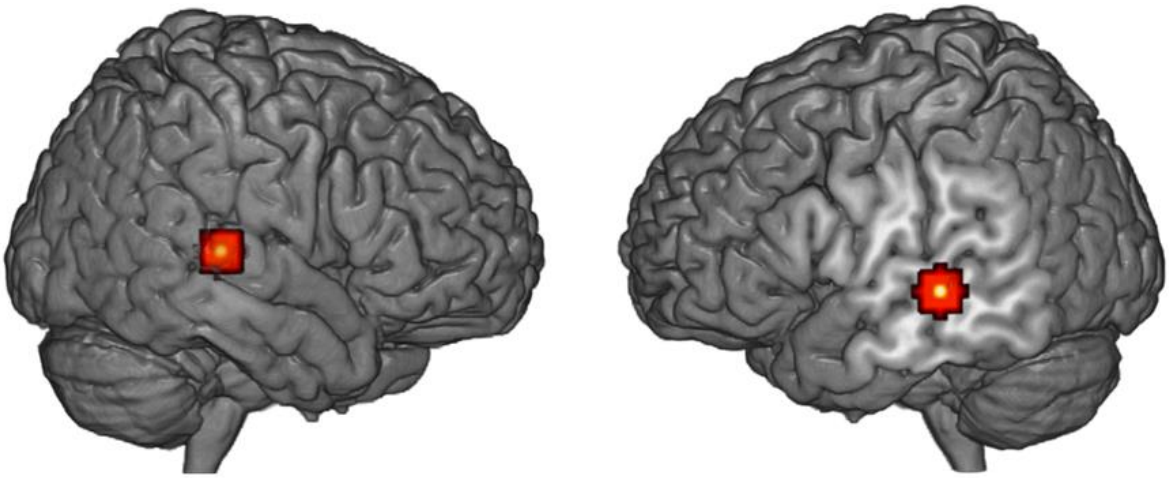


Fig 4.4: Voice sensitive ROIs rendered on a template brain. ROI centre voxels are shown in white. Millimeter coordinates are in MNI space.

4.4.1 Decoding of speaker identity from voice-sensitive voxels under native – but not foreign – speech conditions

Locally-multivariate response patterns in-voice sensitive regions of interest enabled our machine-learning algorithm to decode the identity of the eliciting voice. Crucially, significantly above-chance performance was only achieved in the English (native) language speech condition, but not in the Mandarin (foreign) speech condition. The strongest results were derived from the combination of bilateral voice-sensitive response patterns, but pattern estimates from the left hemispheric voice region alone also enabled above-chance native identity decoding, which was again significantly better than foreign identity decoding. The voice-sensitive ROI in the left hemisphere was centred on left mid-posterior STC, which appears to be involved in the processing of intelligible speech and phonological information (Chang et al., 2010; Desai, Liebenthal, Waldron, & Binder, 2008; Evans et al., 2014; Hickok & Poeppel, 2007; Narain, 2003; Okada et al., 2010; Price, 2010,

2012; Vigneau et al., 2006). The ROI also extended into left mid-posterior MTG, which, likewise, has been implicated in speech comprehension and the processing of lexical-semantic information (Abrams et al., 2013; Davis & Johnsrude, 2003; Hickok & Poeppel, 2007; Okada et al., 2010; Peelle, Gross, & Davis, 2013; Price, 2010, 2012). However, these regions have also been implicated in voice identity processing. Previous mass-univariate results have indicated that mid-posterior left STC regions are sensitive to acoustical variation in voices, and to learned perceptual identity shifts (Andics et al., 2010; Andics et al., 2013; Latinus et al., 2011). Additionally, the left mid-STS appears functionally connected to identity-sensitive areas of the posterior right STS during voice-recognition (Von Kriegstein and Giraud, 2004) and the left mid-MTG appears to be more responsive in voice recognition tasks as compared to recognition of speech envelope noises (Von Kriegstein et al., 2003). More recently, Bonte and colleagues (Bonte et al., 2014) demonstrated above chance decoding of voices from multivariate patters contained within voice and sound-sensitive voxels in the STC, and also showed that performance on a matching-to-sample voice recognition test was well-predicted by decoding accuracy from left STC patterns. Particularly germane to the present study are the recent results of Chandrasekaran and colleagues (2011), who used an fMRI adaptation design to show that a cluster peaking in the posterior MTG (which appeared to very slight overlap with STS) adapted (i.e., showed attenuated BOLD responses) to stimulus blocks which contained stimuli where both speaker identity and word information were repeated. In contrast, when either of these (identity – “who”; lexical – “what”) dimensions varied within a stimulus block, the response in the posterior MTG cluster was greater than that elicited by the stimulus repetition blocks. Critically, no difference was found in the contrast strength between the comparison of repeat blocks and blocks where only identity information varied (i.e., speaker identity changed but the same word was repeated), and

the comparison of repeat blocks and blocks where only word information varied (same speakers – different words). The authors interpreted these findings as evidence for an integrative function of this part of MTG, where speech and voice information are coupled. While it should be noted that the peak of the cluster in their analysis was located slightly more ventrally and far more posterior to our voice-ROI peak, which was found in the mid-posterior STS (in the present study: [-60, -24, -2]; in Chandrasekaran et al: [-62, -41, -4]), our spherical ROI did project into more posterior areas, and into dorsal parts of the MTG.

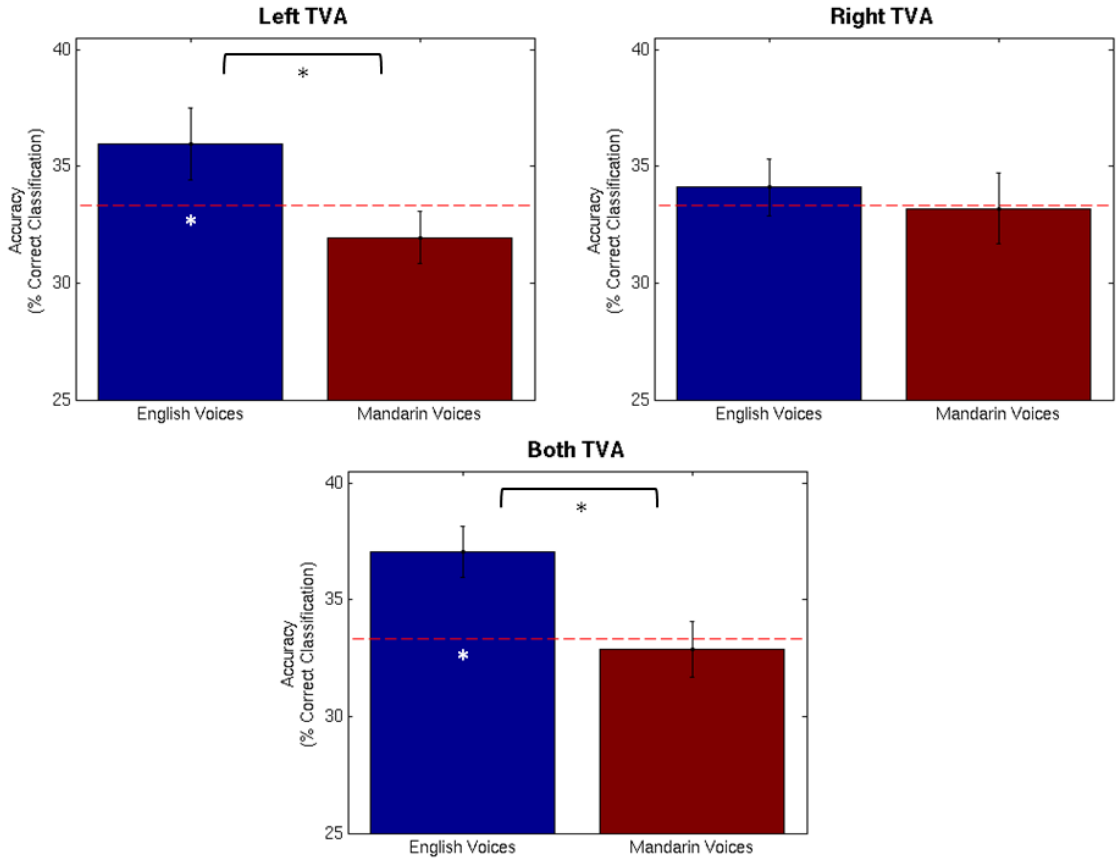


Fig 4.5: Three-way decoding results from voice-sensitive ROIs. Classification analysis was performed on data extracted from ROIs constructed around peaks of voice-sensitive activation as defined in an independent localizer scan. Above-chance decoding of English identities only ($p < 0.05$, white asterisks), was achieved in the left hemisphere voice ROI and when information was combined from bilateral ROIs. English decoding accuracy was significantly greater than Mandarin accuracy in these areas. Black asterisks above the bars denote significant p-values (< .05) associated with the English > Mandarin difference.

