Behavioural and neural insights into the recognition and motivational salience of familiar voice identities.

Elise Kanber

A thesis submitted in partial fulfilment of

the requirements for the degree of

Doctor of Philosophy

of

University College London.

Division of Psychology and Language Sciences

Faculty of Brain Sciences

UCL

Declaration:

I, Elise Kanber, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: **Date:** 22/03/22

Abstract

The majority of voices encountered in everyday life belong to people we know, such as close friends, relatives, or romantic partners. However, research to date has overlooked this type of familiarity when investigating voice identity perception. This thesis aimed to address this gap in the literature, through a detailed investigation of voice perception across different types of familiarity: personally familiar voices, famous voices, and lab-trained voices. The experimental chapters of the thesis cover two broad research topics: 1) Measuring the recognition and representation of personally familiar voice identities in comparison with lab-trained identities, and 2) Investigating motivation and reward in relation to hearing personally valued voices compared with unfamiliar voice identities. In the first of these, an exploration of the extent of human voice recognition capabilities was undertaken using personally familiar voices of romantic partners. The perceptual benefits of personal familiarity for voice and speech perception were examined, as well as an investigation into how voice identity representations are formed through exposure to new voice identities. Evidence for highly robust voice representations for personally familiar voices was found in the face of perceptual challenges, which greatly exceeded those found for lab-trained voices of varying levels of familiarity. Conclusions are drawn about the relevance of the amount and type of exposure on speaker recognition, the expertise we have with certain voices, and the framing of familiarity as a continuum rather than a binary categorisation. The second topic utilised voices of famous singers and their "super-fans" as listeners to probe reward and motivational responses to hearing these valued voices, using behavioural and neuroimaging experiments. Listeners were found to work harder, as evidenced by faster reaction times, to hear their musical idol compared to less valued voices in an effort-based decision-making task, and the neural correlates of these effects are reported and examined.

Impact Statement

The research in the first half of this thesis provides novel insights into the perception of identity from highly personally familiar voices. Previous work states that face recognition is superior to voice recognition, and that voice recognition is largely error-prone. However, research into vocal identity perception – largely due to practical necessity – often uses lab-trained or famous voices. Therefore, our understanding of how good voice recognition could be at its best was previously limited. The studies in this thesis used personally familiar voices and found that participants are able to recognise these voices to a high degree of accuracy from both non-verbal and verbal vocalisations, and from acoustically-manipulated sounds. This provides an important advancement in our understanding of human identity perception, which has consequences for theoretical models of vocal identity processing, as well as affecting how familiarity is approached and understood in future investigations. The findings in the first half of this thesis are relevant to the study of voice and speech perception, both in the lab and in applied settings (e.g. eye- and earwitness testimony). They may also be of interest to domains such as social and evolutionary psychology, such as in understanding the recognition of kin, or in studying the evolution of vocal communication.

The second half of this thesis also provides valuable insights into a previously underexplored facet of human voice processing. That is, investigating the social and emotional significance of particular vocal identities. The human voice delivers a wealth of information upon hearing it, including clues about a person's identity, health, current mood, and personality to name a few. Besides this, often the voices of people familiar to us are attached to individuals that we value. Anecdotally, hearing the voices of people we care about can be a pleasant and soothing experience. Yet the implications that differences in the subjective importance placed on particular voices may have for our understanding of voice processing has remained unexplored in the extant voice perception literature. The studies in this thesis used tasks designed for investigating the value or motivational nature of rewards to explore how voices could influence behaviour or brain activity. These studies highlighted that not all voices are equally appraised, in that some voices are more rewarding than others. These results provide, for the first time, some evidence for the importance of studying the voice as a social stimulus, that can be associated with personal and emotional value. It sheds light on the need for future studies to consider the importance of the social and emotional significance of voices, and the impact that this may have on voice perception and learning.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Carolyn McGettigan, my primary supervisor, without whom getting to this point would not have been possible. Your guidance, mentorship, and commitment to my professional development has been invaluable. Thank you for being so generous with your time, no matter how busy you may have been, and for always meeting me where I was at. I consider myself extremely lucky to have been supervised by somebody that is a great inspiration to me as an academic and a person. Thank you, Carolyn.

Thanks also go to my secondary supervisor, Jon Roiser for many interesting and useful discussions, particularly regarding experimental design. Your time, expertise and patience in explaining concepts is highly appreciated. And to Davide for all of the help and advice at the MRI scanner - we made a great team!

Thank you to members of the VoCoLab past and present (in alphabetical order): Abbie, Clare, Michel, Nadine, Sarah, and Ziyun. I am extremely grateful to have had the fortune of working alongside people that are equal parts knowledgeable, approachable, and generally lovely. I have learnt a huge amount from each of you. Thank you to Bryony for being a PhD buddy from the start, it was a pleasure to do our PhD journeys alongside one another! To Anqi, Gwijde, Shego, Shiran (and Maddie!), Julie, Yue, Max, Anna and Ana, Clara (and Clara!), Rachel, Jonas, Han, Hannah, Begonia, Magda, and Gwen, for making Chandler house a great place to be. Thank you also to all of my participants for giving up their time to take part in what were often long and repetitive experiments, and for their honest feedback (e.g. "*This study is psychotic.*").

To John, Luisa, and Simon for keeping me sane throughout this process – the lessons you taught me will persist long after the PhD, and for that I am grateful. To Aliya (a.k.a. Wendy), Zandria, Imaan, Katherine, Lara, YaoTong, Kyra and Minsung, thank you for celebrating the small wins with me, and sharing in the challenges. I appreciate your patience, support, and your memes.

Lastly, to my family, for showing unwavering belief in me, even if it isn't completely clear what it is I actually do. To my parents, ma and pa, thank you for your sacrifices that made this endeavour possible, for encouraging me to not give up on myself, and for raising me to believe that anything is achievable – *sizi seviyorum, iyi ki varsınız.* To my brothers, Jay and Eren, for their tolerance and for being able to make me laugh during stressful times, and to my grandparents, nan and fuff – I'm not the prime minister but I will be a Dr!

This thesis is dedicated to the memory of my grandmother,
Elif Kanber
(1929 – 2019)

Contents

List of Figures:

List of Tables:

1 Introduction

Human voices are everywhere. Many of the voices encountered in everyday life belong to people familiar to us, such as close friends, romantic partners, or relatives. However, research to date into vocal identity processing has most commonly probed lab-trained familiarity or familiarity with famous voices, and has largely overlooked the voices that are most consistently interacted with, and of greatest personal importance: highly personally familiar voices. In addition to familiarity, particular individuals in one's life may hold value or personal relevance. Yet the social and emotional significance of human voices has also rarely been examined. This thesis is broadly separated into two research topics, with a view to begin to examine some of these underexplored/under-represented aspects of familiar voice perception. It will attempt to address questions surrounding the recognition and representation of different familiar voice identities, including voices at the highest levels of familiarity. In addition, it will probe the social and motivational significance of known, personally valued voices, and the potential effects of hearing them on brain and behaviour. This chapter will provide an overview of the current models of vocal identity processing to summarise our theoretical understanding of familiar and unfamiliar voice processing, and their underlying voice representations. Next, previous research into familiar voice and face recognition will be outlined. Lastly, the current view of the voice as a socially relevant signal will be discussed within a reward processing framework, to describe what is currently understood about this aspect of voices, and to lay the groundwork for the novel investigations in this thesis.

1.1 Vocal Identity Perception

A person's voice can deliver a wealth of information, from speech content to who is speaking (vocal identity), as well as insight into a person's age, sex, health, and more (Latinus & Belin, 2011). The information available upon hearing a voice may differ depending on whether that voice is familiar or unfamiliar to the listener. Unfamiliar voices belong to people we do not know or recognise. Listeners have the ability to discriminate between unfamiliar voices; for instance, to tell whether two heard signals were produced by the same or different people (van Lancker & Kreiman, 1987). One can also form general impressions of a person based on their voice for both familiar and unfamiliar voices, such as inferred personality traits, attractiveness, and current mood (Lavan & McGettigan, 2019). Familiar voices can further be recognised, and are those that belong to people we know. However, familiarity can range from a general feeling

that a voice has been heard before, to identification of the speaker by name (Kreiman & Sidtis, 2011). Moreover, the type and degree of familiarity within speakers that can be recognised by name can be somewhat varied. For instance, the voice of the barista in one's local coffee shop versus the voice of a romantic partner are likely to differ quite substantially in terms of the depth and breadth of experiences with these voices. One aim of the current thesis is to explore recognition of voices of differing familiarity, including voices of people we are most familiar with. Thus, I will begin by outlining existing theoretical models of how voices are recognised, as well as presenting previous literature examining familiar and unfamiliar voice processing, to demonstrate how these voices have been studied thus far, and identify key problems or gaps in our current understanding of the study of voice perception.

1.2 Models of Voice Identity

Being able to recognise voices is something that most human listeners can do, yet it is not an easy task. Producing voiced sounds (those involving vibration of the vocal folds) involves a combination of action of the respiratory system, the vocal folds and the vocal tract (López et al., 2013). A controlled flow of air from the lungs passes between the vocal folds causing them to oscillate, alternating between blocking and opening the airway. This creates changes in air pressure which are perceived as sound. Different configurations of the articulators such as the tongue, lips, jaw, and soft palate in the supralaryngeal vocal tract can be varied to further control the sounds produced, such as speech sounds or accents (Kreiman & Sidtis, 2011). Differences between individuals' voices can be based on variations in this vocal anatomy, or on learned behaviours. Characteristics based on anatomy, such as mean fundamental frequency (F0; linked to the rate of vocal fold vibration and is closely related to perceived pitch) or formant spacing (correlates with length of the vocal tract), are considered to be good indicators of vocal identity as they are fairly consistent over time. Differences based on experience – i.e. how the person's vocal patterns or habits have been learned and developed over time - are features such as accent, speaking rate, and other aspects of voice quality (Kreiman & Sidtis, 2011). Thus, human voices possess a similar basic structure, but variations in acoustic properties (determined by e.g. differences in physiology and/or voice use) allow for unique individual vocal identities. In order to recognise and remember individual voices, humans are tasked with extracting what are often subtle differences in these features across individuals (Latinus & Belin, 2011). Although each individual voice and what it is capable of producing is somewhat constrained by the anatomy of the individual's vocal system (e.g. the rate of vocal

fold vibration can be varied within an individual, but the range of this variability is affected by the length and mass of the vocal folds (Kreiman & Sidtis, 2011), the variety of sounds that can be produced from a single individual's vocal apparatus is extensive. By manipulating aspects of respiratory, laryngeal, and vocal tract systems, human beings are capable of creating an almost infinite number of linguistic utterances, varying the loudness, frequency of speech, and pitch, to add dimension and richness to speech or other vocal sounds. Speaking style can also be adjusted, so that we can talk, shout, laugh, sing, or read from written text, as well as adapting this style to communicate with different audiences (e.g. a colleague vs. a pet; Lavan, Burton, Scott, & McGettigan, 2019). Humans will also never produce exactly the same sound twice, and thus to recognise a voice, listeners must overcome this to produce a stable percept, possibly by extracting invariant features in vocal signals (Lavan, Burton, et al., 2019; Latinus & Belin, 2011). Therefore, being able to perceive identity in spite of both intrinsic factors, such as subtle differences in acoustic properties, as well as the large potential variability from utterance to utterance, is an impressive feat. This highlights the complexity of voice recognition, and various models have been proposed to try to explain how this challenging task is accomplished.

1.2.1 Vocal identity processing as a sequential process: The "Auditory Face" model

One longstanding model adapted from Bruce and Young's (1986) face perception model, was termed the "auditory face" model, and viewed voice processing as a sequential process (Bruce & Young, 1986; Belin, Fecteau, & Bedard, 2004). This model proposed that voice processing first involves low-level auditory analysis. Following this, structural encoding or analysis of the voice occurs. At this stage, three types of vocal information – identity, emotion, and speech – are extracted in partially distinct but interacting systems (Belin, Bestelmeyer, Latinus, & Watson, 2011). This involves perceptual analysis, extracting stable features of the voice (e.g. mean F0) which are then further processed in these three systems. For vocal identity processing, it is proposed that recognition of familiar voices involves matching the extracted features from the structural encoding phase to stored representations in "voice recognition units" (Bruce & Young, 1986; Belin et al., 2004, 2011). The three pathways (identity, affect, speech) are thought to interact with each other at all stages, as well as interacting with homologues in the face processing pathways, which allows for multimodal integration. Lastly, semantic information (e.g. the person's name or occupation) can be accessed when a voice is

recognised, and this is thought to be accomplished via a person identity node (PIN) containing this semantic information (Maguinness, Roswandowitz, & von Kriegstein, 2018).

1.2.2 The Prototype Model of Vocal Identity Processing

One influential model that has a lot of supporting evidence is the prototype model (Lavner, Rosenhouse, & Gath, 2001). This model proposes that there is a multidimensional 'voice space' in the brain where each encountered voice is represented. This voice space is thought to comprise perceptual features that can be used to identify a speaker, such as glottal pulse rate (GPR) or acoustic properties determined by vocal tract length (VTL). At the centre of this voice space is the prototype voice, which is thought to be an average of speakers' features that we have encountered, or a very common voice. Prior exposure to voices in one's environment establishes this prototype, and thus people from a similar geographical location or community may possess a similar prototype pattern (Lavner, Rosenhouse, Gath, 2001). A theory previously often proposed as an alternative to the prototype account is that of an exemplar-based or episodic model of vocal identity processing (see Valentine, 1991 for faces). Under this account, voices are thought to be stored as specific exemplars (rather than an abstracted average) in long-term memory, such that new incoming vocalisations are matched to the nearest matching exemplar for recognition to be achieved (Lavan, Burton, et al., 2019). The existing evidence largely supports prototype-based processing of voices (e.g. see Latinus & Belin, 2011, for the observation of adaptation after-effects supporting prototype processing). Exemplar-based encoding is generally thought to be relevant when a stored representation/prototype has not yet been established, such as when encountering a new voice (Fontaine et al., 2017). Thus, under the prototype model, voice identities are not thought to be represented as distinct values in voice space, rather each voice is proposed to be stored in long-term memory via the acoustic features that deviate from the prototype or norm. These stored deviations are known as 'reference patterns.' Recognising a familiar voice involves comparing an incoming signal to stored reference patterns. If the distance between this signal and a particular reference pattern is less than a perceptual threshold, this is a match and the voice is recognised as belonging to the identity corresponding to that reference pattern. Empirical evidence for separate prototypes for male and female speakers (Latinus, McAleer, Bestelmeyer, & Belin, 2013), as well as average-based representations of individual voice identities (representing within-person variability) has been found (Lavan, Knight, & McGettigan, 2019a).

Evidence in favour of prototype or average-based processing of vocal identity comes from research investigating the recognition of distinctive and average-sounding voices. These studies find that voices that are closer to the prototype (i.e. smaller deviations from the average or 'very common' voice) are more confusable with other average-sounding voices, whereas distinctive voices are those that deviate further from the prototype, and are therefore easier to recognise (Sørensen, 2012). For instance, a voice line-up study used voices selected from a distribution of 60 male Danish speakers containing their mean F0 values. Voices taken from the centre of the distribution were defined as common, and voices selected from the tails were defined as distinctive. Participants listened to 30 seconds of one common and one distinctive voice and returned a week later to attempt to select each voice from a line-up. Distinctive voices were significantly better recognised in the line-up compared to common voices. Further, a study by Barsics & Brédart (2012) explored the amount of semantic information that could be retrieved from distinctive and average faces and voices. The authors found that distinctiveness of both voices and faces led to greater retrieval of semantic and episodic information.