In the knowledge that left mid-posterior STC and MTG appear to contribute to the processing of native speech and voice identity under native speech conditions, it is possible that areas in the more ventral parts of the left-lateralized auditory stream which are sensitive to phonology and are involved in relating meaning to sound might also play a role in the integration of speech and voice information (Chandrasekaran et al., 2011). In other words, speaker individuation in the brain might be carried out by computing differences in voice/indexical information around stored representations of native-language speech sounds. To develop this idea, we refer to a theoretical account of the LFE proposed by Perrachione and colleagues (Perrachione et al., 2007; 2009). This account is in turn modelled on a framework proposed by Valentine (1991) for understanding the "other-race effect" (ORE) for faces, where faces from an observer's racial out-group are recognized more poorly than faces from their in-group. Briefly, in this account, a voice-space may exist in which incoming voice information is encoded based on acoustical variability around stored representations of native language phonological elements. In the case of speakers of a listener's native-language, this process is achieved readily. Where the speaker's language is completely foreign to the listener, however, phonological representations are impoverished, and the listener becomes reliant upon acoustical-based identity encoding alone without the benefit of canonical representations of native speech sounds to anchor identity computations. Indeed, Perrachione and colleagues (Perrachione et al., 2009) adduce this theoretical framework to explain an interesting behavioural finding: binaural native-language speaker identification performance is better predicted by right-ear monaural native-language performance, than left-ear performance. This right-ear advantage is interpreted as reflecting increased left-hemisphere engagement during a native-language voice recognition task, as contrasted with a foreign voice recognition task. Here, we show that left temporal lobe

regions which have been previously implicated in speech processing carry pattern information which, alone, and in combination with right temporal voice-sensitive regions, can allow native-language speaker decoding, but, critically, not foreign-language speaker decoding. Therefore, our decoding results provide a measure of support for Perrachione and colleagues' (2009) interpretation of their behavioural results. Specifically, brain-based decoding of native-language voices, while achieved from bilateral voice-sensitive areas, appeared to rely more on a voice sensitive region in the left mid-posterior STC/MTG which overlapped with areas previously implicated in speech processing (Hickok and Poeppel, 2007). Response patterns elicited by English voices which were carried within this region contained sufficient information to allow a decoding algorithm to label the eliciting voice at a rate above chance. It is possible, then, that English-speaking listeners were sensitive to subtle inter-talker indexical variations around English speech sounds which not only enabled them to recognize English voices better than Mandarin voices, but which also contributed to more efficiently "individuated" brain representations of individual English-speaking voices, enabling reasonably accurate brain-based identity decoding. Conversely, as listeners had no knowledge of Mandarin speech, such subtle individuation would prove more difficult. We would reiterate that we have attempted to control for inter-talker variability in the acoustical information contained within the English and Mandarin speech batteries. However, despite the comparable variances of the two sets, it is still possible that listeners could bring their knowledge of native speech sounds to bear in order to enhance their individuation of native language talkers, as compared to a set of foreign talkers.

4.4.2 Limitations

It is, however, worth commenting on a few caveats. Firstly, it is perhaps curious that classification analyses based upon pattern estimates in right hemispheric voice-sensitive ROIs alone did not allow for successful classification of either English or Mandarin-speaking voices, given previous reports of successful speaker decoding from right STC voxels (Formisano et al., 2008; Bonte et al., 2014). This could perhaps be due to the location of the right-hemisphere peak of voice-sensitivity, which was found in a very lateral part of the mid-posterior STG ([70, -20, 2]). While this voice-sensitive ROI did include portions of STS and MTG, it mainly occupied the mid-posterior STG, and never projected as medially as the left hemisphere ROI, due to its lateral peak. Previous results suggest that the centres of gravity of right STG/STS regions implicated in identity processing may be located more medially, with a broad anteroposterior distribution (e.g., Belin and Zatorre, 2003; Von Kriegstein et al., 2003; Von Kriegstein and Giraud, 2004; Formisano et al., 2008). There is also evidence for a segregation of processing within right superior temporal regions: while mid-posterior areas of the STS may be involved in the bottom-up processing of voices as complex acoustic targets, the anterior portion may respond more selectively in explicit identification tasks (Von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004). Similarly, other reports have suggested that the right anterior temporal lobe might represent voice identities in an acoustically invariant manner (Nakamura et al., 2001; Belin and Zatorre, 2003; Formisano et al., 2008; Andics et al., 2010). It seems, therefore, that identity representations may be distributed intra-hemisphere, beyond the small spherical portions included here, with different foci engaging with different aspects of processing. Multivariate decoding studies support this notion: identity representations may depend on distributed patterns which do not

necessarily arise from contiguous portions of cortex (Formisano et al., 2008) and some voice-sensitive voxels contribute more to the decoding of voice identity than others (Bonte et al., 2014). However, we do note that native decoding accuracy does slightly improve with the inclusion of right hemisphere voice-sensitive voxels. While the region cannot successfully decode identity alone, it may carry some information about the acoustical structure of voices given its mid-posterior location (Von Kriegstein and Giraud, 2004), despite its position relative to previous studies of voice identity, where peaks have been reported in more medial and ventral positions (Von Kriegstein and Giraud, 2004; Andics et al., 2010; Andics et al., 2013).

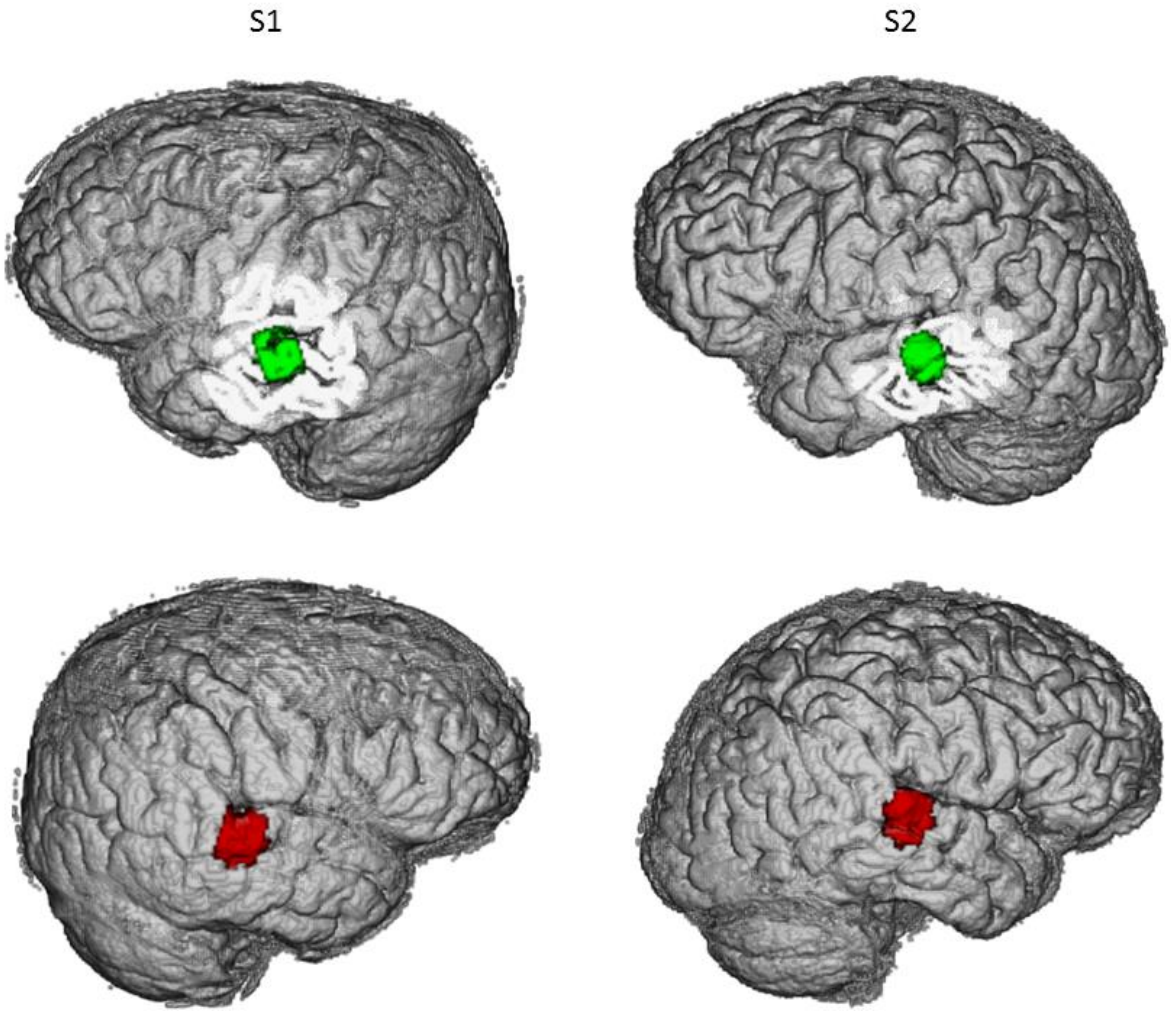
Finally, it should be acknowledged that we have focused our analysis on only one group of listeners. While we have striven to control the level of inter-speaker acoustical variability across stimulus groups, we cannot definitively rule out the possibility that our results are driven by asymmetric variability in an unmeasured feature in one set of stimuli relative to another. The decision to measure only one participant group was taken in order to ensure that those tested had no functional knowledge of the foreign-language under study. While behavioural investigations demonstrate that a LFE is still observable in Mandarin-speaking English learners (Perrachione et al., 2007; 2009), the magnitude of the effect is attenuated, relative to monolinguals. We elected for an approach where we study only a “true” monolingual group to avoid any confounds related to variability in language experience across participant groups. If further investigations in this domain are forthcoming, initially, at least, they would be best conducted under similar conditions in another country (e.g., monolingual Mandarin native-speakers in China).

4.5. Conclusion

We present results which provide a first indication (to our knowledge) as to the neural underpinnings of the LFE for speaker identification, in a group of monolingual, native English-speakers. Voice-sensitive voxels in the temporal lobe carry pattern information which can allow the decoding of the eliciting speaker's identity, but only under speech conditions native to the listener. Recognition of native-speaking voices may be supported by an ability to compute speaker differences around stored representations of native speech sounds carried by the left mid-posterior middle/superior temporal cortex. While future studies will address the caveats listed here by examining potential correlates in other speaker groups, the present findings provide a tentative indication that a listener's language knowledge may influence their cerebral representations of voice identities.

4.6. Supplementary Information

Figure 4.6.1. Spherical ROIs based on group-level voice-sensitive peaks. ROIs (7.5mm radius) are shown for two representative participants, projected onto native-space grey-matter masks. Left hemisphere ROIs are shown in green and right hemisphere ROIs are shown in red.



Chapter 5: General Discussion

In this thesis, I have attempted to address some unresolved issues relating to the Language-Familiarity Effect (LFE) for voice recognition. In **Chapter 2**, we have seen that an LFE for discrimination appears even when listeners cannot fully comprehend the linguistic message in an utterance. Specifically, a group of English and Mandarin native speakers rated pairs of unintelligible speakers as more dissimilar in their native language, as opposed to a foreign language. Speech intelligibility was disrupted via time-reversal; while this process precludes lexical access, it may also preserve important language-specific acoustical cues, such as pitch contour and vowel formant frequencies (Binder et al., 2000). Evidently, listeners are capable of taking advantage of this information in an unintelligible utterance, and bringing it to bear in a speaker discrimination task. This finding is consistent with reports of the LFE in young infants, who have immature language comprehension skills (Sundara and Kuhl, 2006; Johnson et al., 2011).

Next, I wished to determine whether a neural correlate of this apparent discrimination LFE in unintelligible speech could be found. In **Chapter 3**, the English and Mandarin participants from **Chapter 2** participated in an fMRI experiment where they were presented with English and Mandarin reversed-speech stimuli. Using a multivariate representational similarity analysis (RSA) approach, we compared the pattern of dissimilarity among cerebral responses to three hypothetical predictor models: Two of these reflected a pattern of higher inter-speaker neural dissimilarity in the native language condition, and lower dissimilarity in the foreign language condition; another reflected high cross-language dissimilarity and low within-language dissimilarity. While

we found no correlations between brain dissimilarities and the first two models, we did observe clusters of significant correlations with the cross-language model in localized patches of the bilateral superior temporal cortex. While multivariate approaches have previously been used to localize regions of auditory cortex which discriminate intelligible and unintelligible speech (Okada et al., 2010; Abrams et al., 2012; McGettigan et al., 2012; Evans et al., 2014), we show here that these methods can be applied to discriminate unintelligible speech in two different languages.

Finally, in **Chapter 4**, we present what is, to my knowledge, the first investigation of the neural basis of the LFE in intelligible speech. Using short English and Mandarin sentences which were controlled for inter-speaker acoustical variability across groups, monolingual native English speakers participated in an fMRI experiment. Spherical ROIs were defined around peaks of cerebral voice sensitivity in the bilateral superior temporal cortex. The patterns contained within these regions enabled a machine-learning algorithm to decode the heard voice identity in the English-language condition only. Furthermore, the left – temporal voice-sensitive region alone enabled above-chance native decoding, again significantly greater than in the foreign condition. Voice-sensitive regions, overlapping with speech-sensitive regions in the left temporal lobe may carry more differentiable representations of native language talkers, as compared to representations of foreign language talkers.

5.1. The LFE as an analogue to the ORE?

In the introductory chapter, I compared the LFE and the "other-race" effect (ORE) for faces. Discrimination effects appear for both at fairly early developmental stages: infants better discriminate native voices than foreign voice at 4 - 7 months (Sundara and Kuhl, 2006; Johnson et al., 2011); 3-month-olds show a preference for own-race faces (Bar-Haim, Ziv, Lamy, & Hodes, 2006; Kelly et al., 2005; Kelly, Liu, et al., 2007), and at 9 months, show an own-race advantage for face discrimination (Kelly, Quinn, et al., 2007). Both effects might also be attenuated by early experience: the magnitude of the LFE appears to covary with age of foreign language acquisition, where earlier learners show comparable rates of learning for native and foreign talkers, as compared to later learners (Creel and Bregman, 2014). In the case of the ORE, Sangrigoli et al (Sangrigoli, Pallier, Argenti, Ventureyra, & de Schonen, 2005) show that the ORE was reversed in Korean (East Asian) children who were adopted into French (Western Caucasian) families between 3-9 years old. Both effects have, furthermore, been shown to emerge across a range of different identification and discrimination tasks; for a review on the ORE, see Meissner and Brigham (2001) and see **Chapter 1** for relevant literature pertaining to the LFE.

Most germane to this thesis, however, is the manner in which both effects have been conceptualized in terms of similarity-based identity encoding models. Specifically, the model of Valentine (1991; Valentine and Endo, 1992), has been invoked previously in examinations of the LFE (e.g., Thompson, 1987; Perrachione and Wong, 2007; Perrachione et al., 2009) and the ORE (Valentine, 1991; Caldara and Abdi, 2006). To re-iterate, Valentine's face model posits that a face is encoded in a multidimensional space

whose dimensions represent relevant diagnostic features. These features appear to be elaborated over the course of an individual's perceptual experience, and, as such, may be tuned for dimensions which most reliably discriminate identity in the cultural group to which the has acquired the most experience, over the course of their life. Consequently, when an individual attempts to perform recognition tasks on a set of faces belonging to an unfamiliar cultural group, performance may suffer, as the face-space is not optimized for these faces. Valentine presented both exemplar and norm-based explanations of this model: in the norm-based version, the centre of the face space is occupied by a "mean" or "prototype" face, which represents the average of all faces encountered (weighted in terms of the number of encounters) over the course of the individual's experience. New faces are encoded with reference to this prototype, and their position in the space may be determined by their distance from (or similarity to) the mean. Therefore, typical faces will be represented closely to the average, whereas distinctive faces will be more distant. Importantly, as the dimensions of the space may be tuned for faces of an observer's own race group, newly encountered own-race faces will be better individuated and encoded in a more distributed manner than other-race faces within the face space. In the exemplar-based version, the centre of the space is unimportant; faces are encoded according to their (dis)similarity to other faces in the space. However, again, as own-race faces are better individuated, they will enjoy a richer distribution around the dimensions of the face space, in contrast to other-race faces, which will be tightly clustered, due to poor individuation. It is exactly this type of representational asymmetry which is proposed to underlie the "they-all-look-alike" effect in other-race face perception, which may lead to other-race face recognition decrements (Feingold, 1914; Vizioli et al., 2010).

Three relatively recent findings provide compelling evidence for the plausibility of this model with regard to the ORE. Caldara and Abdi (2006) performed neural network simulations where auto-associative models were trained on either East Asian or Caucasian faces. When a network was trained on a particular face-race, its projections of different faces from that race group were more distributed than on the untrained face-group (note that this was the case for both East Asian and Caucasian learning). Furthermore, when performing gender classification on untrained or "other-race" faces, the networks made more mis-categorizations than they did with trained or "own-race" faces. This work provided an important quantitative characterization of the Valentine model.