1.2.3 Integrative Model of Vocal Identity

Maguinness, Roswandowitz, and von Kriegstein (2018) expanded upon the prototype and sequential models in an attempt to reconcile a shortcoming of the "auditory face" model: specifically, that this model could not account for clinical findings where dissociations have been reported in individuals with particular brain lesions who cannot discriminate between unfamiliar voices, whilst recognition of familiar voices remains intact (e.g. Van Lancker, Cummings, Kreiman, & Dobkin, 1988). This suggests a partial dissociation between the perceptual processing of unfamiliar voices and recognition of familiar voices. In adapting these models to explain this partial dissociation, this integrative account details how unfamiliar voices may become familiar stored representations and this is worth discussing here. This account also highlights the likely brain regions underpinning the different voice processing stages. Within this model, the first step is termed "identity feature analysis." This stage is where features supporting identity (as well as emotion and speech processing) are analysed perceptually. This is thought to be supported by the posterior STG/S, the planum temporale, and Heschl's gyrus. Next, in mid regions of the STG/S, and similarly to the prototype model, these features are compared to the prototype. The deviations from this prototype are extracted and compared to stored reference patterns (reference pattern comparison). The model separates at this point, depending on whether the voice is recognised as familiar or not. If the distance

between the extracted deviations from the incoming signal and the stored reference patterns is smaller than a certain threshold, then a voice can be recognised, and this likely takes place in the anterior STG/S. However, for unfamiliar voices where there is not yet a stored reference pattern or representation, a reference pattern needs to be established. The authors of this model argue that reference patterns are established via an iterative loop involving early identity feature analysis and comparison to the prototype. This has been termed the "perceptual voice identity processing loop" (Maguinness, Roswandowitz, & von Kriegstein, 2018). The idea is that reference patterns will be refined with continued iterations through this perceptual processing loop. Moreover, it is proposed that distinctive voices may require fewer iterations through this loop as these voices deviate further from the prototype. With continued exposure, a unique reference pattern is stored for the new voice, joining other stored reference patterns.

The likely brain regions implicated in supporting the hierarchy of processes posited in the integrative model have been identified using functional neuroimaging experiments. An early study by Belin, Zatorre, Lafaille, Ahad, & Pike (2000) used functional magnetic resonance imaging (fMRI) to measure brain activation whilst participants passively listened to vocal sounds (speech & non-speech) and environmental sounds (e.g. nature sounds, animals, musical instruments, cars etc.). The authors found greater responses to vocal compared to non-vocal sounds in bilateral regions of the superior temporal sulci/gyri, with the maximum voice-sensitive activation observed in the upper central part of the STS (Belin et al., 2000). These voice-selective regions have become known as the temporal voice areas (TVAs; Belin, Zatorre, & Ahad, 2002; von Kriegstein & Giraud, 2004; Pernet et al., 2015). These regions along the STS and STG have commonly been implicated in various stages of voice processing. For instance, more posterior portions of the temporal lobe, including Heschl's gyrus, planum temporale, and posterior STS (pSTS) have been associated with acoustic processing of vocal identity, such as vocal pitch or vocal tract parameters (e.g. von Kriegstein Warren, Ives, Patterson, & Griffiths, 2006; von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010; Kreitewolf, Gaudrain, & von Kriegstein, 2014). More anterior portions of the right STS have been implicated more in later stages of voice processing, namely the recognition or identification of speakers in the mid to anterior STS/G (Belin & Zatorre, 2003; Andics et al., 2010; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). This supports a common finding of a posterior to anterior gradient in the superior temporal lobe involved in voice processing, whereby more posterior regions are thought to be implicated in general sensory processing, whereas moving anteriorly towards the temporal poles, regions are associated with recognising

specific identities (Luthra, 2021). Maguinness, Roswandowitz and von Kriegstein (2018) therefore identify these more posterior portions as being involved in the earliest stages of voice processing. Later stages in the model (comparisons to prototype, voice recognition) are proposed to be supported by mid (mSTS/G) to anterior (aSTS/G) portions of the STS/G (Maguinness, Roswandowitz, & von Kriegstein, 2018).

1.3 Previous Research into the Processing of Identity

Early studies explored how well listeners can recognise voices, finding that humans can recognise familiar voices extremely well under normal listening conditions, compared to the abilities of other non-human animals in recognising conspecifics. The number of voices humans can learn to recognise is extensive, and there is thought to be no upper limit to this ability (Kreiman & Sidtis, 2011). The integrative model outlined above details differences in the processing of unfamiliar and familiar voices, and proposes a theory as to how unfamiliar voices may become familiar over time. Studies of vocal identity also often make distinctions between familiar and unfamiliar voices, attempting to understand differences in recognition/discrimination ability for these two categories, or the factors important for recognition/discrimination. To do this, researchers typically manipulate some aspect of the voice – or rely on naturally varying aspects of the signal – and observe the effects of these manipulations on voice recognition/discrimination (Kreiman & Sidtis, 2011). Some manipulations have included: manipulating voice acoustics (Lavner, Gath, & Rosenhouse, 2000; Gaudrain, Li, Ban, & Patterson, 2009), varying linguistic factors (e.g. using familiar vs unfamiliar languages, forward vs. reversed speech; Zarate, Tian, Woods, & Poeppel, 2015; Levi, 2017; Perrachione, 2018; Winters, Levi, & Pisoni, 2008), duration of utterances (e.g. single syllable vs. sentences; Schweinberger, Herholz, & Sommer, 1997), and type of speech (e.g. neutral vs. whispered speech; Yarmey, Yarmey, Yarmey, & Parliament, 2001), to name a few. To understand the existing knowledge about voice recognition and familiar and unfamiliar voice processing, I will outline the available existing research into vocal identity, including both behavioural and neuroimaging findings. Research investigating identity from faces may provide relevant insights into likely effects in voices due to the commonalities proposed between face and voice processing (Yovel & Belin, 2013; Kuhn, Wydell, Lavan, McGettigan, & Garrido, 2017). For example, the voice has often been referred to as an "auditory face" (Belin, Fecteau, & Bedard, 2004). Therefore, observations from the face perception literature will also be discussed. The following section will outline studies exploring differences in

recognition of familiar and unfamiliar voices, including research investigating the factors affecting voice recognition and the robustness of familiar voice representations. Next, I will note that there are different types of familiar voices and explore the differences between these types, proposing a need for research to acknowledge that familiarity is not a binary concept.

1.4 Factors Affecting Person Identity Processing

Typically, behavioural research into person identity processing has often attempted to understand it via measuring the effects of experimental manipulations of the stimuli or familiarity on discrimination or identification ability, using a range of identity perception-related tasks. Research in the face perception literature has manipulated aspects of face images to explore how recognition or discrimination are affected (e.g. altering facial features). The recognition of familiar faces has been found to be robust to changes in viewpoint, lighting, image resolution, as well as naturally occurring variability (e.g. within-person variability; Ramon & Gobbini, 2018; Guntupalli & Gobbini, 2017). There is also some evidence to suggest that features such as identity and gender can be processed automatically from familiar but not unfamiliar faces (Yan, Young, & Andrews, 2017). Research by Bruce, Henderson, Newman, & Burton (2001) showed participants CCTV footage of university lecturers with whom the participants were either familiar or unfamiliar, alongside a photograph of a face. Participants were required to judge whether the face photograph matched the person in the CCTV footage. It was found that familiar participants were able to decide whether the identities were a match or mismatch with high accuracy, despite the degraded video footage, whereas unfamiliar participants were much poorer. Differences in the underlying representations or stored reference patterns for familiar and unfamiliar people have been proposed to explain differences in recognition or discrimination ability for faces and voices (Burton, Kramer, Ritchie, & Jenkins, 2016; Lavan, Burton, et al., 2019). For unfamiliar or newly-learned people, stored reference patterns either do not exist or are relatively unstable. In the presence of low or no familiarity, there is a proposed increased reliance on visual properties of specific unfamiliar faces and on distinguishing low-level acoustic properties for voices (McGettigan, 2015). Familiar person recognition, on the other hand, involves comparing the observed signal to stored reference patterns in memory. Therefore, when there are changes to visual/acoustic properties, this may affect unfamiliar people to a greater extent, whereas for familiar people, the relatively stable representations built up over time allows them to better contend with changes in image/voice properties (Lavan, Burton, et al., 2019).

In the voice perception literature, similar effects have been found, whereby familiar voices are better recognised than unfamiliar voices, and familiar voices are more robust to manipulated or naturally varying changes to the vocal signal. For instance, research has found that the type of vocalisation can dramatically affect voice recognition or discrimination (Lavan, Scott, & McGettigan, 2016; Lavan, Short, Wilding, & McGettigan, 2018; Peynircioğlu, Rabinovitz, & Repice 2017; Bartle & Dellwo, 2015; Wagner & Köster, 1999; Reich & Duke, 1979). A study by Peynircioğlu, Rabinovitz, and Repice (2017) tested unfamiliar listeners' ability to discriminate whether two voices belonged to the same speaker or different speakers, and used both spoken and sung speech. The authors found that discrimination was most accurate when determining whether two spoken excerpts were from the same talker, followed by discriminating between two sung excerpts. The poorest performance was observed for cross-modality same-different judgements (i.e. spoken – sung speech pairs). Also using unfamiliar listeners, other research studies have found better recognition from spoken speech versus when it was whispered (Bartle & Dellwo, 2015), and that vocal disguise (e.g. hyper-nasality, hoarse voice) was highly detrimental to unfamiliar participants' ability to decide whether two excerpts were produced by the same or different speakers on each trial (Reich & Duke, 1979). Studies using familiar voices have generally shown better performance for known others, whilst also finding impairments in performance in less-than-optimal conditions. For instance, Wagner and Köster (1999) used voices of the participants' colleagues reading text from a blackmailer's telephone call. This text was read both in a normal voice and using a falsetto voice as a disguise. Participants were required to name the voices they were familiar with and were not told that the speakers would sometimes be using a falsetto voice. The falsetto and normal voice clips were intermixed within the testing session. The authors found that for normal speech, listeners could recognise the speakers with very high (97%) accuracy, whereas this fell to 4% when speakers used a falsetto voice. Exploring familiarity with a language, research has also found that for unfamiliar listeners, discrimination of vocal identity is easier and more accurate when speakers are using the same language across stimuli, than in the case of cross-language discriminations (Wester, 2012). Furthermore, recognition of familiar lab-trained voices is more accurate the more phonologically similar their language is to one's native language (Zarate et al., 2015).

1.4.1 Familiar vs. Unfamiliar Voice Perception

All of the studies mentioned in the previous section have either used familiar listeners or unfamiliar listeners as participants but surprisingly do not make a direct comparison between familiar and unfamiliar voice processing, such as comparing groups listening to the same voices, or presenting participants with both familiar and unfamiliar identities within a single task. As unfamiliar voices by definition cannot be recognised, it is not possible to use a recognition task with these voices. Thus, discrimination tasks have often been used to probe unfamiliar voice processing, and identification or recognition tasks for familiar voices (Sidtis & Kreiman, 2011). Thus, because of the differences in task designs, this makes it difficult to compare performance for familiar and unfamiliar voice processing. Despite it being possible to use discrimination tasks with both familiar and unfamiliar voices, these have not typically been employed with familiar listeners. However, one study that directly compared performance of familiar and unfamiliar participants using the same discrimination task was Lavan and colleagues (2016). In this study, the voices of five talkers who were lecturing staff at the university were recorded. Participants were either students who were taught by these 5 speakers (familiar) or students from other departments not taught by the speakers (unfamiliar). The stimuli used were vowels and laughter produced under differing levels of volitional control. In the discrimination task, pairs of stimuli were presented and participants had to decide whether they thought they were produced by the same or different speakers. Familiar listeners displayed greater accuracy overall compared to unfamiliar listeners, and were better in a separate voice identity recognition task involving forward and reversed speech produced by the same lecturers (where unfamiliar listeners were given only a brief pre-exposure to the talkers before the test). However, in the discrimination task both groups were similarly negatively affected when having to make judgements across vocalisation types and when there were differences in volitional control. This finding supplied some information about the potential limits of familiar voice processing. In particular, the familiar listeners' experience with their lecturers' speech was confined to specific contexts, i.e. during lectures, whereas their experience with non-verbal vocalisations was likely to be more limited. As a result, identity perception of forward speech was highly accurate for these listeners, whereas judgements involving non-verbal vocalisations was impaired. This shows that identity perception performance can be quite fragile/unstable, even for speakers that are known. This may be in part due to having to learn how to generalise one's representation of a vocal identity across the different types of sounds a speaker is capable of producing. The ability to generalise arguably allows listeners to create a stable percept of a voice, enabling these voices to be recognised under a wide range of contexts. The following section will further outline an existing body of

research that investigates the extent to which familiarity aids listeners' ability to contend with the natural variability that exists within individual voices, using 'voice sorting' tasks.

Over the past decade, sorting tasks have been used with face and voice stimuli to explore familiar and unfamiliar participants' ability to group stimuli by identity (Jenkins, White, Van Montfort, & Burton, 2011). In these tasks, participants are presented with a number of photographs or voice clips, and are asked to sort these into piles/clusters according to identity. The task has been particularly useful in allowing researchers to test how listeners cope with the natural variability in everyday speech by measuring two key aspects of identity perception: 1) the ability to tell different identities apart, and 2) the ability to match different instances of the same person together ("telling together"). Furthermore, because sorting does not require explicit recognition or naming judgements to complete the task, it can readily be used with both familiar and unfamiliar observers/listeners. In the first published face identity sorting study, images of two Dutch celebrities were used. In one experiment, English participants unfamiliar with the celebrities were used, and in another, familiar Dutch participants were used. It was found that the unfamiliar participants sorted the faces into a median of 7.5 identities, whereas familiar participants perceived 2, the true number in the set. For unfamiliar participants, piles rarely contained photos from both identities. Thus, both familiar and unfamiliar participants had no issue in telling people apart, rather it was an issue in telling people together (i.e. knowing that two images belong to the same identity). The finding that familiar listeners are able to perceive closer to the true number of identities has now been replicated several times in the face processing literature (Andrews, Jenkins, Cursiter, & Burton, 2015; Redfern & Benton, 2017; Zhou & Mondloch, 2016).

Sorting tasks have also been used with voices (Lavan, Burston, & Garrido, 2019a; Lavan, Burston, et al., 2019b; Lavan et al., 2020; Stevenage, Symons, Fletcher, & Coen, 2020). Similarly to faces, when voices are unfamiliar, many more identities are perceived, indicating that natural variability within a single speaker is perceived as between-person variability when a listener is unfamiliar. Lavan, Burston, and Garrido (2019a) used two famous voices from the TV show "Orange is the New Black." Participants were either familiar (had watched the show) or unfamiliar (had not watched the show), and were required to sort the voice excerpts into piles according to identity. Consistent with face sorting studies, familiar listeners created fewer clusters (between 3 and 4) compared to unfamiliar listeners (between 4 and 9). Mirroring the findings in the face perception literature, both unfamiliar and familiar listeners did not present

with any difficulty in "telling people apart." Rather, unfamiliar listeners' errors manifested as errors in "telling people together", that is, knowing that different exemplars were produced by the same speaker. Thus, these sorting studies highlight that familiarity aids the ability to tell different instances of the same person together, to withstand natural within-person variability whereas unfamiliar listeners are impaired. Even research informing listeners that there were only two identities in the set did not lead to accurate identity performance in unfamiliar listeners, highlighting the difficulty of this task when unfamiliar (Lavan et al., 2020). Relating back to the models of vocal identity outlined previously, the differences in performance for familiar and unfamiliar listeners suggest differences in the listeners' underlying voice representations for the identities encountered in these tasks. That is, for familiar voices, reliable voice representations may be established, allowing listeners to contend with natural variability, whereas representations may not yet exist or be weak/partially formed for unfamiliar or low-familiar voices.

Moreover, as previously outlined above, manipulating stimulus properties (e.g. spoken vs. sung speech, language of speech) can produce different effects on voice identity recognition (Peynircioğlu, Rabinovitz, and Repice, 2017; Wester, 2012). This has also been observed in voice sorting studies that manipulate the vocalisations used or the type of familiarity. For instance, Stevenage and colleagues (2020) used voices of teaching staff at the authors' university, finding that familiar listeners (students taught by these speakers) sorted the identities into fewer clusters, indicating better performance, compared to unfamiliar listeners (not taught by the speakers). In this study, familiar listeners were better at both telling the voices together and telling apart compared to unfamiliar listeners, which differs from previous findings. As mentioned above, it is commonly found that whilst familiar listeners are better at telling voices together, both familiar and unfamiliar listeners are adept at telling voice identities apart. This discrepancy was explained by the authors as potentially being due to the use of personally familiar voices in their study compared to publicly familiar (i.e. famous) voices used in previous studies. Other research has found that familiar and unfamiliar listeners alike made more errors in telling together and telling apart voice clips when they were highly expressive (e.g. shouting, growling; Lavan, Burston, et al., 2019b). These results are also similar to Lavan and colleagues' (2016) study finding that performance was negatively affected for vocalisations that familiar listeners had less experience with (Lavan, Scott, & McGettigan, 2016). Therefore, the existing research suggests that there are advantages to familiarity for recognition of both faces and voices, and that this can be explained by differences in the

underlying representations for familiar and unfamiliar others. However, research has also suggested that representations may vary in quality depending on the amount and type of prior experience listeners have with them, implying that differences can exist not only in perception between familiar and unfamiliar voices, but also between different types or levels of familiarity.

1.4.2 Familiarity as a Continuum, Not a Binary Concept

Whilst unfamiliar voices are defined as those we have not encountered before, voices defined as "familiar" in the study of identity processing have included lab-trained voices, celebrity voices, and personally familiar voices. While these voices are all familiar, there are key differences between them, as has been alluded to in the previous section and which will be outlined here.

A frequently used category of familiar voices that only exist in experimental settings are lab-trained voices. The use of these voices allows for control over the level of familiarity that listeners have, as well as over the content and quality of the testing materials. The length and type of training varies, and training most commonly involves learning only the voice of the individual alongside a name, which differs from naturalistic experiences of learning the voice of a newly-encountered person typically alongside what they look like via social interaction. Some studies do provide associated faces, traits, or semantic information during training of new voices, but importantly, these voices do not belong to individuals one has had personal interactions with.

In addition to lab-trained voices, famous people are commonly used in person perception research. Familiarity with famous people is created through exposure in the media (Kreiman & Sidtis, 2011), and this category of people has been an attractive choice for studying voice perception/processing as it allows for the creation of stimulus sets that include voices familiar to a large group of participants. However, listeners can be differentially familiar with these voices, which can be difficult to control (Kreiman & Sidtis, 2011). However, this is also true of naturally acquired personally familiar voices. Famous people make up a unique category of familiar people as often details related to them (i.e. person knowledge) are stored, setting these apart from lab-trained voices. However, similarly to lab-trained identities, participants have typically had no prior personal interactions with the famous people used in experiments.

Lastly, individuals that we have real world experience and interactions with are termed personally familiar (Sugiura, 2014). Personally familiar people include, but are not limited to, work colleagues, mere acquaintances and friends, which have been termed common personally familiar people, as well as family members, close friends, romantic partners, and the self, which are highly personally familiar or "unique" personally familiar people (Sugiura, 2014). Hearing the personally familiar voice also often generates information connected to the speaker, such as person knowledge, memories of previous encounters/interactions, emotions towards the individual, social context (i.e. relationship to me), and historical details, which can affect our behaviour towards, and interactions with these individuals (Kreiman & Sidtis, 2011). It should be noted that within this category, experiences with these speakers can differ significantly. For instance, colleagues may have been heard in restricted contexts, allowing limited semantic knowledge of them beyond the workplace, and one's own voice can sound different when hearing it recorded compared to hearing it as one speaks (due to bone conduction). It is important to consider these variations when designing experimental manipulations of familiarity.

In sum, the experiences we have with different types of familiar people can be quite diverse, and this may affect the processing and underlying representations of these voices stored in memory. Comparing different types of familiar people within the same study can give some insight into whether this might be the case. In the next paragraphs I review some of this research in the extant voice and face perception literature, to explore and describe whether observable differences exist in the processing of different types of familiar people.

For faces, a study by Herzmann, Schweinberger, Sommer, Jentzsch (2004) explored priming effects using personally familiar (lecturing staff), famous, and unfamiliar faces, as well as measuring skin conductance responses (SCRs), whilst participants made familiarity judgements. The authors found larger SCRs for personally familiar faces compared to famous and unfamiliar faces. In the priming task, participants were required to choose whether a target face was familiar or unfamiliar, and this target was preceded by either the same identity (primed), or a different identity (unprimed). In this task, the smallest priming effects were observed for unfamiliar faces, but there were no differences in priming effects between personally familiar and famous faces. This finding of no difference between the two different types of familiar faces used has since been proposed to be due to the type of personally familiar

faces used in this study. The use of lecturing staff may not best represent the personally familiar category, due to not being "familiar enough" with these individuals. Another study by Walton and Hills (2012) also compared personally familiar, famous, and unfamiliar faces, but used the faces of the participants' parents, who are arguably substantially more familiar than lecturing staff. The face distortion aftereffect was examined, which is a phenomenon whereby adaptation to distorted faces (e.g. spatially compressed image) causes undistorted faces to be perceived to be distorted in the opposite direction (e.g. expanded; Walton & Hills, 2012). One image per condition (personally familiar, famous, unfamiliar) was used across both test and adaptor conditions, though this image was distorted in accordance with the experimental design. Adaptation aftereffects were examined for test faces (personally familiar, famous, unfamiliar) following adaptor images that were personally familiar, famous, or unfamiliar. That is, aftereffects for all test face conditions were examined following every type of adaptor (personally familiar, famous, unfamiliar). In this way, effects of the type of familiarity on adaptation could be examined within and between identities. The authors found larger differences in adaptation between the personally familiar face and the other two types of face (famous, unfamiliar). Specifically, adaptation aftereffects for all test face conditions (personally familiar, famous, unfamiliar) were overall very small following adaptation to a personally familiar face. For the famous and unfamiliar faces, adaptation aftereffects were most pronounced when the familiarity of the adaptor image and test image were matched, i.e. famous (distorted) adaptor – famous (undistorted) test image. However, for personally familiar faces, this within-familiarity adaptation effect was significantly smaller. The authors argued that due to a stronger knowledge of the possible variability in personally familiar faces, adaptation is less effective, possibly as a function of knowing that these faces cannot be distorted in the un-naturalistic ways as used for the adaptor images. Similarly, Liccione and colleagues (2014) examined reaction times for judging the familiarity of personally familiar (participant's relative or significant other), famous, and unfamiliar faces. The faces were either presented upright or inverted. Reaction times were significantly faster overall for personally familiar faces compared to the other two types of faces, and there were no effects of inversion (upright/inverted) on speed of categorisation for personally familiar faces. Famous faces were categorised significantly faster than unfamiliar faces, however both famous and unfamiliar face familiarity judgements were slower when these faces were inverted compared to upright. There is also some evidence that different types of familiar faces may be processed differently in the brain. One study used fMRI to compare neural activity when viewing personally familiar (parents/partner/own), famous, and unfamiliar faces, finding that the regions observed and the

extent of brain activity differed based on the type of familiarity (Taylor, Arsalidou, Bayless, Morris, Evans, & Barbeau, 2009). For example, all familiar faces led to activation in the fusiform gyrus, but one's partner's face showed additional activation in the parahippocampal gyrus, insula, amygdala, and thalamus. Moreover, famous faces revealed predominantly right lateralised activation, whereas personally familiar faces (partner, own face) were processed bilaterally, and activation was modulated by factors such as how much time is spent with the individuals. Therefore, these findings illustrate that familiarity is not a binary concept i.e. familiar or not, but that there are differences depending on the degree of familiarity and the range of experiences we have with different familiar people.

For voice perception, comparisons between voices of different degrees of familiarity, particularly those comparing personally familiar voices to other types of familiar voices, are rare. Of the few existing studies, research has either focused on the degree of familiarity within a particular familiar voice "category" (e.g. duration of training for lab-trained voices), or has compared recognition of voices belonging to different types of familiar voice categories (comparing personally familiar to lab-trained recognition). For instance, one study explored the effects of the amount of lab exposure on voice recognition and speech intelligibility, finding that there was little difference in recognition accuracy for participants who had received 10, 20, or 60 minutes of training (Holmes, To, & Johnsrude, 2021). Taking a different approach, Fontaine, Love, and Latinus (2017) explored voice recognition of famous and lab-trained voices in two experiments. Participants were required to decide who was producing a single vowel sound on each trial, in a three alternative forced choice task. Some of these voice excerpts were the original speakers' voices, and others were "speaker averages" created by morphing multiple utterances from the same speaker (Fontaine, Love, & Latinus, 2017). Identification was improved for famous voices, coupled with a decrease in reaction times, when the number of utterances included in the speaker averages increased. This effect was not observed for lab-trained voices. In fact, performance decreased with increasing averageness. These results suggested a qualitative difference in the recognition of famous and lab-trained voices. That is, for lab-trained voices, a stable voice representation or reference pattern had not been formed, whereas for famous voices, an abstract representation may have been extracted presumably due to having more experience with celebrity voices via the media. This is supported by research finding that during the learning of voices from variable exemplars, listeners can automatically extract speaker averages and further recognise these voices from their averages, despite having never heard them (Lavan, Knight, & McGettigan, 2019a).

Moreover, a recent study by Plante-Hébert, Boucher, and Jemel (2021) was interested in the differences between recognition (as defined as recognising whether a person is familiar/unfamiliar) and explicit identification (being able to state who the person is). The researchers used electroencephalography (EEG) with personally familiar and lab-trained voices to explore the time course of these processes. Personally familiar voices were close friends, family members, or romantic partners who could be identified by the participants, whereas lab-trained voices were not explicitly trained but rather became more familiar with repetition over the course of the experiment (thus facilitating recognition but not identification). Differences in event-related potentials (ERPs) were found for personally familiar voices compared to lab-trained (or frequently heard stimuli) and unfamiliar voices. That is, different responses were observed for personally familiar voices compared to unfamiliar ones in two separate time windows: the visual P2 component, occurring between 200-250ms, thought to reflect an early recognition stage (this was found in central-frontal sites), and a late positive component, occurring between 450-850ms, thought to be reflective of speaker identification. The lab-trained (or frequently heard voice) showed responses in a different time window (N250 component, between 300-350ms post-voice presentation). This study, whilst not a direct test of recognition, probed differences in the time course of processes underlying different types of familiarity, namely a feeling of familiarity vs identification, finding that neural responses differed as a function of the type of familiarity.

In the literature, the type or degree of familiarity with a voice is not always acknowledged to have an effect on the observed effects. In such cases, all familiar voices are grouped into one category and the label "familiar" is used without further qualification of the extent and type of this familiarity. This assumes that all familiar voices are processed in the same way. However, the above research comparing people with differing types of familiarity within the same experiment suggests that this may not be the case, and underlines two important issues. Firstly, it illustrates that familiarity is not a binary concept i.e. familiar vs. unfamiliar, but that the range of experiences one can have with different familiar others can result in differences in the recognition or processing of their voices/faces. Very few studies in the voice perception literature have compared and contrasted voices that differ in the extent of familiarity, highlighting a potential oversimplification in our understanding of familiar voice processing. Secondly, these results also often reveal a separation in observed effects for highly personally familiar people, compared with other less familiar people. For instance, the research on face identity processing outlined above found smaller adaptation effects for personally familiar

faces, that recognition was unaffected by face inversion, as well as different neural signatures of personally familiar people compared to other, less familiar people (Walton & Hills, 2012; Liccione et al., 2014; Taylor et al., 2009; Plante-Hébert et al., 2021). These results have been concluded to be due to stronger or more robust representations of personally familiar people stored in memory, and may suggest a "special" status for personally familiar people.

One of the reasons why few studies in the vocal identity literature have used personally familiar voices may be the constraints in obtaining voice recordings from highly familiar individuals (McGettigan, 2015). Famous voices have more of a many-to-one mapping (i.e. many participants are familiar with a particular celebrity), whereas personally familiar voices are familiar to smaller, more individualised groups. Of the studies that use personally familiar voices, common PF voices are often used, with university lecturers and their colleagues/students as listeners being most routinely chosen (Sugiura, 2014). "Unique" personally familiar voices, such as voices of close friends, family members, and romantic partners (often involving a one-to-one mapping of voice identities to familiar listeners in experiments) are even less frequent in vocal identity research. This suggests an incompleteness in our understanding of the processing of the full range of familiar voices that exist, particularly those that play the most important roles in our social lives (Ramon & Gobbini, 2018).