At the neurophysiological level, the model has enjoyed similar support. Using an EEG-Repetition Suppression design, Vizioli and colleagues (Vizioli et al., 2010) found that neural adaptation was increased for repeated presentations of individual own-race faces only. More specifically, when own-race face identity was held constant within a paired trial, a response within the temporal window of the face-sensitive N170 component was attenuated, but not for different own-race identities. However, differences in adaptation strength did not arise between same-identity pairs or different-identity pairs in the other-race face conditions, indicating that the brain represents other-race identities less efficiently. In addition, the same lead author (Vizioli, 2012) conducted an fMRI experiment where responses to individual own- and other-race faces were extracted from within the face-sensitive fusiform face area (FFA). Dissimilarity among these individual face responses was measured by computing 1 minus the pearson's correlation across FFA voxels, and computing the average difference in own- and other-race dissimilarity. Notably, in the dominant FFA of the observers, higher dissimilarity was observed for own-

race faces than for other-race faces. Taken together, these studies provide quantitative support for Valentine's dimensional model, and evidence of its neurophysiological plausibility. Note, however, that although Caldara and Abdi (2006) quantify the dimensional space (in this case reducing it to projections onto the first three eigenvectors of the neural network model), they do not apply specific categorical labels to the dimensions (e.g., does a dimension represent the hair or the eyes?).

As has been discussed in the introductory chapter, a conceptual model of this nature has been applied by others to the LFE (Thompson, 1987; Perrachione and Wong, 2007; Perrachione et al, 2009). Following recent evidence from behavioural and neurophysiological investigations of voice processing, voices may be represented about a central prototype in a similar way which is proposed to underlie the norm-based variant of Valentine's model (Bruckert et al., 2010; Latinus and Belin, 2011; Andics et al., 2010; Andics et al., 2013; Latinus et al., 2013). The space may include dimensions relating to acoustical properties of voices such as fundamental frequency, formant frequencies and harmonicity (Baumann and Belin, 2010; Latinus et al., 2013). The dimensions of the space might also, however, be affected by early linguistic input; infants perceptually narrow to the sounds of their own language at a very young age, moving from "language-universals" (Gomez et al., 2014) to a point where they express a heightened sensitivity to speech sound changes in their nascent mother tongue, as opposed to foreign speech (Kuhl, 2004; Kuhl et al., 2008). Therefore, the voice space model should flexibly account for the language spoken by a talker, by utilizing the listener's phonological representations of their native language. Where a listener cannot recruit 'canonical' representations of how the phonemes which form words in a language should sound, then the ability to

individuate speakers is hindered, indeed whether from single words (Zarate et al., 2015) or from running speech (Perrachione et al., 2011).

Notably, although this representational similarity scheme has been proposed by different authors, none have explicitly tested it for the LFE, unlike the ORE. Behaviourally, **Chapter 2**, does exactly this, by showing that listeners rate native-language voice pairs as more dissimilar in their native language, even when they cannot understand speech. Neurophysiologically, **Chapter 4** provides some evidence that cortical identity representations might be better differentiated in a listener's native language, as contrasted with a foreign language. Indeed, it is this heightened differentiation which enables a classification algorithm to evince higher accuracy in one condition over another. Notably, a left-lateralized voice-sensitive ROI permitted above-chance decoding only for speakers of a listener's native language. This ROI projected into the mid-posterior STC, which may play a role in phonological processing and the processing of intelligible speech (Narain et al., 2003; Hickok and Poeppel, 2007; Okada et al., 2010; McGettigan et al., 2012; Evans et al., 2014), and the mid-posterior part of the middle-temporal gyrus (MTG) of the left temporal lobe, anterior to a region which shows reduced BOLD adaptation to variation in both speaker and word information relative to repetitions of both information (Chandrasekaran et al., 2011). In addition to their respective involvement in phonological and lexico-semantic processing (Hickok and Poeppel, 2007; Price, 2009; Price, 2012), therefore, mid-posterior STC and MTG may play an integrative role in processing speech and voice information. Specifically, as has been proposed previously (Thompson, 1987, Perrachione and Wong, 2007; Perrachione et al., 2009; Perrachione et al., 2011), stored representations of a language's phonology might enable a listener to better individuate talkers, depending on indexical variability around those stored representations. In other

words, knowledge of a language's sounds may enhance the listener's ability to tell talkers apart, as knowledge of that language's sounds will allow the listener to compute inter-talker variations in pronunciation. Where such speech-sound representations are absent – as in the case of foreign languages – individuation and recognition are impaired. This theoretical representational asymmetry should be reflected in voice and speech sensitive regions of the brain, as shown here.

Taken together, these studies have provided some evidence for the plausibility of an LFE/ORE analogy, within the Valentine similarity framework. Note too that **Chapter 3**, may be interpreted as supporting the 'voice-space' coding scheme: while we found no evidence of a native-foreign asymmetry in similarity representations, we did find that the acoustically dissimilar English and Mandarin reversed speech voices were differentiated in bilateral superior temporal cortices. This is consistent under the terms of a voice-space whose dimensions may represent acoustical parameters which differentiate voices. Furthermore, this 'separation' of voices could represent an analytical stage of voice processing which may go on to underpin the LFE at later stages, somewhat resembling the analytical stages described in models of voice processing (Belin et al., 2004; Belin et al., 2011). This is speculative at present, but it may form a basis for future investigations of the differences between neural processes sub-serving voice discrimination and voice identification.

However, it should be noted that neither the work presented here, nor the aforementioned neurophysiological studies of the ORE provide specific information about the nature of the representational schema. As described above, the Valentine model comes in two flavours: the norm-based and the exemplar-based. Recent

neurophysiological evidence from studies of face (Leopold et al., 2001; Leopold et al., 2006) and voice-processing (Andics et al., 2010; Andics et al., 2013; Latinus et al., 2013) suggest mechanisms supporting a norm-based representation of identity, but the data presented here does not explicitly assess this. Indeed, this was neither the focus of this work, nor the work of Vizioli and colleagues vis-à-vis the ORE (Vizioli et al., 2010; Vizioli, 2012), but future studies should strive to probe this further.

5.2. Limitations

A few caveats must be discussed with regard to the work contained in this thesis. Firstly, the results presented in the fMRI experiment in **Chapter 3** suggest that the asymmetric behavioural similarity structure presented in **Chapter 2** was apparently not reflected in the brains of listeners. This behaviour-brain discrepancy may perhaps be accounted for by the fact that listeners were presented with voices under two very different sets of circumstances: in one instance, they were asked to explicitly rate the dissimilarity of pairs of voices, and given the opportunity to listen to each pair as many times as they felt necessary. In the other, they were presented with individual voices, at relatively short intervals, and were tasked with detecting an infrequently-occurring pure-tone. Given the known influence of task demands in auditory neuroimaging experiments (Bonte, Valente, & Formisano, 2009; Bonte et al., 2014; von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004) it is perhaps unsurprising that this discrepancy should emerge. As discussed in the chapter itself, it would, however, present a considerable methodological challenge to prepare an in-scanner task closely matched with the behavioural task. Another reason for the brain-behaviour discrepancy may simply be related to the magnitude of the LFE for reversed-speech. As can be seen in **Chapter 2**, although seemingly reliable, the

discrimination LFE for unintelligible speech is quite small, especially when contrasted with effect sizes obtained in previous identification studies with intelligible speech (e.g., Perrachione et al., 2011). The scanning protocol presented here, therefore, may have been insufficiently sensitive as to allow detection of the cerebral underpinnings of a small effect. A further potential explanation concerns the asymmetric amount of exposure the participants had to the voices prior to the fMRI and behavioural experiments. The same participants took part in both experiments, but, prior to fMRI scanning they received no exposure to the voice set. When they subsequently proceeded to complete the behavioural session on a separate testing day, however, they may have been able to bring to bear their prior exposure to the voices in completing the dissimilarity rating experiment. For example, in previous studies of the LFE, and in voice-recognition in general, participants tend to perform better in recognition of voices from learned speech items, as compared to performance in generalization to new speech items from learned talkers (e.g., Perrachione et al., 2007).