Nonetheless, a few studies do exist that used unique personally familiar voices. For instance, a study by Yarmey, Yarmey, Yarmey, and Parliament (2001) first recorded the voices of a group of speakers and then identified potential participants who would be familiar with their voices – this was done by asking the speakers to create a list of people that differed in terms of their familiarity to them, including names of people of high, moderate, and low familiarity. The named individuals were contacted and asked to participate as listeners in the experiment. A high familiar listener was a best friend or immediate family member of the speaker, a moderate familiar listener was a co-worker, team mate, or casual friend of the speaker, and low familiar people were acquaintances such as a neighbour who the speaker would not have interacted with for more than a few minutes in any week in the past year. Unfamiliar speakers were also included. In this study, it was found that with increasing familiarity came better recognition when listening to normal speech (i.e. conversational tone of voice, no background noise etc.). For whispered speech, accuracy was negatively affected for voices at all familiarity levels. However, for the high familiar voices, recognition was still high for whispered speech (79% correct responses vs. 89% accurate for normal speech), whereas for moderate familiarity and

below, performance dropped from between 61 and 75% accurate for normal speech to between 20 and 35% accurate for whispered speech. Another study by Huckvale and Kristiansen (2012) recruited friends who were in the same cohort on a University course as the speakers and listeners. The authors found that identification performance remained relatively high even when the voices were acoustically distorted (Huckvale & Kristiansen, 2012; cited in Krix, Sauerland, & Schreuder, 2017). This is in contrast to voice identity studies that used similar challenging conditions with less familiar voices and found much more detrimental effects on recognition (e.g. Wagner & Köster, 1999). Therefore, in both of these studies, personal familiarity with the voice allowed for recognition that withstood changes to the signal such as through whispering or artificial manipulation of acoustics. Personal familiarity for faces has previously been proposed to be characterised by view-invariant representations, detailed person knowledge, as well as evoking emotional responses. Similarly, PF voices have been associated with an ability to generalise across within-person variability to maintain stable recognition, as well as being associated with "packets" of information such as biographical knowledge, memories with the person, etc. (Kreiman & Sidtis, 2011). However, the use of personally familiar voices is relatively rare in vocal identity research compared to those involving unfamiliar, lab-trained or famous speakers. Whilst all of these categories of voices are imperative to understanding vocal identity processing, recruiting voices of highly familiar, socially important individuals such as those used in Yarmey and colleagues (2001) and Huckvale and Kristiansen's (2012) studies is fundamental to furthering our understanding of the full potential of the human voice processing system, the underlying representations of these voices stored in memory, as well as the potential wider implications of hearing these voices on aspects other than vocal identity (Ramon & Gobbini, 2018).

1.5 The Wider Benefits of Personally Relevant Familiar Voices

In addition to an improved ability to recognise who is speaking when familiar with a speaker, familiarity has also been observed to give rise to other behavioural benefits. One of these benefits is in enhancing speech comprehension. Specifically, when presented at the same time as a competing talker, or other background sounds, familiar voices have been reported to be more intelligible than unfamiliar voices (Nygaard & Pisoni, 1998; Johnsrude et al., 2013; Nygaard, Sommers, & Pisoni, 1994; Kreitewolf, Mathias, & von Kriegstein, 2017; Souza, Gehani, Wright, & McCloy, 2013; Holmes & Johnsrude, 2020; Domingo, Holmes, & Johnsrude, 2019). As the sections above demonstrate, the focus in the study of human vocal

communication has traditionally been on understanding the perception and production of speech (McGettigan, 2015), and while the "auditory face" model includes pathways for speech, emotion, and identity perception, it does not address how these processes might be modulated by the social and affective significance of the talker being heard. To elaborate, it is assumed that a rich array of affective and other types of social information accompany personally known, relevant voices (Sidtis & Kreiman, 2012). Familiarity has previously been associated with increased attention, evoked memories, motivation to approach/avoid, and fundamentally, a feeling of personal relevance (Purhonen, Kilpelainen-Lees, Valkonen-Korhonen, Karhu, & Lehtonen, 2004; Kreiman & Sidtis, 2011), which makes familiar individuals perfect candidates to study when attempting to understand the meaning or value of voices beyond perceptual processing and recognition.

Human beings are in constant contact with others, and there are few social events in everyday life that do not involve interacting with voices (Sidtis & Kreiman, 2012). Voices that are engaged with most regularly are probably those that are personally known and important to the listener (Tye-Murray, Spehar, Sommers, & Barcroft, 2016). Anecdotally, hearing the voice of a loved one, such as over the phone, can be a soothing and pleasant experience. This may imply that particular familiar voices may be personally relevant signals that are capable of influencing the internal state of the listener (Bliss-Moreau, Owren, & Barrett, 2010). Human faces have long been regarded as meaningful social stimuli (Jack & Schyns, 2015). However, unlike faces, little is empirically known about the cognitive and neural underpinnings of the familiar voice as a socially meaningful and affective signal in one's environment (McGettigan, 2015). Exploring this facet of familiar voices is the second broad aim of this thesis.

As mentioned earlier, neuroimaging studies using personally familiar voices predominantly exploit common personally familiar voices, such as work colleagues or lecturing staff, focusing on the perceptual discrimination and recognition of these identities rather than asking questions about their social or emotional significance. Nevertheless, these studies do allude to recruitment of systems involved in social communication in addition to core auditory regions (See Nakamura et al., 2001; Shah et al., 2002). In the face perception literature, where personally familiar faces have been arguably easier to source in the form of photographs, research presenting participants with their romantic partner's face have found significant activation in brain systems implicated in reward, affect, and motivation that are not seen for low familiar faces (Bartels & Zeki, 2000; Acevedo, Aron, Fisher, & Brown, 2012). The

recruitment of affective processing systems in the brain suggests that participants experience (positive) emotional responses to viewing personally familiar others, while engagement of reward and motivation regions implies that personally familiar faces may be similar to other stimuli that humans will work to receive (e.g. food/sex). These findings have been observed in both early and later stages of love, and appear to be independent of sexual preference and culture (Zeki & Romaya, 2010; Xu et al., 2011; Sugiura, 2014). Other neuroimaging research using the familiar faces of one's own children has also found engagement of systems implicated in reward and motivation, as well as those concerned with attachment and bonding (Leibenluft, Gobbini, Harrison, & Haxby, 2004; Wittfoth-Schardt et al., 2012). This illustrates that there may be something "special" about personal familiarity that could be reflected in the brain's response to these individuals.

As personally familiar people form a unique category in one's repertory of known voices, cues to these individuals may be processed differently to unknown or less valued individuals, in a manner that reflects their increased personal and social relevance. Emergent from the idea that people may be motivated to interact with others that will afford positive social outcomes, as well as the subjective value individuals place on those that they love, it can be hypothesised that hearing a loved one's voice may be socially rewarding and consequently induce responses in structures of the dopaminergic reward system in the brain. This idea is indirectly supported by evidence from animal studies finding connectivity between auditory cortices and reward regions (e.g. orbitofrontal cortex/OFC) in primates that are thought to be important for responding to social auditory signals (Camalier & Kaas, 2011), as well as some evidence in humans showing reward responses to hearing other types of auditory stimuli, such as music (Salimpoor et al., 2013). The aforementioned research into PF faces and reward also supports this hypothesis (e.g. Bartels & Zeki, 2000). Therefore, it is reasonable to predict that voices, particularly those of high social importance, may show reward-related patterns of activity in the brain. A more detailed account of the evidence in favour of the socially rewarding nature of voices is outlined in the following sections.

1.5.1 Reward Processing

The processing of rewards plays a vital role in daily life. To understand whether personally familiar voices can be socially rewarding, we first need to understand what constitutes a reward, both in the brain systems engaged, and the types of behaviour it can elicit. Rewards

are broadly defined as any stimulus, object, or event that has positive valence and the potential to influence our behaviour (Smith & Delgado, 2015). More specifically, three psychological components of reward processing have been described. Rewards: 1) have the ability to produce associative learning, 2) have the capacity to affect decision-making and induce approach behaviour (via motivational salience), and 3) elicit pleasure or other positive emotions (Schultz, 2015). If a stimulus is rewarding, humans (and non-human animals) will learn to repeat the behaviours that led to the reward, with the expectation of receiving that rewarding experience again (Schultz, 2015). Next, rewards are motivating, and effort is exerted to receive them. Motivation for rewards has also been termed 'wanting' (desire) or incentive salience. Incentive salience is a psychological process that makes reward cues appealing and induces approach behaviour (Berridge & Robinson, 2016). A similar term, motivational salience, is also sometimes used, but motivational salience is a process that can induce both approach or avoidance behaviour, whereas incentive salience refers only to positive reinforcers (Puglisi-Allegra & Ventura, 2012). Lastly, rewards are experienced as pleasurable. This is important, as if a stimulus is pleasurable, it may contribute to the desire to experience this stimulus/event again. This psychological component has been termed 'liking' (pleasure; Berridge & Robinson, 2003). The incentive-sensitisation theory of addiction posits that changes to systems mediating incentive salience can result in an increase in 'wanting' for drugs (for instance), without increasing 'liking', and thus dissociations can exist within these psychological components of reward (Berridge, 2012). Reward processing is not a unitary concept, but rather follows various stages that unfold over time (Novak, Novak, Lynam, & Foti, 2016). These include anticipatory and consummatory processing, which may be underpinned by partially distinct but overlapping systems of neural reward (Liu, Hairston, Schrier, & Fan, 2011; Husain & Roiser, 2018).

1.5.2 Reward System in the Brain

The brain structures that make up the reward system include the ventral tegmental area (VTA), the basal ganglia – comprising the ventral (nucleus accumbens/NAcc, olfactory tubercle) and dorsal striatum (caudate nucleus and putamen), globus pallidus, substantia nigra, ventral pallidum and subthalamic nucleus – as well as the prefrontal cortex, anterior cingulate cortex (ACC), insular cortex, hippocampus, amygdala, hypothalamus and thalamus (Yager, Garcia, Wunsch, & Ferguson, 2015; Berridge & Kringelbach, 2015).

An important role for dopamine (DA) in reward processing has been proposed. Dopamine is a neurotransmitter and dopaminergic neurons can be found in the mesencephalon (midbrain), diencephalon, and olfactory bulb, with these neurons making up less than 1% of the total quantity of neurons in the brain (Arias-Carrión, Stamelou, Murillo-Rodríguez, Menéndez-González & Pöppel, 2010). Nearly all of the existing DA neurons are found in the ventral midbrain. Two pathways that originate in the VTA are the mesolimbic and mesocortical DA pathways (Arias-Carrión & Pöppel, 2007). In the mesolimbic pathway, DA neurons project from the VTA to the nucleus accumbens (NAcc) and olfactory tubercle (i.e. the ventral striatum), as well as connections with amygdala and hippocampus. The mesocortical DA pathway involves dopaminergic projections from the VTA to regions in the prefrontal cortex, cingulate, and perirhinal cortex. Together, these pathways are also known as the mesocorticolimbic system (Arias-Carrión & Pöppel, 2007). DA systems are important for learning and reward-seeking (motivation), but may not have a role in hedonic aspects of reward (i.e. pleasure; Berridge & Robinson, 2016). Liking or pleasure has instead been recently associated with "hedonic hotspots" in the brain. Candidate regions for these hedonic hotspots have included limbic prefrontal cortex, including orbitofrontal cortex and insula, as well as the ventral pallidum, which has been found to be the only known region whereby lesions are capable of turning 'liking' into 'disgust' (Berridge & Kringelbach, 2015). In summary, opioid and endocannabinoid neurotransmitters are considered to be more important for 'liking'/pleasure, whereas dopamine is associated with 'wanting'/motivation.

1.5.3 Social Rewards

Rewards can either be primary or secondary reinforcers; primary reinforcers are those that serve a biological/evolutionary function (e.g. food, sex), and are thus unconditioned, whereas secondary rewards (e.g. money) are not directly related to survival, gaining value instead through learned associations with primary reinforcers (Sescousse, Caldú, Segura, & Dreher, 2013). Most rewards in modern life are secondary reinforcers. As human beings are inherently social, opportunities for experiencing rewards often operate within a social context (Fareri & Delgado, 2014). As mentioned, one aim of this thesis is to explore the question of whether voices can function as social rewards. Social rewards are generally defined as positive experiences that involve other people (Bhanji & Delgado, 2014). Neuroimaging data has suggested that social rewards may be processed similarly to other non-social rewards (Bhanji & Delgado, 2014). For instance, research using fMRI to compare responses to both social

(positive feedback), and monetary rewards found similar Blood Oxygen Level Dependent (BOLD) activation for both types of reward (Izuma, Saito, & Sadato, 2008). In a social feedback condition, participants viewed a photograph of their own face, and received 'feedback' which came in the form of a single word indicating the first impression independent raters had of them. Positive feedback was associated with an increase in dorsal striatal activity (in both the caudate nucleus and putamen), similarly to responses to monetary gain. Comparably, another study examining social rewards in the brain illustrated that being liked by others led to increased activation in the ventral striatum, as well as in regions including ventral medial prefrontal cortex (vmPFC), ventral midbrain, and cingulate cortex (Davey, Allen, Harrison, Dwyer, & Yücel, 2010). Thus, it appears that social rewards may be processed in regions implicated in processing other types of – non-social – rewards (Leotti & Delgado, 2011).

Whilst the above research demonstrates that contexts involving interactions with others can engage regions of corticostriatal reward circuitry, the question remains as to whether specific individuals themselves may be rewarding. Both faces and voices (amongst other physical cues to other people) contain a large amount of socially relevant information, regardless of familiarity with the individual (McGettigan, 2015). There is some evidence, primarily in the face perception literature to suggest that other people may be rewarding stimuli (Fareri & Delgado, 2014). Both familiar (personally relevant) and unfamiliar (attractive) faces have been shown to be socially rewarding. For instance, an attractive face may be intrinsically rewarding as it theoretically signals mate value, and this could motivate approach behaviours (Daniel & Pollman, 2014). Correspondingly, seeing the face or hearing the voice of a personally familiar other may also be a pleasurable, rewarding experience (Bartels & Zeki, 2000). The existing research exploring the socially rewarding nature of familiar and unfamiliar faces and voices will be outlined below.

1.5.4 Personally Relevant Familiar Others and Reward

1.5.4.1 Faces:

Engagement with personally relevant familiar individuals may be a socially rewarding experience, over and above any notion of conventional attractiveness. Perhaps prior affiliation with these individuals allows them to become highly valued or salient signals, leading to the experience of reward. Support for this concept stems from the face-processing literature, which

reports investigations into the neural systems implicated in viewing a loved one's face, whether this be a maternal figure, or a romantic partner (Bartels & Zeki, 2000). One such study examined neural activity to viewing photographs of one's romantic partner compared to viewing photographs of a familiar acquaintance, where the participants were in early-stage intense romantic love (Aron et al., 2005). Contrasting responses to the beloved's photograph with those of a neutral familiar other, the researchers identified increases in the BOLD response in key brain structures that have previously been implicated in other types of reward, such as monetary gain (Carter, Macinnes, Huettel, & Adcock, 2009). Regions included the medial and dorsal caudate, posterior cingulate cortex, and a region in the right midbrain around the VTA. It was argued that seeing a romantic partner's face engages neural systems associated with reward in order to motivate individuals to perform social behaviours towards their beloved (Aron et al., 2005). The methodology presented here was replicated in a Chinese sample (Xu et al., 2011). This study similarly described increased activity in the right VTA and caudate nucleus when viewing a significant other's face (Xu et al., 2011). Activation was also observed in other regions associated with reward processing, including the OFC and NAcc. Deactivations were also reported in both studies, specifically in the amygdala (Aron et al., 2005; Xu et al., 2011). This is a finding that has been consistently replicated in many studies of romantic love, and has been attributed to love reducing fearful responses (i.e. increased feelings of safety when with the partner; Aron et al., 2005).

Similarly, neuroimaging research has also investigated a later stage of romantic love. Acevedo and colleagues (2012) probed whether brain activity for long-term love (married >10 years) showed a similar profile of reward to early-stage romantic love. Married couples were required to view face images of their beloved as well as a close friend and low-familiar person whilst undergoing functional MRI. Significant activations were reported in dopamine-rich regions such as the VTA and dorsal striatum (similarly to early-stage love), as well as substantia nigra (SN), when comparing responses to the face of the beloved to a close friend or highly familiar acquaintance. Additionally, globus pallidus (GP), SN, dorsal raphe, and ACC (amongst other regions) activations were recorded that overlapped with activation in response to a maternal figure. These regions, especially the GP, have been linked to attachment and pair-bonding, the authors proposing that romantic love promotes pair-bond maintenance through sustained reward over the course of a romantic relationship (Acevedo et al., 2012). Hence, responses to a partner's face in both an early stage, and at a later phase are shown to engage similar systems of reward, yet longer-term romantic love appears to be associated with additional systems

linked to pair-bonding and attachment. However, the mechanisms underlying these longitudinal changes are not currently known. This raises the question of how cues to individual people, such as a romantic partner, gain value over time, as the bond between them strengthens and familiarity increases. One study by Xu and colleagues (2012) explored this using a longitudinal design, finding that relationship status after 40 months (i.e. staying together vs. breaking up) could be predicted from activation in the caudate tail displayed in early-stage love.

1.5.4.2 Voices:

In voices, it is less clear whether hearing a personally relevant familiar voice is rewarding in the same way as has been observed for viewing valued faces, nor is it known how a voice may gain significance over time. As outlined previously, an integrative model of vocal identity postulates that a voice becomes familiar and subsequently identifiable over time with repeated exposure to that voice, until a representation is formed (Maguinness, Roswandowitz, & von Kriegstein, 2018). However, this model, and existing research more broadly, has failed to capture the importance of the value of a personally familiar voice, over and above its identification as a known other. One aspect of Maguinness et al.'s integrative model that could serve as a potential candidate for processing the social significance of voices is what has been highlighted as the "semantic processing" stage in the model (proposed to be supported by an "extended system" of brain regions. The argument is that brain structures in this system (e.g. precuneus/posterior cingulate, amygdala, inferior frontal gyrus; Shah et al., 2001; Blank et al., 2014) further encode the "meaning" of voices, or evaluate one's feelings or relationship towards familiar voices. Nonetheless, the empirical evidence to support these assertions is currently absent.

Very few studies have examined the value of personally relevant voices, however there is some evidence to suggest that reward related responses may be expected in the vocal as well as the visual domain. An fMRI study by Ortigue, Bianchi-Demicheli, Hamilton, and Grafton (2007) used subliminal priming with participants' romantic partners, close friends, and strangers' names as the prime words and either words (nouns), nonwords, or blanks were used as the target stimuli. Participants were asked to indicate whether the targets were English words on each trial. Subliminal priming with a particular masked stimulus can have effects on performance or behaviour in a subsequent task. In particular, performance might be quicker or more accurate following particular primes. The researchers found that when the romantic

partner's first name was the prime, responses to the target word were significantly quicker. Responses to the presentation of the romantic partner subliminal prime were found in caudate nucleus and VTA amongst other regions. Participants also completed the passionate love scale (PLS) that measures the level of self-reported passionate love a person has with another (Hatfield & Sprecher, 1986). Scores on the PLS were correlated both negatively with reaction times and positively with activation in the VTA, meaning that the more in self-reported love participants were, the greater the activation observed in response to the partner's name in this region. These findings are in line with aforementioned research that showed PLS scores correlating with activation in the VTA when viewing a partner's face (Aron et al., 2002; Acevedo et al., 2012). Thus, if a name cue to an individual is sufficient to facilitate responses and engage reward and motivation systems in a similar fashion to faces, it is reasonable to predict that the voice of a personally relevant other, which is arguably more indicative of identity (compared to first names that can cue associations to multiple individuals) should also engage reward circuitry in the brain.

More directly related, a few studies utilising mother-child dyads also allude to the valued voice functioning as a socially rewarding stimulus. For instance, Seltzer, Ziegler, and Pollack (2010) investigated whether tactile and verbal contact from a child's mother could affect levels of cortisol and oxytocin in children after the child had engaged in a stressful task. Cortisol is a biomarker of stress, and Oxytocin has been found to be important for pair-bonding, the formation of trust, and stress regulation (Olff et al., 2013). The authors found the largest reductions in this hormone both when children were comforted through tactile and verbal means, and when they were solely comforted by their mother's voice. Increases in levels of Oxytocin were also observed in these conditions, relative to a baseline of not being comforted at all. Recent evidence has pointed to a potential role for the release of Oxytocin mediating reward responses in the brain (Hung et al., 2017; Scheele et al., 2013). Additionally, children comforted by their mothers via an instant text messaging service failed to display these biological responses – in contrast, these stress-relieving benefits were found when children were comforted by their mother's voice over the phone (Seltzer, Prososki, Ziegler, & Pollack, 2012). Thus, it appears that there is something about the voice of a highly significantly individual, not explained by the linguistic content of their communications, which is sufficient to bring about meaningful, positive biological changes in these participants.

Related research also using mother-child pairs explored the neural correlates of voice-related reward processing in children – Abrams and colleagues (2013) obtained resting state functional MRI data from both children with Autism spectrum disorders (ASDs) and typically developing (TD) controls. Functional connectivity of bilateral pSTS (a voice-selective region) was examined, and key differences were found between the groups. Compared to controls, children with ASD exhibited significant underconnectivity between left hemisphere pSTS and structures implicated in the brain's reward system, such as the VTA, ventral and dorsal striatum, and OFC. The authors also noted that the strength of connectivity between these systems predicted scores on the ADOS/ADI social communication subtests in this group. Hence, it was concluded that this connectivity may mediate social communicative skills in the neurotypical population, and impede the development of these skills in ASD. However, as this research examined resting state activity, it was unclear whether the two groups in the study would have responded differently to voices, particularly those belonging to highly familiar others. Therefore, drawing concrete conclusions about these differences and whether neural activity could predict communication abilities was difficult (Brock et al., 2013). Nonetheless, evidence exists with other auditory stimuli that has observed increases in functional connectivity between auditory cortex and the NAcc (a key reward region) in adults whilst listening to increasingly pleasurable music, providing some possible support for this conclusion (Salimpoor et al., 2013).

In an attempt to resolve the shortcomings of their prior study, Abrams and colleagues (2016) presented TD children with their mother's voice, female control voices, and environmental sounds whilst they underwent fMRI. In contrast to female control voices, the mother's voice induced greater activation in voice selective (STS) regions as well as those implicated in affect (amygdala), reward (ventral striatum, OFC), and salience (anterior insula, cingulate). Further to this, social communication skills in these TD children were correlated with the strength of connectivity between pSTS and regions of reward and salience detection. Hence, as in their previous study, the connection strength between voice selective cortex and reward circuitry appeared to be linked to communicative abilities in children, although the mechanisms underlying this are not yet currently known (Abrams et al., 2013). Secondly, the mother's voice, a highly valued signal in a child's life, engaged regions of reward and salience in the brain, similarly to the faces of personally valued others outlined previously. Therefore, it may be reasonable to predict that voices of other uniquely valued individuals may also engage similar regions.

Collectively, the limited research into the processing of highly familiar voices outlined above suggests that these voices may be powerful social signals in one's environment, with the potential to induce both positive biological changes, and engage regions of the dopaminergic reward pathway, similar to findings in the face perception literature. Moreover, preliminary findings point to the co-activation of voice selective and reward regions as impacting social communication. Thus, investigating the value of familiar voices in the brain and exploring the potential implications for social and cognitive functioning is necessary to identify the mechanisms underpinning the perception of socially relevant voices.

In summary, it has been demonstrated that faces personally known to the viewer can lead to recruitment of reward circuitry in the brain. In the voice perception domain, however, formal investigations of this have proven inadequate. Yet, the limited evidence that does exist suggests that a similar profile of reward may be observed here. For instance, research with children listening to their mother's voice highlighted both biological benefits (which may be linked to systems of reward), as well as illustrating the link between the auditory system and that of reward, relating this to successful social functioning. However, whereas the face perception literature has examined reward more directly by examining the ability for valued faces to motivate behaviour, no such studies directly explore this with voices. This leaves a gap in the literature for a more direct study of whether valued voices can be rewarding, by examining components that make up a reward, and how this presents in the brain systems engaged.

1.6 The Current Thesis

This chapter has outlined what is currently known about the processing of vocal identity in familiar and unfamiliar voices, as well as demonstrating the large variation that exists within what constitutes a familiar individual. I have shown that the type and extent of familiarity can have an effect on the conditions under which voices or faces can be recognised, and what this subsequently means for their underlying stored representations. The second half of this chapter used previous research in both the voice and face perception literatures to make predictions about the potential broader implications of hearing known, valued voices, such as their ability to serve as socially rewarding stimuli. The use of highly familiar voices and the possible implications of high familiarity with voices on identity perception has rarely been examined. Moreover, investigations of voice processing have largely neglected the personal relevance of

individual voices to listeners. Thus, large gaps in the literature exist surrounding the exploration of voices at the upper end of the familiarity spectrum, as well as in studying the emotional and social significance of hearing familiar valued others.

Therefore, the chapters in this thesis will follow two lines of inquiry. First, the potential perceptual benefits associated with personal familiarity, including effects on identity recognition and speech intelligibility/comprehension, will be explored. Particularly, these benefits will be explored under perceptually challenging listening conditions. The effects of the amount of training with a lab-trained voice on recognition ability and decision-making will also be examined. In this way, these questions (described in Chapters 2 and 3) will aim to further our understanding of vocal identity processing involving different types of familiarity, and how this is related to underlying voice representations. The second line of inquiry turns to investigate whether voices that are personally relevant to the listeners can be socially rewarding and evoke motivated behaviour, in a behavioural study (Chapter 4), followed by an exploration of the neural underpinnings of this in a functional MRI study (Chapter 5). Taken together, the experiments in this thesis aim to expand on our current knowledge of the perception of familiar voices, particularly with regard to our view, or definition, of familiarity and the social and emotional significance of particular familiar voices.

2 Familiarity Benefits for Voice and Speech Recognition

I declare here that a version of Chapter 2 has been published in the Journal of Experimental Psychology:General (Kanber, E., Lavan, N., & McGettigan, C. (2021). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology: General*, advance online publication. doi: 10.1037/xge0001112)

This chapter describes a set of three experiments that explore how differences in familiarity affect voice recognition and speech intelligibility. It directly compares voice recognition and speech recognition ability for voices that are highly personally familiar (romantic couples) to voices learned in a lab setting (lab-trained voices). The existing literature suggests that voice recognition can be improved with increasing familiarity, but that even recognition of familiar voices is error-prone when the listener is faced with perceptual challenges. However, the use of voices at the highest levels of familiarity (e.g. close friends, partners, family members) is uncommon in vocal identity research due to recruitment restraints, yet these are the voices we are likely to encounter most consistently, in a variety of contexts and over a prolonged period of time. Research suggests that these voices may thus be more robustly represented in memory, and more able to contend with challenges to perception. Therefore, this chapter explores the first broad aim of the thesis, using highly familiar voices to examine what perceptual benefits a personally familiar voice may afford a listener, compared to one that is lab-trained or unfamiliar, in three perceptually challenging experiments.

2.1 General Introduction

Humans are voice experts in that we have the ability to produce and understand speech. In addition, the human voice also conveys a wealth of socially-relevant paralinguistic information, including cues to a speaker's identity. Recognising who is speaking, though a crucial skill for communication, has been shown to be both challenging and error-prone. A speaker never produces exactly the same sound twice, and constantly adjusts the sound of their voice to express intentions, adapt to a range of speaking situations or to cater to different audiences. This means that an individual speaker has the capacity to sound potentially very different depending on the context (Latinus & Belin, 2011; Lavan, Burton, Scott &

McGettigan, 2019). As well as variability in a speaker's voice, external factors such as being in a noisy environment can also present challenges to both voice and speech perception (Smith et al., 2018; Lee, Shim, Yoon, & Lee, 2009). These factors complicate voice recognition and discrimination.

However, as outlined in Chapter 1, familiarity has been found to improve person perception, with various studies finding improved recognition of both faces (Burton, Jenkins, & Schweinberger, 2011; Noyes & Jenkins, 2019) and voices (Latinus & Belin, 2011; Kreiman & Sidtis, 2011; Lavan, Burton, et al., 2019 for a review) when a person is familiar to the viewer/listener. For example, accurate recognition of familiar faces has been observed under various conditions that have been found to disrupt unfamiliar face matching (e.g. changes in lighting, viewpoint, facial expression; Kok, Taubert, Van der Burg, Rhodes, & Alias, 2017). Other benefits to familiarity have included faster detection of familiar faces in a visual search paradigm (Tong & Nakayama, 1999), as well as some evidence to suggest more rapid detection of social cues from familiar faces (Visconti di Oleggio Castello, Guntupalli, Yang, & Gobbini, 2014).

For voices, there are many reports that familiarity improves the ability to comprehend speech, particularly in noisy environments (e.g. Kreitewolf, Mathias, & von Kriegstein, 2017; Johnsrude et al., 2013; Nygaard & Pisoni, 1998). For recognising identity, familiarity has also been found to produce perceptual benefits. A few studies have used voice sorting tasks to explore how well listeners tell voices apart, but also the extent to which people can "tell voices together" (see General Introduction for a summary of voice/face sorting research). Briefly, in these tasks, listeners are presented with a number of voice excerpts and are asked to group together excerpts that they think are produced by the same speakers. Unfamiliar listeners are observed to create many more clusters than there are voices, i.e. perceiving more identities than the veridical number in the set. However, familiar listeners create fewer clusters and are closer to the true number of identities featured (Lavan, Burston, & Garrido, 2019a; Stevenage, Symons, Fletcher, & Coen, 2020). Therefore, unfamiliar listeners in these tasks find it challenging to group together different instances of the same voice as one singular identity, misperceiving within-person variability as between-person variability, whereas familiar listeners are more adept. Results such as these have been interpreted as familiar listeners having built up more robust and refined stored mental representations, enabling them to better manage both between- and within-person variability (Lavan, Burston, & McGettigan, 2019a).

However, whilst familiar voices are recognised/discriminated more accurately when compared to unfamiliar voices, performance of familiar listeners varies depending on the task design, stimuli, and degree of familiarity. The recognition of people by voice has been found to be poorer than the recognition of people by face (Barsics, 2014; Brédart, Barsics, & Hanley, 2009; Hanley, Smith, & Hadfield, 1998), and familiarity does not necessarily guarantee reliable voice identification, particularly when the vocal signal to be recognised is not neutral, conversational speech. For instance, a voice sorting study by Lavan, Burston, and colleagues (2019b) explored the effect of expressiveness on familiar and unfamiliar participants' ability to sort voice excerpts by identity. High-expressiveness voice excerpts included vocalisations such as shouting or speaking in a strained voice. For both the low- and high-expressiveness versions of the task, familiar participants created closer to the true number of identities in the set, whereas unfamiliar participants created many more clusters. However, for the high-expressiveness version, it was found that voice excerpts belonging to different identities became more confusable for familiar and unfamiliar listeners alike. Overall, expressiveness did not have an effect on the overall number of identities perceived, with familiar listeners perceiving closer to the true number in the set in both versions. However, for the high-expressiveness version, both familiar and unfamiliar listeners had more difficulty telling voices apart, mixing excerpts from different speakers into the same perceived identity. Moreover, the ability to group different instances of the same voice together (i.e. telling voices together), was impaired for familiar listeners in the high-expressiveness version of the task. Other studies have also found reductions in recognition ability for familiar and unfamiliar listeners when the task involves generalising across different vocalisation types (Lavan, Scott, & McGettigan, 2016).