Secondly, **Chapter 4** presents evidence of a neural correlate of the LFE for intelligible speech, using English and Mandarin stimuli. However, the presented experiment recruited only one sample of listeners, from one language background (monolingual English speakers). The decision to test only one listener group was made after taking into account a number of practical considerations: Firstly, any Mandarin speakers recruited at the University of Glasgow would necessarily have a high level of functional knowledge of English, given that they would have been recruited as students of the University and would have to fulfil English language requirements for admission. (Note that this is not a problem typically faced by researchers examining the face ORE in only one country. In these experiments it may be sufficient to simply ensure that participants have spent

relatively little time in the country of testing, and have received limited exposure to faces in the other-race condition.) This did not present as considerable an issue in the reversed-speech experiments in earlier chapters, as these employed stimuli which were largely unintelligible to both sets of listeners. Secondly, there is some evidence that tone-language speakers possess better general voice recognition abilities than non-tone language speakers (Xie and Myers, 2015), which may have proven to be a confound in terms of cross-group comparisons. On this basis, a decision was made to study only one group of participants, but to ensure that both English and Mandarin voice stimuli were relatively similar with regards to inter-talker variability. While we cannot be absolutely sure that all sources of acoustical variability were accounted for with our summary measures, we have made strong attempts to control variability as far as possible. Note that the procedures we use follow very closely those used in Johnson et al. (2011) who tested a set of (infant) participants from only one language background. Nonetheless, future attempts should be made to study the neural basis of the LFE in participants from a range of different language backgrounds, as has been done with the classic behavioural effect.

Finally, while this thesis contains results which generally provide support for an extension of the Valentine model to the LFE, the exact basis of the effect is still somewhat unclear. For example, as discussed in the introduction, the hallmark native-language recognition advantage observed in the LFE could arise from voice memory traces stored within a "voice-space" elaborated around vocal acoustics, and shaped according to linguistic experience; or, the advantage may hinge on the use of robust, invariant "prototypical"/"abstract" representations of native language speech sounds (which are unavailable under foreign speech listening conditions) to compute inter-speaker

differences in vocal information for the benefit of voice recognition. Indeed, the effect may even involve the combined influence of prototypical representations of speech and voice information, and future studies should strive to tease these two factors apart.

Furthermore, on the foundations of the LFE, on the one hand, results from behaviour, including those contained in **Chapter 2** (Fleming et al., 2014), provide support for a phonological basis for the effect, which does not necessarily depend on intelligible speech. For example, Zarate et al. (2015) find that English pseudo-speech (or "nonsense-speech") supports better voice recognition than Mandarin speech in English-speaking listeners. Pseudo-speech has no meaning, but is constructed from valid phonological elements within a language. On the other hand, the neuroimaging results presented here might be interpreted as reflective of the role of intelligibility – apparent asymmetric cerebral representations of different native and foreign talkers were only found under clear speech conditions. However, as I have acknowledged above, the absence of this finding under unintelligible speech conditions may be reflective of the task which was used. Despite this, the behavioural findings from **Chapter 2** and the neuroimaging findings from **Chapter 4** both provide support for an extension of the Valentine model, in that access to native speech sounds appears to support enhanced individuation of native speakers. An intriguing next step in neuroimaging investigations of the LFE would be to extend the findings of Zarate and colleagues (2015) in varying the amount of linguistic information available in stimulus batteries. Stimulus content manipulations of this kind will prove vital in determining the point in the para-linguistic/linguistic hierarchy at which cerebral representations of native and foreign talkers begin to diverge.

5.3. Future Directions

Since it was first reported nearly 30 years ago, the LFE has continued to stimulate the interests of psycholinguists and cognitive experimental psychologists alike, giving rise to a literature which continues to yield new contributions. I will now propose some lines of thought for future investigation, particularly with regard to the scope for neuroimaging investigation.

In the context of recent advances in our understanding of the neural basis of voice recognition, it would be interesting to firstly build on the decoding results presented in **Chapter 4**. For example, as I discussed in **Chapter 1**, and as we have seen in the results of **Chapter 2**, the magnitude of a listener's particular LFE may be affected by the knowledge they have of the foreign language which is under study (Creel and Bregman, 2014), or by the relationship between their native language and the foreign language (Zarate et al., 2015). This represents an obvious opportunity to extend the present results – I would speculate that the distinctiveness of cerebral representations of different speakers would be affected by both of these issues. A study of this nature would provide an even stronger test of the adapted form of the Valentine face-encoding framework – for example, cerebral representations of different foreign speakers may become more differentiable or "native-like", as a function of inter-language phonological overlap, or knowledge of a foreign language.

A related line of investigation would involve the extent to which representations of different speakers which are purportedly content-invariant (i.e., regardless of what the speaker is saying) would vary depending on whether stimuli are cross-classified *across*-

languages as opposed to within-languages, but over different speech tokens. Formisano and colleagues previously showed (Formisano et al., 2008) that distributed (but right-lateralized) auditory cortical patterns could be used by a decoder to not only discriminate speaker identity, but also to generalize performance to patterns elicited by trained speakers uttering untrained vowels. It would be interesting to determine whether this finding could be extended with more complex stimuli, such as words or sentences, and, crucially, to explore whether training a decoder to identify a speaker of one language would enable it to generalize to brain patterns elicited by hearing the same speaker talking in a different language. Note that this would be similar to a study conducted by Correia et al., (Correia et al., 2014), where Dutch-English bilinguals were scanned whilst listening to word lists presented in both languages. There, local pattern content in the left anterior temporal lobe enabled cross-classification of individual words, whether classifier training was performed on responses to Dutch words and generalized to English, or vice-versa. For example, training on trial responses to the Dutch word “paard” could enable a classification algorithm to correctly classify an untrained response to the same word translated into English (“horse”). In the case of decoding bilingual speaker identities, comparing within-language discriminative maps to cross-language maps might yield some interesting differences, as in the work of Correia and colleagues (2014). Furthermore, related to an earlier point, one might observe a high degree of overlap in within and across-language speaker generalization discriminative maps when participating *listeners* themselves are bilingual, in contrast to monolingual listeners. A related line of investigation would be to combine the approaches specified above, in order to examine decoding generalization performance in a “parametric” fashion as modelled on the behavioural study of Zarate and colleagues (Zarate et al., 2015). To illustrate, consider the example of a monolingual native English speaker who is scanned whilst listening to

speech stimuli recorded by 5 male talkers across a range of speech conditions; in other words, exactly as executed in Zarate et al. (2015). Based on those behavioural findings, we might expect that training of a decoder on patterns elicited by English speech in the brain of an English-speaking monolingual would lead to better decoding generalization to "unseen" patterns elicited by the same speakers reading German speech, as compared with generalization to Mandarin speech. Such a hypothetical finding could be accounted for by the heightened typological similarity shared between English and German, as compared to that shared between English and Mandarin (Schiller and Koester, 1996; Schiller et al., 1997; Zarate et al., 2015).

As a final thought, while fMRI has been used in the present experiments to characterize the spatial localization of differences between brain representations of native and foreign speakers, it is also critical to understand the neural timescale of the LFE. For example, results from MEG suggest that speech and voice information are integrated pre-attentively in bilateral auditory cortex, approximately 120-140ms following stimulus presentation (Knösche, Lattner, Maess, Schauer, & Friederici, 2002). More recently, an EEG study by Zhang and colleagues (Zhang et al., 2015) found an integration of talker and lexical tone properties in a later time window (500-800ms) over central and frontal electrodes. While recent studies have used EEG to study the neural timecourses and oscillatory fingerprints associated with speaker learning and recognition (e.g. Bonte et al., 2009; Zäske, Volberg, Kovács, & Schweinberger, 2014), they typically use stimuli derived from the speech sounds of participants' native languages, as with much of the fMRI literature on voice recognition. It is possible, therefore, that the profile of electrophysiological responses thought to be involved in voice recognition shall vary depending on whether a listener is acquainted with the language spoken by a voice.

5.4 General Conclusion

In this thesis, I have attempted to examine the differences between representations of native and foreign speakers, using behavioural and brain imaging methods. Adding to the growing body of literature on the "Language-Familiarity" Effect (LFE) in voice recognition, I have presented evidence that suggests that these representations do indeed differ: behaviourally, listeners better differentiate speakers when those speakers utter time-reversed native-language speech. In the brain, different voices which speak a listener's native language may be more efficiently represented than different foreign voices. I have argued for a parallel between the LFE and the "Other-Race" Effect (ORE) in face recognition, within Valentine's (1991) similarity-based person encoding model; perhaps more importantly, I have shown that, to approach a complete understanding of human voice recognition, cognitive scientists must continue to take account of the now well-known interactions between the speaker and speech information carried by the human voice.