Therefore, whilst familiar listeners can capitalise on past experience with a voice to promote stable identity recognition, experiences with different voices can vary and thus the magnitude of familiar voice benefits and the contexts in which these benefits are observed may also vary as a result. Research that illustrates this comes from studies comparing different familiar voices in the same task. Fontaine, Love, and Latinus (2017) compared the recognition of famous and lab-trained voices, and created several speaker averages using voice morphing. Listeners were found to be quicker and more accurate in recognising the identity of the speaker as the number of voice excerpts in the average increased, but this was only observed for famous voices. The recognition of lab-trained voices was not improved by increasing averageness. The prevailing prototype model states that familiar voices are stored in memory as unique reference patterns.

The use of averaging retains the idiosyncrasies of a speaker's voice that make it unique in comparison to other voices, whilst smoothing the variability in vocalisations (Fontaine et al., 2017). Therefore, the improvement in recognition for famous voices was proposed to be due to existing mental representations for these speakers, whereby the increasing number of exemplars in the average corresponded well to these stored reference patterns. Other recent research has also found evidence for average-based representations of individual identities, where representational voice spaces may exist for individual speakers, with that speaker's average/prototype at the centre (Lavan, Knight, & McGettigan, 2019a). For lab-trained voices, on the other hand, stable stored representations do not yet exist, and thus listeners were found to rely more on exemplar-based coding. Exemplar-based coding in voice perception refers to the idea that known voices are stored as exemplars in long-term memory, which is in contrast to norm-based coding which argues for the existence of an abstracted average or prototype developed from prolonged and repeated exposure to different vocalisations from a speaker (Lavan, Burton et al., 2019). For lab-trained voices, this was exemplified by better recognition of previously heard voice excerpts, and thus, averaging is unhelpful for the recognition of these voices. This highlights that the differences in the type and extent of familiarity can affect voice recognition and perception.

Moreover, the shift from exemplar-based coding of unfamiliar/newly-learned voices to stored prototype-based representations for familiar voices is also thought to underpin why familiarity benefits are observed (when compared to low-familiar voice recognition), particularly under perceptually challenging conditions. These stored representations are thought to encode information about how individual voices can sound different in different situations, and thus enable listeners to contend with this variability to maintain accurate recognition (Lavan, Burton et al., 2019). However, the contexts within which familiar voices have been experienced previously may be important, and indeed the previous studies by Lavan and colleagues exploring voice sorting and participants' abilities to make across-vocalisation judgements concluded that recognition/discrimination was impaired for both unfamiliar and familiar listeners alike due to the types of stimuli used. That is, listeners may have experienced highly expressive speech or vocalisations such as laughter less often than neutral conversational speech, and thus stored representations for these voices may have been under-specified for the particular vocalisations used (Lavan, Scott, & McGettigan, 2016).

Little is known about potential differences in the robustness of stored representations within different types of familiar voices, and the effects that this could have on perception. In order to form a robust representation of a speaker's voice, a listener must be able to incorporate information about the ways in which a speaker's voice varies within that individual, in addition to knowing how that speaker's voice differs from other speakers (Lavan, Burton, et al., 2019; Stevenage et al., 2020; Burton, Kramer, Ritchie, & Jenkins, 2016). In the face perception literature, there is some evidence that within-person variability may be partially idiosyncratic, meaning that different faces vary in different ways (Burton, Kramer, Ritchie, & Jenkins, 2016). Whilst much research previously has attempted to identify invariant features of faces or voices that allow for a system for recognising individuals, it has been noted that rather than viewing variability as noise to be to be coped with/filtered out, this variability may instead be diagnostic and aid flexible recognition (Burton et al., 2016). Therefore, even if average-based representations are formed, information about the ways in which a face or voice varies may not be discarded. As an individual becomes increasingly familiar with a voice over time through repeated and varied social interactions, stored representations are refined to include information about both between- and within-person variability. However, despite the multifarious nature of the methods and stimuli used in vocal identity research, familiarity has primarily been studied using either famous voices, or voices trained to be familiar in a lab setting (Kreiman & Sidtis, 2011; McGettigan, 2015). Experience with celebrity or lab-trained voices varies, but is often limited and confined to specific contexts. For example, for lab-trained voices, familiarity is confined to the particular stimuli used during training, and experience with famous voices is acquired via the media. Whereas outside of an experimental setting, naturally acquired familiar voices are predominantly experienced in a wide variety of contexts, including those that are highly social in nature, involving shared memories, knowledge, and experiences. This is something that is usually comparatively lacking or entirely absent for famous or lab-trained voices used in vocal identity research, as recreating the conditions necessary for robust voice representations in a lab setting is difficult. Therefore, we may expect that the most robust representations should exist for those that we are most familiar with (e.g. close friends, romantic partners, family members), and thus the type of familiarity (i.e. lab-trained or famous voices) most frequently used in previous research may have led to an underestimation of the extent of human voice recognition capabilities by overlooking these voices. This is something the experiments in the current chapter aimed to test.

In this chapter, I report three experiments that aimed to examine whether personal familiarity with a voice (using the voices of participants' romantic partners) can afford benefits for voice and speech perception, and how this compares to lab-trained familiarity. Is the recognition of highly familiar voices similarly impaired by perceptual challenges? A discussion is included as to the significance of the findings for the underlying representations of different types of familiar voices. In Experiment 1, a voice identity task was used in which listeners attempted to recognise three speakers from very brief filler sounds (e.g. "umm", "uhh"). Experiment 2 also used a voice recognition task. In this experiment, two acoustic properties (fundamental frequency and formant spacing) were modulated to varying degrees and recognition of the same three voices was tested. In Experiment 3, a speech intelligibility task was used to examine the common finding that familiarity with a voice can facilitate improved intelligibility of speech when heard in a noisy environment. I predict that across all three experiments, a high degree of personal familiarity would result in significantly enhanced voice recognition or speech intelligibility – specific predictions for each experiment are described in the relevant sections below. The study design and analyses were preregistered on the Open Science Framework (https://osf.io/utche).

2.2 General Methods

As the same groups of participants took part in all three experiments in this chapter, I first report information on participant demographics, stimulus materials, the vocal identities used, and the order in which the experiments were completed, in a General Methods section. This will be followed by individual Methods sections for each experiment.

2.2.1.1 Participants

Sixty-four participants in total (32 female, mean age = 27.95 years, SD = 6.50 years, range = 18-40 years) were recruited to take part in the study. Half of these participants were couples in romantic relationships (Sixteen couples: 32 participants, 1 male and 1 female per couple, mean age = 26.31 years, SD = 6.10 years, range = 18-37 years) and the other half were matched control participants (32 controls: 16 female, mean age = 29.22 years, SD = 6.66 years, range = 18-40 years). Couples first visited the lab together, to allow recordings of their voices to be obtained, and then participated in the three experiments via the online testing platform Gorilla.sc (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018). Control

participants did not need to provide any voice recordings, and solely completed the three tasks on Gorilla.sc.

Couples had been in a romantic relationship with their partner for a minimum of six months (mean length of relationship = 63.78 months, SD = 51.49 months, range = 6-204 months) and reported speaking to each other frequently (mean = 34.66 hours per week, range = 4 – 88 hours), thus it was assumed that these participants were highly familiar with their partner's voice. One female participant in the Couples group did not complete the experiments online, leaving a total of 31 participants in this group. Participants that failed a vigilance check (those scoring less than 75% or 6/8 correct on the vigilance trials in each experiment) were excluded per experiment. Feedback from several participants highlighted that the vigilance check for Experiment 3 was confusing, therefore participants that failed the vigilance check in this experiment were not excluded. However, if participants failed the checks in Experiment 3 and another experiment, their data was excluded from both experiments. Participant exclusions in each experiment are reported in the relevant sections below.

2.2.1.1.1 Inclusion of a Control Group:

To ensure that the observed effects were not due to differences in the specific voices used e.g. the couples' voices being systematically more distinctive or memorable than the lab-trained voices, control participants were recruited. Each control group participant was sex-matched to a couples group participant, so that each version of the experiment created for a member of a couple was repeated with a corresponding member of the control group.
Where an individual participant's data were removed from the couples group, the corresponding control group participant's data were also removed, to maintain a one-to-one match between the voice identities presented to the two groups. In order to minimise data loss through participant exclusion, control participants who failed the in-task vigilance checks (see below) were removed and replaced with new control participants until a full set of 31 usable datasets corresponding to the couples group was obtained.

All participants were native English speakers, had normal or corrected-to-normal vision, and reported no hearing difficulties. Couples were all speakers of Standard Southern British English (SSBE) as were the other test voices, such that accent would be controlled across all of the voices used in the studies. Participants were recruited via the UCL Psychology Subject

Pool and social media. On completion of the tasks, participants were compensated at a rate of £7.50/hr of participation. Ethical approval was obtained via the UCL research ethics committee (approval code: SHaPS-2019-CM-030) and informed consent given by all participants.

2.2.1.2 Materials

I obtained voice recordings from the 16 romantic couples (i.e. the 32 participants), 6 adult voices from the freely-available LUCID corpus of speech materials (3 female; Baker & Hazan, 2011), and 2 further adult voices recruited from within the Department of Speech, Hearing and Phonetic Sciences at UCL (1 female). Recordings of the couples were used in the experimental tasks to represent personally-familiar voices (i.e. the romantic partners), while the LUCID identities were used to represent, for each participant, 1 lab-trained identity plus two further unfamiliar identities used in 1) the familiarisation task and 2) Experiments 1 and 2 (3 female, 3 male in total; see Procedure). The two further voices were recruited to obtain recordings of unfamiliar identities reading the coordinate response measure (CRM) sentences (see Read sentences).

2.2.1.2.1 Spontaneous speech

Spontaneous speech was elicited from the couples by asking them to perform the DIAPIX task (Baker & Hazan, 2011). This task involves pairs of participants engaging in an interactive "spot the difference" task: each individual receives only one image in a picture pair, and the aim is to locate all 12 differences between the pictures through discussion of their respective images. In a preliminary session, I recorded each couple discussing a total of three DIAPIX image pairs. The members of the couple were seated in separate sound-attenuating chambers. Each participant wore Beyerdynamic DT297PV headsets fitted with cardioid microphones to enable discussion of their images, and so that I could record their speech (without interference from their partner). Speech was recorded and digitised at a sampling rate of 44100Hz. Both participants were required to click with their mouse at the location of each difference so these could be scored. Each session lasted as long as it took to find all 12 differences, or until a 10-minute timer ended.

Short excerpts (1.5-2s) of fluent and meaningful spontaneous speech, as well as conversational filler sounds (e.g. "um", "mm"), were selected from each member of each couple, as well as from the additional SSBE speakers' DIAPIX recordings (3 female, obtained via the LUCID

corpus; Baker & Hazan, 2011). Fillers were selected on the basis that they were not lexical (e.g. “mmm”, “umm”, “uhuh” would be included; “yeah” or “yep” would not be included). All stimuli were saved as mono WAV files using PRAAT (Boersma & Weenink, 2010), normed for RMS amplitude, and finally converted into mp3 format for use on the online testing platform Gorilla.sc. These stimuli were used for the familiarisation and fillers task (Experiment 1; see *Assignment of voice identities to tasks*).

2.2.1.2.2 Read Sentences

Sentence stimuli included:

- 50 items from the LUCID corpus (e.g. “My brother Paul ran towards the beach.”), produced by the couples (personally-familiar voices) and by 4 LUCID corpus speakers (these served as lab-trained and unfamiliar voices “Anna”/“Adam” and “Clara”/”Charlie”). These sentences were used in the voice modulation task (Experiment 2).
- 50 items from the CRM database (Bolia, Nelson, Ericson, & Simpson, 2000), produced by the couples (personally-familiar voices) and the two recruited novel speakers (see Materials). CRM sentences take the form “Ready [call sign], go to [colour] [number] now.” The call signs used were “Baron”, “Eagle”, and “Laker”, colours were “red”, “green”, “blue”, and “white”, and the numbers were one to eight. These items were used in the speech perception task (Experiment 3).

All newly-recorded items were recorded in a sound-attenuating chamber, using a Røde NT-1A microphone connected to an RME fireface UC audio interface at a sampling rate of 44100Hz. Stimuli were normed for RMS amplitude and converted into mp3 format as required for online testing.

2.2.1.2.3 Assignment of voice identities to tasks

In all the tasks described for the couples group, recordings of each participant’s romantic partner represented the personally-familiar voice, while other, previously-unknown voices were used as lab-trained and unfamiliar identities. To control for basic acoustic cues across the identities, all voices used per participant were of the same sex as the romantic partner. The assignment of these unknown identities to the voice conditions was as follows:

- *Familiarisation of the lab-trained voice:* The participant’s romantic partner represented the personally-familiar voice. One of the LUCID corpus speakers was used as the lab-

trained voice ("Anna" or "Adam"), and one further LUCID speaker of the same sex was used as an unfamiliar identity ("Someone else"; 4 total LUCID speakers (2 female, 2 male) used here).

- *Experiments 1 & 2:* The participant's romantic partner represented the personally-familiar voice. The familiarised LUCID corpus speaker was used as the lab-trained voice ("Anna" or "Adam"), plus a previously-unheard LUCID speaker was introduced as a new identity ("Clara" or "Charlie"; the 2 remaining LUCID speakers (1 female, 1 male) were assigned here).
- *Experiment 3:* The participant's romantic partner represented the personally-familiar voice. A further novel, unfamiliar identity was introduced, using recordings from one of the 2 speakers recruited from UCL Speech, Hearing and Phonetic Sciences (see Materials).

Note that for each participant in the control group, the personally-familiar voice of one couples group member was presented as a second lab-trained identity, labelled either "Beth" or "Ben".

2.2.1.2.4 Vigilance stimuli

A text-to-speech online tool (https://text2speech.us/) was used to generate computerised voices reading "Please press the left key", and "Please press the right key." These were used in vigilance trials (8 per task; 4 of each instruction) to check participants' attention to the auditory stimuli.

2.2.1.3 Procedure

2.2.1.3.1 Online testing session

Approximately 1-2 weeks after recording the stimuli, each of the participants in the couples group completed the perceptual tasks independently (i.e. not in the presence of their partner) on the online testing platform Gorilla.sc (Anwyl-Irvine et al., 2018). A link to a personalised version of the study was sent to participants via email. Participants in the control group were recruited via the online recruitment platform Prolific.co (www.prolific.co) and also completed the tasks on Gorilla.sc. Participants set the volume of the stimuli to a comfortable listening level and were required to pass a headphone screening to ensure that participants were wearing headphones and able to hear the stimuli presented (Woods, Siegel, Traer, & McDermott, 2017).

Each trial of the screening task involves judging which of three tones is the quietest. In each triplet, one tone is presented 180 degrees out of phase across the stereo channels. This makes the task simple with headphones, but difficult without, due to phase cancellation when listening over loudspeakers.

In each of the three main tasks, eight vigilance trials were included to ensure participants were paying sufficient attention to the audio stimuli. These trials required participants to press the left or right arrow keys on their keyboard in accordance with the audio instruction (see vigilance stimuli), instead of clicking a response option with their mouse. Participants that failed to respond correctly at least 75% of the time on these trials were excluded from the relevant task.