6. References

Abrams, D. a, Ryali, S., Chen, T., Balaban, E., Levitin, D. J., & Menon, V. (2013). Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(7), 1703–14. http://doi.org/10.1093/cercor/bhs165

Ahrens, M.-M., Awwad Shiekh Hasan, B., Giordano, B. L., & Belin, P. (2014). Gender differences in the temporal voice areas. *Frontiers in Neuroscience*, *8*(July), 228. http://doi.org/10.3389/fnins.2014.00228

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage*, *13*(5), 903–919. http://doi.org/10.1006/nimg.2001.0746

Andics, A., Gácsi, M., Faragó, T., Kis, A., & Miklósi, A. (2014). Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology : CB*, *24*(5), 574–8. http://doi.org/10.1016/j.cub.2014.01.058

Andics, A., McQueen, J. M., & Petersson, K. M. (2013). Mean-based neural coding of voices. *NeuroImage*, *79*, 351–60. http://doi.org/10.1016/j.neuroimage.2013.05.002

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage*, *52*(4), 1528–40. http://doi.org/10.1016/j.neuroimage.2010.05.048

Arnott, S. R., Heywood, C. a., Kentridge, R. W., & Goodale, M. a. (2008). Voice recognition and the posterior cingulate: An fMRI study of prosopagnosia. *Journal of Neuropsychology*, *2*(1), 269–286. http://doi.org/10.1348/174866407X246131

Assal, G., Aubert, C., & Buttet, J. (1981). Asymétrie cérébrale reconnaissance de la voix. *Revue Neurologique*, *137*, 255–268.

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and Nurture in Own- Race Face Processing. *Psychological Science*, *17*(2), 159–163.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–20. http://doi.org/10.1007/s00426-008-0185-z

Beauchemin, M., González-Frankenberger, B., Tremblay, J., Vannasing, P., Martínez-Montes, E., Belin, P., … Lassonde, M. (2011). Mother and stranger: an electrophysiological study of voice processing in newborns. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(8), 1705–11. http://doi.org/10.1093/cercor/bhq242

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology (London, England : 1953)*, *102*(4), 711–25. http://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–35. 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. *Neuroreport*, *14*, 2105–2109. http://doi.org/10.1097/00001756-200311140-00019

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Research. Cognitive Brain Research*, *13*(1), 17–26.

Belin, P., Zatorre, R., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–12. http://doi.org/10.1038/35002078

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford–Kowal–Bench) sentence lists for partially-hearing children. *British Journal of Audiology, 13*, 108–112.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*. http://doi.org/10.1162/jocn.1996.8.6.551

Bethmann, A., & Brechmann, A. (2014). On the definition and interpretation of voice selective activation in the temporal cortex. *Frontiers in Human Neuroscience*, *8*(July), 499. http://doi.org/10.3389/fnhum.2014.00499

Bethmann, A., Scheich, H., & Brechmann, A. (2012). The Temporal Lobes Differentiate between the Voices of Famous and Unknown People: An Event-Related fMRI Study on Speaker Recognition. *PLoS ONE*, *7*(10), e47626. http://doi.org/10.1371/journal.pone.0047626

Binder, J. R., Frost, J. a, Hammeke, T. a, Bellgowan, P. S., Springer, J. a, Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex (New York, N.Y. : 1991)*, *10*(5), 512–28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10847601

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(36), 12906–15. http://doi.org/10.1523/JNEUROSCI.2091-11.2011

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., … Murphy, D. G. M. (2011). Early specialization for voice and emotion processing in the infant brain. *Current Biology : CB*, *21*(14), 1220–4. http://doi.org/10.1016/j.cub.2011.06.009

Boersma, P & Weenink, D. (2015) Praat, a system for doing phonetics by computer (http://www.praat.org).

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(13), 4548–57. http://doi.org/10.1523/JNEUROSCI.4339-13.2014

Bonte, M., Valente, G., & Formisano, E. (2009). Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations. *Journal of Neuroscience*, *29*(6), 1699–1706. http://doi.org/10.1523/JNEUROSCI.3694-08.2009

Brainard, D. (1997). The Psychophysics Toolbox, *Spatial Vision 10*, 433-436.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95. http://doi.org/10.1016/j.cognition.2013.09.010

Bricker, P. D., & Pruzansky, S. (1966). Effects of Stimulus Content and Duration on Talker Identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441–1449.

Brosch, T., Bar-David, E., & Phelps, E. a. (2013). Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological Science*, *24*(2), 160–6. http://doi.org/10.1177/0956797612451465

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology (London, England : 1953)*, *77 (Pt 3)*, 305–327. http://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., … Belin, P. (2010). Vocal Attractiveness Increases by Averaging. *Current Biology*, *20*, 116–120. http://doi.org/10.1016/j.cub.2009.11.034

Caldara, R., & Abdi, H. (2006). Simulating the “other-race” effect with autoassociative neural networks: further evidence in favor of the face-space model. *Perception*, *35*(5), 659–670. http://doi.org/10.1068/p5360

Capilla, A., Belin, P., & Gross, J. (2013). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(6), 1388–95. http://doi.org/10.1093/cercor/bhs119

Carlin, J. D., Nili, H., Calder, A. J., Rowe, J. B., & Kriegeskorte, N. (2011). A Head View-Invariant Representation of Gaze Direction in Anterior Superior Temporal Sulcus. *Current Biology*, *21*(21), 1817–1821.

Carlin, J. D., Rowe, J. B., Kriegeskorte, N., Thompson, R., & Calder, A. J. (2012). Direction-sensitive codes for observed head turns in human superior temporal sulcus. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(4), 735–44. http://doi.org/10.1093/cercor/bhr061

Chandrasekaran, B., Chan, A. H. D., & Wong, P. C. M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–700. http://doi.org/10.1162/jocn.2011.21631

Chang, C. & Lin, C. (2011). LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3), 27:1-27:27.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–32. http://doi.org/10.1038/nn.2641

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

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(1), 332–8. http://doi.org/10.1523/JNEUROSCI.1302-13.2014

Coutanche, M. N. & Thompson-Schill, S. L. (2012). The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. *Neuroimage, 61,* 1113-1119.

Creel, S. C., & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing ,*Language and Linguistics Compass, 5*, 190–204.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(8), 3423–3431. http://doi.org/23/8/3423 [pii]

Davis, T., LaRocque, K. F., Mumford, J. a, Norman, K. a, Wagner, A. D., & Poldrack, R. a. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, *97*, 271–83. http://doi.org/10.1016/j.neuroimage.2014.04.037

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–88. http://doi.org/10.1162/jocn.2008.20081

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body, *Science, 293*, 2470–2473.

Duanmu, S. (2000). *The Phonology of Standard Chinese* (Oxford Press, New York).

Eady, S. J. (1982). Differences in the F0 Patterns of Speech: Tone Language Versus Stress Language. *Language and Speech*, *25*, 29–42. http://doi.org/10.1177/002383098202500103

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of Emotional Information in Voice-Sensitive Cortices. *Current Biology*, *19*(12), 1028–1033. http://doi.org/10.1016/j.cub.2009.04.054

Evans, S., Kyong, J. S., Rosen, S., Golestani, N., Warren, J. E., McGettigan, C., … Scott, S. K. (2014). The pathways for intelligible speech: multivariate and univariate perspectives. *Cerebral Cortex (New York, N.Y. : 1991)*, *24*(9), 2350–61. http://doi.org/10.1093/cercor/bht083

Feingold, C. A. (1914). *The influence of environment on identification of persons and things. Journal of Criminal Law and Police Science,5,* 39–51.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, *111*(38), 13795–13798. http://doi.org/10.1073/pnas.1401383111

Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, *3*(3), 333–343. http://doi.org/10.1111/1467-7687.00128

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science (New York, N.Y.)*, *322*(5903), 970–3. http://doi.org/10.1126/science.1164318

Gabrieli, J. D. E. (2009). Dyslexia: a new synergy between education and cognitive neuroscience. *Science (New York, N.Y.)*, *325*(5938), 280–3. http://doi.org/10.1126/science.1171999

Gainotti, G. (2013a). Is the right anterior temporal variant of prosopagnosia a form of "associative prosopagnosia" or a form of "multimodal person recognition disorder"? *Neuropsychology Review*, *23*(2), 99–110. http://doi.org/10.1007/s11065-013-9232-7

Gainotti, G. (2013b). Laterality effects in normal subjects' recognition of familiar faces, voices and names. Perceptual and representational components. *Neuropsychologia*, *51*(7), 1151–60. http://doi.org/10.1016/j.neuropsychologia.2013.03.009

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., … Duchaine, B. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, *47*(1), 123–31. http://doi.org/10.1016/j.neuropsychologia.2008.08.003

Giordano, B. L., Pernet, C, Charest I., Belizaire, G., Zatorre, R. J. & Belin, P. (2014). Automatic domain-general processing of sound source identity in the left posterior middle frontal gyrus. *Cortex, 58,* 170-185.

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(9), 2025–37. http://doi.org/10.1093/cercor/bhs162

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*, 448–458. http://doi.org/10.3758/BF03199567

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices, *Bulletin of the Psychonomic Society, 17*(5), 217–220.

Gómez, D., Berent, I., Benavides-Varela, S., Bion, R. A. H., Cattarossi, L., Nespor, M., & Mehler, J. (2014). Language universals at birth. *Proceedings of the National Academy of Sciences, 111*(16), 5837-5841.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, *8*(2), 145–6. http://doi.org/10.1038/nn1392

Griffiths, T. D. & Warren, J. D. (2002). The planum temporale as a computational hub*, Trends In Cognitive Sciences, 25*(7), 348–353.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2005.11.006

Grossmann, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron*, *65*(6), 852–8. http://doi.org/10.1016/j.neuron.2010.03.001

Hailstone, J. C., Ridgway, G. R., Bartlett, J. W., Goll, J. C., Buckley, A. H., Crutch, S. J., & Warren, J. D. (2011). Voice processing in dementia: a neuropsychological and neuroanatomical analysis. *Brain : A Journal of Neurology*, *134*(Pt 9), 2535–47. http://doi.org/10.1093/brain/awr205

Halai, A., Parkes, L. M., & Welbourne, S. (2015). Dual-echo fMRI can detect activations in inferior temporal lobe during intelligible speech comprehension. *NeuroImage*, *122*, 214–221. http://doi.org/10.1016/j.neuroimage.2015.05.067

Hanson, H. M. (1997). Glottal characteristics of female speakers: acoustic correlates. *The Journal of the Acoustical Society of America*, *101*(1), 466–481. http://doi.org/10.1121/1.417991

Henson, R. (2006). Efficient Experimental Design for fMRI. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols and W. Penny (Eds), *Statistical Parametric Mapping: The analysis of functional brain images* (pp 193-210). Elsevier, London.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Neuroscience*, *8*(May), 393–402.

Hsieh, L., Gandour, J., Wong, D., & Hutchins, G. D. (2001). Functional Heterogeneity of Inferior Frontal Gyrus Is Shaped by Linguistic Experience. *Brain and Language*, *76*(3), 227–252. http://doi.org/10.1006/brln.2000.2382

Imaizumi, S., Mori, C. A. K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., … Nakamura, K. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport, 8,* 2809–2812.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–11. http://doi.org/10.1111/j.1467-7687.2011.01052.x

Johnson, K. (2008). Speaker Normalization in Speech Perception. In D. B. Pisoni and R. E. Remez (Eds), *The handbook of speech perception* (pp 363-389). Blackwell, Oxford.

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*, 4302–4311. http://doi.org/10.1098/Rstb.2006.1934

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050–60. http://doi.org/10.1121/1.4730893

Kelly, D. J., Liu, S., Ge, L., Quinn, P. C., Slater, A. M., Lee, K., … Pascalis, O. (2007). Cross-Race Preferences for Same-Race Faces Extend Beyond the African Versus Caucasian Contrast in 3-Month-Old Infants. *Infancy*, *11*(1), 87–95. http://doi.org/10.1080/15250000709336871

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084–9. http://doi.org/10.1111/j.1467-9280.2007.02029.x

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., … Pascalis, O. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*(6), 31–37. http://doi.org/10.1111/j.1467-7687.2005.0434a.x

Kisilevsky, B. S., Hains, S. M. J., Brown, C. a, Lee, C. T., Cowperthwaite, B., Stutzman, S. S., … Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior & Development*, *32*(1), 59–71. http://doi.org/10.1016/j.infbeh.2008.10.002

Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H. H., … Wang, Z. (2003). Effects of experience on fetal voice recognition, *Psychological Science, 14*(3), 220–224.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *NeuroImage*, *13*(4), 646–53. http://doi.org/10.1006/nimg.2000.0738

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 2698–2704. http://doi.org/10.1098/rspb.2012.0311

Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early Parallel Processing of Auditory Word and Voice Information. *NeuroImage*, *17*(3), 1493–1503. http://doi.org/10.1006/nimg.2002.1262

Koster, O., & Schillert, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, *4*(1), 18–28.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4. http://doi.org/10.3389/neuro.06.004.2008

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

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*(11), 831–43. http://doi.org/10.1038/nrn1533

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science, 255*(5044), 606-608.

Kuhl, P., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annual Review of Neuroscience*, *31*, 511–34. http://doi.org/10.1146/annurev.neuro.30.051606.094321

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*(July), 175. http://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., Crabbe, F., & Belin, P. (2011a). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(12), 2820–8. http://doi.org/10.1093/cercor/bhr077

Latinus, M., Crabbe, F., & Belin, P. (2011b). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(12), 2820–8. http://doi.org/10.1093/cercor/bhr077

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013a). Norm-based coding of voice identity in human auditory cortex. *Current Biology : CB*, *23*(12), 1075–80. http://doi.org/10.1016/j.cub.2013.04.055

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013b). Norm-based coding of voice identity in human auditory cortex. *Current Biology : CB*, *23*(12), 1075–80. http://doi.org/10.1016/j.cub.2013.04.055

Leopold, D. a, Bondar, I. V, & Giese, M. a. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–575. http://doi.org/10.1038/nature04951

Leopold, D. a, O'Toole, a J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94. http://doi.org/10.1038/82947

Levi, S. V., & Schwarz, R. G. (2013). The development of language-specific and language-independent talker processing. *Journal of Speech, Language, and Hearing Research.*, *56*, 913–925.

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: electrophysiological evidence. *Neuroreport*, *12*, 2653–2657. http://doi.org/10.1097/00001756-200108280-00013

Levy, D. a., Granot, R., & Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology*, *40*(2), 291–305. http://doi.org/10.1111/1469-8986.00031

Mang, E. (2001). A Cross-language Comparison of Preschool Children's Vocal Fundamental Frequency in Speech and Song Production. *Research Studies in Music Education*, *16*(1), 4–14. http://doi.org/10.1177/1321103X010160010201

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: parallels with face recognition. *Journal of Experimental Child Psychology*, *27*, 153–165. http://doi.org/10.1016/0022-0965(79)90067-5

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*, 676–684. http://doi.org/10.1037/0278-7393.17.1.152

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional Representation of Personal Quality of Vowels and its Acoustical Correlates. *IEEE Transactions on Audio and Electroacoustics, 21*, 428–436.

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

McGehee, F. (1937). The Reliability of the Identification of the Human Voice. *The Journal of General Psychology*, *17*(2), 249–271. http://doi.org/10.1080/00221309.1937.9917999

Mcgettigan, C., Evans, S., Rosen, S., Agnew, Z. K., Shah, P., & Scott, S. K. (2012). An Application of Univariate and Multivariate Approaches in fMRI to Quantifying the Hemispheric Lateralization of Acoustic and Linguistic Processes, 636–652.

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what's wrong, what's right and what's left? *Trends in Cognitive Sciences*, *16*(5), 269–76. http://doi.org/10.1016/j.tics.2012.04.006

Mehler, J., Bertoncini, J., Barriere, M., & Jassik-gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception, 7*, 491–497.

Meissner, C. a., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. http://doi.org/10.1037//1076-8971.7.1.3

Misaki, M., Kim, Y., Bandettini, P. a, & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. http://doi.org/10.1016/j.neuroimage.2010.05.051

Moon, C., Cooper, R., & Fifer, W. (1993). Two-Day-Olds Prefer Their Native language. *Infant Behavior & Development*, *16*(4), 495–500.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. http://doi.org/10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–78.

Mumford, J. A., Turner, B. O. Ashby, F. G. & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage, 59*(3), 2636-2643.

Nakamura, K., Kawashima, R., Sugiura, M., & Kato, T. (2001). Neural substrates for recognition of familiar voices : a PET study. *Neuropsychologia, 39*, 1047–1054.

Narain, C. (2003). Defining a Left-lateralized Response Specific to Intelligible Speech Using fMRI. *Cerebral Cortex*, *13*(12), 1362–1368. http://doi.org/10.1093/cercor/bhg083

Natu, V., Raboy, D., & O'Toole, A. J. (2011). Neural correlates of own- and other-race face perception: spatial and temporal response differences. *NeuroImage*, *54*(3), 2547–55. http://doi.org/10.1016/j.neuroimage.2010.10.006

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language Discrimination by Newborns : Toward an Understanding of the Role of Rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, *43*(1), 1–19. http://doi.org/10.1006/jmla.2000.2698

Neuhoff, J. G., Schott, S. a, Kropf, A. J., & Neuhoff, E. M. (2014). Familiarity, expertise, and change detection: Change deafness is worse in your native language. *Perception*, *43*(2), 219–222. http://doi.org/10.1068/p7665

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–76.