2.2.1.4 Familiarisation of the lab-trained voice

In order to directly compare the recognition of a lab-trained voice and one that is personally-familiar, listeners first needed to be trained to recognise a new voice before completing the perceptual tasks. Of the spontaneous speech excerpts extracted from the DIAPIX task recordings, 24 excerpts each were chosen for the personally-familiar voice and the lab-trained voice. For use in a passive exposure phase, these were arranged into two 12-excerpt sequences, with each sound clip separated by 1s of silence. For use in a test phase, a further 20 spontaneous speech stimuli were selected from all three identities (personally-familiar, lab-trained, unfamiliar).

In the familiarisation, participants in the couples group were first passively exposed to the lab-trained voice, as well as re-acquainting themselves with their partner's voice. Text presented on-screen read: "This is Anna" or "This is Adam" (matched to their romantic partner's sex) as well as the instruction to "listen carefully and try to memorise how this voice sounds." Participants listened to the two 12-clip excerpts per voice, and always heard the lab-trained voice first and their partner's voice second. After listening to the two sequences of spontaneous speech from both identities, participants were tested on recognition of the two voices. The 60 test stimuli (20 each from the partner, the lab-trained voice, and an unfamiliar voice) were presented in a fully randomised order. Each trial consisted of a short voice clip, followed by three text response options: "My partner", "Anna(/Adam)", or "Someone else" - responses were made via a mouse-click to select one of these options. Audio-visual feedback

(correct/incorrect) was given on every trial to aid learning of the new voice. This task lasted approximately 5-10 minutes. After this training, listeners were able to recognise the lab-trained voice with good accuracy (80.65% correct, SD = 2.5%, chance = 33%). Control participants performed the same familiarisation task, however the "personally-familiar" voice was introduced as a lab-trained identity labelled either "Beth" or "Ben" for this group. Thus, these listeners learned to recognise two identities: "Anna"/"Adam" and "Beth/Ben". Recognition accuracy after training was also high in this group, for both lab-trained voices ("Beth"/"Ben": mean = 82.58%; mean "Anna"/"Adam" = 81.77%). Both voices were thus recognised with similar accuracy and ease, and at a comparable level to the recognition of "Anna"/"Adam" by the couples.

Following the training, participants either performed voice identity recognition from non-verbal vocalisations (Experiment 1) or voice identity recognition from acoustically modulated voices (Experiment 2) first. The order of Experiments 1 and 2 was counterbalanced across participants. Before the start of the first experiment, listeners were introduced to a novel and thus unfamiliar voice "Clara"/"Charlie" and presented with one example speech token from this speaker - this was their only exposure to this speaker before the task began. Note this was a different unfamiliar talker from the one used in the familiarisation.

The speech perception task (Experiment 3) was always completed last. For this task, a final unfamiliar talker was used but was not introduced to the participant, by name or otherwise. Participants did not receive any feedback on their performance during Experiments 1-3, and all experiments were completed within the single online testing session.

2.3 Experiment 1: Voice identity recognition from non-verbal vocalisations

2.3.1 Introduction

In this experiment, identity recognition from vocal stimuli that contained minimal linguistic cues (i.e. conversational filler sounds such as "uhh", "umm") was examined. Perhaps unsurprisingly, perceiving identity from voice excerpts that are short in duration and lack meaningful linguistic information is more challenging than recognising a speaker from longer excerpts involving linguistic content (Schweinberger, Herholz, & Sommer, 1997; Bricker &

Pruzansky, 1966). However, the limited existing research has established an advantage of knowing the talker on vocal identity perception, even under less than optimal conditions. For instance, research using naturally varying non-verbal vocalisations such as laughter, cries, and coughs has provided some evidence for such familiarity benefits. A study by Zarate, Tian, Woods, and Poeppel (2015) observed above-chance recognition of 5 voices from non-speech vocalisations (e.g. cries, grunts, coughs, laughs) after brief familiarisation training also using non-speech vocalisations. Additionally, Lavan, Scott, and McGettigan (2016) used paired combinations of vowels and spontaneous/volitional laughter, and participants who were either familiar or unfamiliar with the voices performed a speaker discrimination task. The authors found that familiar listeners demonstrated an enhanced ability to discriminate between pairs of non-verbal vocalisations produced by their university lecturers compared to unfamiliar listeners who did not know the lecturers. However, it is important to note that accuracy in both of these studies was low, and the study by Zarate and colleagues included a range of conditions, including sentences, finding that recognition from non-verbal vocalisations was the poorest. Lavan and colleagues (2016) proposed that the overall poor performance across both familiar and unfamiliar listeners may have been due to a lack of familiarity with the particular types of vocalisations used in their study. For example, familiar listeners may not have had much experience with their lecturers' laughter and thus this may have impacted recognition negatively (Lavan, Burston, et al., 2019a).

Thus, while familiarity has been shown to produce benefits for voice recognition, differences in the extent or content of prior experience can render identity perception fallible under certain conditions. In order to have a robust stored representation of a voice, a listener may need to have experience with the full range of vocalisations a speaker is capable of producing, in a wide variety of contexts (Lavan, Burton, et al., 2019). Accordingly, for individuals with whom we are personally-familiar (e.g. romantic partners, as in the current experiments), costs to performance should be reduced compared to lab-trained voices because stored representations should be built from more comprehensive exposure to the speaker's vocal repertoire. Therefore, using non-verbal filler sounds as representative of vocalisations with minimal linguistic cues, I aimed to test this prediction.

2.3.2 Methods

2.3.2.1 Stimuli

20 filler sounds (mean duration = 0.59s) were extracted from the DIAPIX task recordings per identity (personally-familiar voice [lab-trained "Beth"/"Ben" for controls], lab-trained voice "Anna"/"Adam", and the unfamiliar voice "Clara"/"Charlie") for this task, as well as 8 vigilance stimuli. The personally-familiar voice was always the romantic partner of one participant from the couples group. The lab-trained and unfamiliar voices were the same for all couples and control participants (where female participants heard male voice identities, and *vice versa*). Examples of stimuli used in this experiment can be found at: https://osf.io/g2jk6/.

2.3.2.2 Procedure

In this experiment, participants heard a total of 60 filler sounds produced by the three speakers (personally-familiar, lab-trained, unfamiliar) and eight vigilance trials in a randomised order. On each trial, a filler sound was presented, followed by a prompt asking participants to select the identity they thought had produced it from three response options ("My partner", "Anna"/"Adam", "Clara"/"Charlie") via mouse-click. For control group participants, the three response options were "Beth"/"Ben", "Anna"/"Adam", and "Clara"/"Charlie". Vigilance trials required participants to respond with a keypress (left or right arrow key) instead of selecting a text response option with their mouse. This task lasted approximately 5 minutes.

2.3.2.3 Data Analysis

Unbiased hit rates (H_u scores) were calculated for each of the three familiarity conditions (personally-familiar, lab-trained, unfamiliar) to correct for any disproportionate usage of certain response categories (Wagner, 1993). Calculating H_u scores involves multiplying the conditional probability that a stimulus is correctly detected given that it is present (i.e. the number of hits / the true number of stimuli in that category) by the conditional probability that the stimulus is correctly detected divided by the total number of uses of that stimulus category (i.e. the number of hits / the total number of times that response category was chosen).

Taking personally-familiar voice trials as an example case:

$$\frac{Hits('My\ Partner')}{True\ N\ ('My\ Partner'\ trials)} \ x \ \frac{Hits('My\ Partner')}{Total\ N\ ('My\ Partner'\ responses)}$$

H_u scores were arcsine transformed (Wagner, 1993). Data were analysed using linear mixed models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For the LMMs, model estimates and associated confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero are significant. Following my pre-registered analysis plan and to keep the statistical models as simple as possible while still being able to address the research question, I have not included group as a factor in the analyses for any experiment. Including group as a factor would have introduced higher-order interactions for all analyses, making the reported effects less easy to understand. I therefore analyse and report the findings of the couples and controls separately.

2.3.3 Results

Data from 4 couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 27 participants per group were retained for the statistical analyses.

2.3.3.1 Couples

To assess the impact of the three types of familiarity (personally-familiar; lab-trained; unfamiliar) on voice identity recognition performance based on the non-verbal filler sounds, an LMM was run with H_u scores for recognition performance as the outcome variable. In this confirmatory analysis, familiarity was entered into the model as a fixed effect, and random intercepts of participant and voice identity were added as random factors. Statistical significance was established via likelihood ratio tests comparing the full model that contained all fixed and random effects to a reduced model where the relevant effect had been dropped.

Familiarity had a significant effect on voice identity recognition ($\chi^2(2) = 20.33$, $p < .0001$), with post-hoc comparisons (via the *emmeans* package in R) indicating that listeners were significantly better at recognising their partner's voice (raw mean = 91.3%, SD = 9.7%) compared to the lab-trained ($p = .001$; raw mean = 64.4%, SD = 13.8%, E = -0.60 , CI = [-0.93, -0.28]) and unfamiliar identities ($p < .001$; raw mean = 47.2%, SD = 16.7, E = -0.69, CI = [-1.01, -0.37]; see Figure 1a). Figure 2a illustrates responses as a confusion matrix – this shows

both a high hit rate and low false alarm rate for the personally-familiar voice, while the lab-trained and unfamiliar voices were more frequently confused with one another.

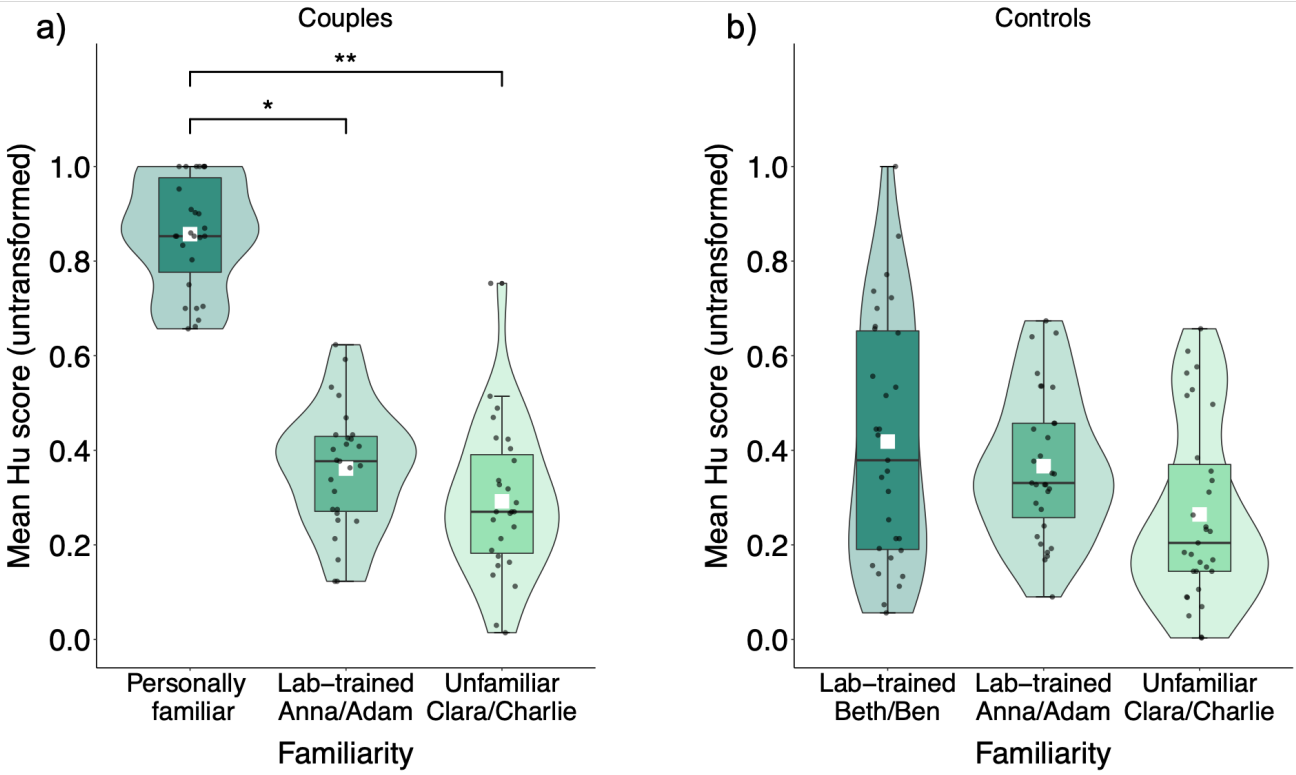

Figure 1. Box plots display median Hu scores (unbiased hit rates) for each of the three speakers in the fillers task (Experiment 1) for (a) the Couples group and (b) the Control group. The boxes range from the first to third quartiles. Whiskers extend to no further than 1.5*interquartile range above and below the 1st and 3rd quartiles. The lighter shaded violin portion of the plots display the probability density of the data, allowing us to visualise the distribution of the data. Wider parts represent a higher probability of data existing at those values, and thinner parts reflect a lower probability of data taking on those values. Points represent individual participants' Hu scores for each speaker identity. White squares display the group means per condition. ** $p < .001$, * $p = .001$.

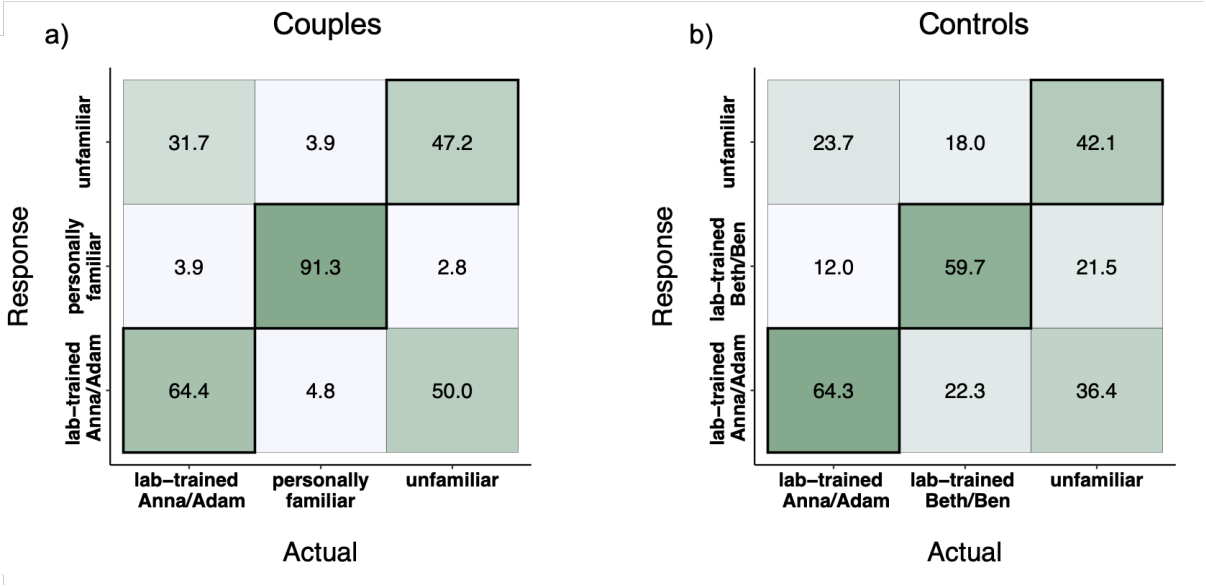

Figure 2. Confusion matrix displaying a) the Couples group and b) the Control group's responses per condition for the recognition of voice identity from non-verbal vocalisations (Experiment 1). Each cell shows the percentage of trials in which a presented voice ("Actual") was perceived as one of the three target identities ("Response"). Cells on the diagonal (indicated by a darker border) reflect correct responses (hits); darker greens indicate higher percentages.