Obleser, J., Meyer, L., & Friederici, A. D. (2011). Dynamic assignment of neural resources in auditory comprehension of complex sentences. *NeuroImage*, *56*(4), 2310–2320. http://doi.org/10.1016/j.neuroimage.2011.03.035

Ockleford, E. M., Vince, M. a, Layton, C., & Reader, M. R. (1988). Responses of neonates to parents' and others' voices. *Early Human Development*, *18*(1), 27–36.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., … Hickok, G. (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. *Cerebral Cortex*, *20*(10), 2486–2495. http://doi.org/10.1093/cercor/bhp318

Oldfield, R. C. (1971). The Assessment and Analysis of Handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776. http://doi.org/10.1016/S0896-6273(02)01060-7

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(6), 1378–87. http://doi.org/10.1093/cercor/bhs118

Perani, D., Dehaene, S., Grassi, F., Cohen, L., Cappa, S., Dupoux, E., … Mehler, J. (1996). Brain processing of native and foreign languages. *Neuroreport*, *7*(15-17), 2439–2444.

Perea, M., Jiménez, M., Suárez-coalla, P., Fernández, N., Viña, C., & Cuetos, F. (2014). Ability for Voice Recognition Is a Marker for Dyslexia in Children. *Experimental Psychology, 61*(6), 480–487.

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

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science (New York, N.Y.)*, *333*(6042), 595. http://doi.org/10.1126/science.1207327

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1950–60. http://doi.org/10.1037/a0015869

Perrachione, T. K., Stepp, C. E., Hillman, R. E., & Wong, P. C. M. (2014). Talker Identification Across Source Mechanisms: Experiments With Laryngeal and Electrolarynx Speech. *Journal of Speech, Language and Hearing Research*, *57*, 1651–1665.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–910. http://doi.org/10.1016/j.neuropsychologia.2006.11.015

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2011). Voice cells in the primate temporal lobe. *Current Biology*, *21*, 1408–1415. http://doi.org/10.1016/j.cub.2011.07.028

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, *11*(3), 367–74. http://doi.org/10.1038/nn2043

Philippon, A. C., Cherryman, J., Bull, R. A. Y., & Vrij, A. (2007). Earwitness Identification Performance : The Effect of Language , Target , Deliberate Strategies and Indirect Measures. *Applied Cognitive Psychology, 21,* 539-550.

Pierce, L. J., Klein, D., Chen, J.-K., Delcenserie, A., & Genesee, F. (2014). Mapping the unconscious maintenance of a lost first language. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48), 17314–9. http://doi.org/10.1073/pnas.1409411111

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*(1-2), 109–125.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as “asymmetric sampling in time.” *Speech Communication*, *41*(1), 245–255. http://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*, 1071–1086. http://doi.org/10.1098/rstb.2007.2160

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*, 62–88. http://doi.org/10.1111/j.1749-6632.2010.05444.x

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–47. http://doi.org/10.1016/j.neuroimage.2012.04.062

Puschmann, S., Uppenkamp, S., Kollmeier, B., & Thiel, C. M. (2010). Dichotic pitch activates pitch processing centre in Heschl's gyrus. *NeuroImage*, *49*(2), 1641–1649. http://doi.org/10.1016/j.neuroimage.2009.09.045

Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *The Journal of the Acoustical Society of America*, *122*(6), 3688–96. http://doi.org/10.1121/1.2799903

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology. Human Perception and Performance*, *23*(3), 651–66.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977–2987. http://doi.org/10.1016/j.visres.2006.03.002

Rossion, B., & Jacques, C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. *NeuroImage*. http://doi.org/10.1016/j.neuroimage.2007.10.011

Sangrigoli, S., Pallier, C., Argenti, a-M., Ventureyra, V. a G., & de Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science : A Journal of the American Psychological Society / APS*, *16*(6), 440–444. http://doi.org/10.1111/j.0956-7976.2005.01554.x

Schall, S., Kiebel, S., Maess, B. & von Kriegstein, K. (2015). Voice Identity Recognition: Functional Division of the Right STS and Its Behavioral Relevance. *Journal of Cognitive Neuroscience, 27*(2), 280-291.

Schiller, N., & Koster, O. (1996). Evaluation of a foreign speaker in forensic phonetics : a report. *Forensic Linguistics*, *3*(1), 176–185.

Schiller, N. O., Koster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics*, *4*(1).

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G. & Zaeske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 15-25.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain : A Journal of Neurology*, *123 Pt 12*, 2400–2406. http://doi.org/10.1093/brain/123.12.2400

Scott, S. K., & McGettigan, C. (2013). Do temporal processes underlie left hemisphere dominance in speech perception? *Brain and Language*, *127*(1), 36–45. http://doi.org/10.1016/j.bandl.2013.07.006

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447–1469. http://doi.org/10.1037//0096-1523.28.6.1447

Shu, H. & Anderson, R. C. (1999). Learning to read Chinese: The development of metalinguistic awareness. In *Reading Chinese Script: A Cognitive Analysis*, eds Wang, J., Inhoff, A. & Chen, H.

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J. (2003). D istinct prefrontal cortex activity associated with item memory and source memory for visual shapes, *Cognitive Brain Research, 17*, 75–82.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Current Biology*, *19*(6), 498–502. http://doi.org/10.1016/j.cub.2009.01.066

Stevens, A. a. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, *18*(2), 162–171. http://doi.org/10.1016/j.cogbrainres.2003.10.008

Sullivan, K. P. H., & Kügler, F. (2001). Was the knowledge of the second language or the age difference the determining factor ? *Forensic Linguistics*, *8*(2).

Sullivan, K. P. H., & Schlichting, F. (2000). Speaker discrimination in a foreign language : first language environment , second language learners. *Forensic Linguistics*, *7*(1).

Sundara, M. & Kuhl, P. K. (2006). Distinguishing voices: Are infants' abilities affected by age or language experience?, *Journal of the Acoustical Society of America, 120*, 3135.

Thompson, C. P. (1987). A Language Effect in Voice Identification. *Applied Cognitive Psychology, 1*, 121–131.

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, *33*, 210–216. http://doi.org/10.1016/j.evolhumbehav.2011.09.004

Turner, B. O., Mumford, J. A., Poldrack, R. A., Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *Neuroimage, 62*(3). 1429-1438.

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

Valentine, T., & Endo, M. (1992). Towards an Exemplar Model of Face Processing: The Effects of Race and Distinctiveness. *The Quarterly Journal of Experimental Psychology Section A*, *44*(4), 671–703. http://doi.org/10.1080/14640749208401305

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics*, *13*, 19–38.

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185–95.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. Phonagnosia: a dissociation between familiar and unfamiliar voices., 24 Cortex; a journal devoted to the study of the nervous system and behavior 195–209 (1988). http://doi.org/10.1016/S0010-9452(88)80029-7

Van Lancker, D. R., & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain and Language*, *32*(2), 265–77.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 665–74. http://doi.org/10.1080/01688638908400923

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houdé, O., … Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, *30*(4), 1414–1432. http://doi.org/10.1016/j.neuroimage.2005.11.002

Vizioli, L. (2012). *Clarifying the neurophysiological basis of the other-race effect*. University of Glasgow, United Kingdom.

Vizioli, L., Rousselet, G. a, & Caldara, R. (2010). Neural repetition suppression to identity is abolished by other-race faces. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(46), 20081–6. http://doi.org/10.1073/pnas.1005751107

Von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research. Cognitive Brain Research*, *17*(1), 48–55.

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

Von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), e326. http://doi.org/10.1371/journal.pbio.0040326

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. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(2), 629–38. http://doi.org/10.1523/JNEUROSCI.2742-09.2010

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790. http://doi.org/10.1016/j.specom.2012.01.006

Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *The Journal of the Acoustical Society of America*, *137*(1), 419–32. http://doi.org/10.1121/1.4904699

Yeong, S. H. M., & Rickard Liow, S. J. (2012). Development of phonological awareness in English-Mandarin bilinguals: a comparison of English-L1 and Mandarin-L1 kindergarten children. *Journal of Experimental Child Psychology*, *112*(2), 111–26. http://doi.org/10.1016/j.jecp.2011.12.006

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–71. http://doi.org/10.1016/j.tics.2013.04.004

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Nature Scientific Reports*, *5*, 11475. http://doi.org/10.1038/srep11475

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research*, *268*(1-2), 38–45. http://doi.org/10.1016/j.heares.2010.04.011

Zäske, R., Volberg, G., Kovács, G., & Schweinberger, S. R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(33), 10821–31. http://doi.org/10.1523/JNEUROSCI.0581-14.2014

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex : music and speech, *Trends in Congitive Sciences, 6*(1), 37–46.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: moving beyond the dichotomies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1087–104. http://doi.org/10.1098/rstb.2007.2161

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., … Wang, W. S.-Y. (2015). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*. http://doi.org/10.1016/j.neuroimage.2015.08.064

Zhang, J., Kriegeskorte, N., Carlin, J. D., & Rowe, J. B. (2013). Choosing the rules: distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(29), 11852–62. http://doi.org/10.1523/JNEUROSCI.5193-12.2013