2.3.3.2 Controls

If the observed results for the couples group were due to relative familiarity of the couples with the personally-familiar and lab-trained voices, and not due to systematic differences in distinctiveness or recognisability of these voices *per se*, there should be no significant differences in control participants' performance for these two identities in both vocal identity tasks (i.e. recognising identity from non-verbal filler sounds, and from modulated sentences).

To assess the impact of the three voice identities on recognition accuracy, an LMM was run with the same fixed and random effects, and model comparison, as reported for the couples group. Statistical significance was again established via likelihood ratio tests comparing the full model that contained all fixed and random effects, to a reduced model that did not include familiarity. Note that familiarity was still defined with 3 levels, corresponding to lab-trained "Beth"/"Ben" (i.e. personally-familiar for couples), lab-trained "Anna"/"Adam", and the unfamiliar voice, respectively.

Comparing the full model to the reduced model revealed no significant differences in performance between the three identities (two lab-trained (Beth/Ben: E = 0.70, CI = [0.59, 0.80]; Anna/Adam: E = -0.07, CI = [-0.36, 0.22]) and one unfamiliar (E = -0.22, CI = [-0.50, 0.07]) voice; $\chi^2(2) = 2.45$, p = .294; See Figure 1b). This shows that there was no overall difference in distinctiveness between the two lab-trained voices. This analysis therefore shows that the effects observed for the couples group are a result of the familiarity with the personally-familiar partner's voice, and not artefacts of the stimuli used in this task.

Raw recognition accuracy for the two lab-trained voices was 64.3% (Anna/Adam; SD = 15.2%) and 59.7% (Beth/Ben; SD = 23.6%), and 42.1% for the unfamiliar voice (SD = 19.4%); however, there were frequent categorisation errors (see Figure 2b).

2.3.4 Discussion

Above chance performance was observed for participants in both the Couples and Controls groups (see Appendix A), however, there were key distinctions between the groups in terms of how well the voices were identified. Couples participants were extremely accurate in recognising their romantic partner's voice from non-verbal filler sounds (raw accuracy = 91.3%), whereas frequent confusions were made in this group when attempting to recognise the lab-trained and unfamiliar voices (see Figure 2a). Control participants confused all three voices, suggesting that the romantic partner voices were not more easily recognisable (e.g. more distinctive) but that the differences in observed accuracy were due to differences in the degree of familiarity with these voices (see Figure 2b). Performance for the lab-trained voices was more consistent with previous studies exploring voice recognition from brief, non-verbal vocalisations, that found performance that was above chance (33.4% accurate; chance = 20%) but highly error-prone (Zarate et al., 2015).

The results from Experiment 1 therefore illustrate that the type of familiarity one has with a voice can affect the extent to which it can be recognised. That is, listeners excelled at recognising a personally familiar voice (their romantic partner) in a task where the available cues to recognition were greatly reduced. However, these same listeners experienced considerable difficulty in perceiving identity from lab-trained voices. Although recognition

was overall weaker for lab-trained voices, brief exposure through training appears to be sufficient for listeners to "get by" in distinguishing voices from each other with above chance accuracy, but this ability is relatively under-developed and susceptible to interference. This may give insight into the nature of the representations that exist for these voices. That is, limited experience with lab-trained voices may mean that their associated representations are not robust, which in turn constrains/inhibits the extent to which these voices can be recognised accurately, particularly under conditions where the available cues to recognition are reduced (Fontaine, Love, & Latinus, 2017). In contrast, protracted and varied experience with personally familiar voices results in representations are robust and this promotes stable, and highly accurate recognition.

2.4 Experiment 2: Voice identity recognition in the context of acoustic modulation

2.4.1 Introduction

Experiment 1 highlighted that personally familiar voices were recognised with higher accuracy across naturally-produced non-verbal utterances, due to the presence of a more robust perceptual representation of that voice. We can also ask questions about what is contained within these representations by manipulating voice acoustics and exploring the effects of this on vocal identity perception. Thus, Experiment 2 used short sentences in which two acoustic cues had been altered to varying degrees. In this way, Experiments 1 and 2 taken together could probe the robustness of voice representations for recognition, as well as gain an insight into how listeners' stored representations are tuned to the manipulated acoustic cues, and how differences in familiarity may affect this tuning.

As the acoustic properties associated with individual speakers vary from person to person, and are partly constrained by individual vocal anatomy, it is only logical that previous investigations into vocal identity have attempted to determine which acoustic cues may be important for recognition. For example, one study suggested that glottal pulse rate (GPR, related to the fundamental frequency) and vocal tract cues are highly relevant for vocal identity perception, particularly as they alter the perceived size and sex of a speaker (Smith & Patterson, 2005). Moreover, GPR and vocal tract length (VTL) have been found to be perceptually salient cues for recognising a speaker. Gaudrain, Li, Ban, and Patterson (2009) explored the extent to

which these features could be altered until listeners no longer perceived two voice samples as being produced by the same unfamiliar speaker. The authors found that listeners were more sensitive to changes in VTL, as this cue could be modulated to a smaller degree compared to GPR before listeners perceived excerpts as two different identities.

Another study by Lavner, Gath, and Rosenhouse (2000) tested listeners' ability to recognise personally familiar voices (members of a Kibbutz in which the participants lived) producing vowel sounds. There were 20 voices to be identified and participants were given a list of 29 names to choose from (9 of which were not actually recorded). Of the voices correctly identified, participants were presented with acoustically modulated versions of these vowels. Modifications included shifting individual formants, and altering fundamental frequency, amongst other acoustic modifications. Modulation of vocal tract properties (i.e. formant frequencies) were identified as being most disruptive for recognition, although different combinations and weightings of acoustic features were diagnostic for different individual voice identities (Lavner, Gath, and Rosenhouse, 2000).

Exploring how the modulation of specific acoustic features affects vocal identity perception can provide information as to the relative importance of various acoustic cues to recognition. It can also further illuminate the nature and robustness of underlying representations for personally familiar voices. In the current experiment, I predicted that acoustic modulation would have a differential effect on voice recognition for personally familiar and lab-trained voices. However, predicting the direction of the expected effect is less clear: On one hand, increased knowledge of one's partner's voice might allow a listener to accept larger modulations of voice acoustics without a cost to recognition. On the other hand, more in-depth knowledge of a speaker's vocal repertoire might reduce the range of acoustic properties that would be accepted as belonging to that personally familiar voice compared to a lab-trained one.

2.4.2 Methods

2.4.2.1 Stimuli

This experiment used 50 read sentences extracted from the LUCID corpus materials, produced by the same identities as used in Experiment 1. Sentences were acoustically modulated with STRAIGHT (Kawahara & Irino, 2004) in the MATLAB environment (see Gaudrain, 2018) to simultaneously introduce changes in glottal pulse rate and vocal tract length (by modulating

F0 and formant spacing) in semitones (a semitone is a twelfth of an octave). STRAIGHT works by decomposing the speech signal into three components: the F0 contour, an aperiodicity matrix, and the spectral envelope. To apply changes to F0, the F0 vector is modified. To effect changes in VTL, the spectral envelope (and the aperiodicity matrix) are rescaled along the frequency axis. Compressing results in an increase in VTL, and expanding effects a reduction in VTL. As a result, all centre frequencies are shifted, as are their widths, proportionally (Kawahara & Irino, 2004). F0 was altered by two or four semitones in either direction, and VTL by one or two semitones, so that with every upward semitone shift in VTL, there was an accompanying two-semitone downward shift in F0, and vice versa (Gaudrain et al., 2009; see Figure 3a). The overall effect of the combined modulations was to create voices that sounded relatively more masculinised (i.e. lower pitch and longer vocal tract) and feminised (i.e. higher pitch and shorter vocal tract) than the original voice. Examples of each of the modulation steps, from 1 male and 1 female speaker, are publicly available on the open science framework (OSF) and can be accessed at: https://osf.io/g2jk6/. Once processed with STRAIGHT, 12 stimuli were randomly selected for each step for both the personally-familiar ("Beth"/"Ben") voice and the lab-trained ("Anna"/"Adam") voice – as there were only 50 recorded sentences available, two randomly selected items from each modulation step and from the unshifted voice recordings were repeated once each during the task. Six tokens per step were selected for the unfamiliar voice.

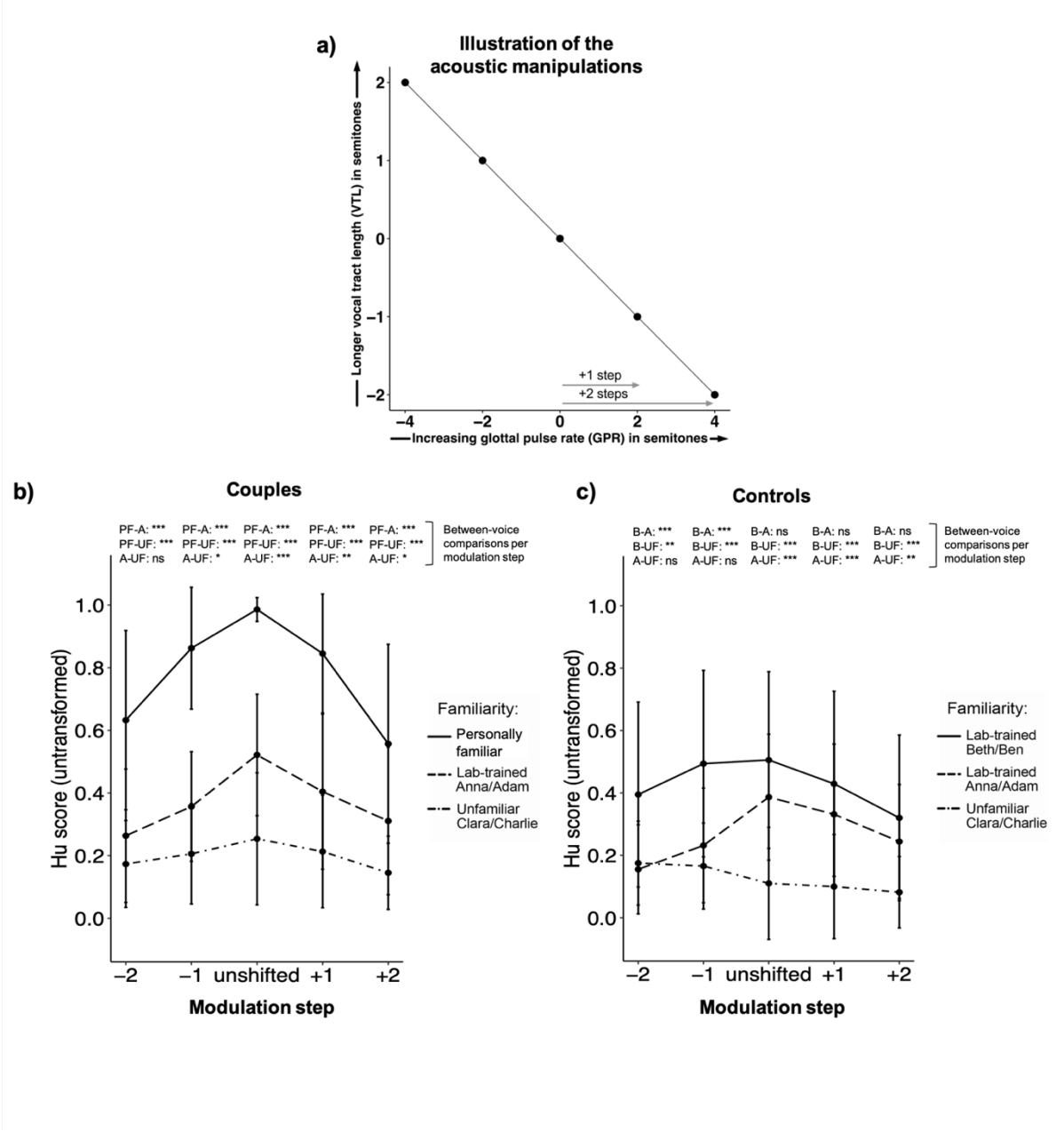

Figure 3. a) Acoustic manipulations made to the voices in Experiment 2. Points represent the five modulation steps used, plotted as combined shifts in glottal pulse rate (GPR) and vocal tract length (VTL) relative to the original voice recordings (i.e. 0,0). Increases in GPR (in semitones) correspond to sounds with higher subjective pitch. For vocal tract length, a positive shift in VTL (in semitones) gives the percept of a longer vocal tract. Orange (lighter) arrows show how the acoustic manipulations corresponded to the modulation "steps" described in the analyses. b) and c) Mean Hu scores are displayed per familiarity condition (Personally-familiar/Lab-trained "Beth"/"Ben", Lab-trained "Anna"/"Adam", unfamiliar) and modulation step (x-axis) for couples (left) and controls (right). Error bars display one standard deviation around the mean. Asterisks denote significance of between-voice comparisons at each

modulation step; PF = personally-familiar, A = Lab-trained "Anna"/"Adam", B = Lab-trained "Beth"/"Ben", UF = unfamiliar; *** p < .0001, ** p < .001, * p < .01, ns = not significant.

2.4.2.2 Procedure

In this task, participants were presented with the 150 modulated and unmodulated stimuli (60 each for the personally-familiar and lab-trained voices, 30 for the unfamiliar voice) and eight vigilance trials, in a fully randomised order. Immediately prior to performing the task, participants were told that they would hear manipulated and original versions of the three voices (Partner, Anna/Adam, Clara/Charlie) and were instructed to decide on each trial who was speaking, regardless of any modifications made to the voices. On each trial, a sentence was presented, followed by a prompt asking participants to select the speaker they thought they had heard from three text response options ("My partner", "Anna"/"Adam", "Clara"/"Charlie") via mouse-click. For controls, the three response options were "Beth"/"Ben", "Anna"/"Adam", and "Clara"/"Charlie". Vigilance trials required participants to follow an instruction to respond with a keypress ("please press the left/right arrow key"), instead of selecting a text response option with their mouse. The task took approximately 15 minutes to complete.

2.4.2.3 Data Analysis

Unbiased hit rates (H_u scores) were calculated for each of the three familiarity conditions (personally-familiar, lab-trained, unfamiliar) to correct for any disproportionate usage of certain response categories (Wagner, 1993). H_u scores were arcsine transformed (Wagner, 1993). Data were analysed using linear mixed models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For the LMMs, model estimates (E) and associated confidence intervals (CIs) are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero indicate significant effects. Following my pre-registered analysis plan, I analyse and report the findings of the couples and controls separately.

2.4.3 Results

Data from 2 couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 29 participants per group were retained for the statistical analyses.

2.4.3.1 Couples

Averaging across all modulation steps per speaker identity showed that the mean overall performance for the personally-familiar voice was 79.5% (SD= 14.0%), with mean scores on the individual modulation steps ranging over 58.6% - 98.9%. Mean overall performance for the lab-trained voice was 56.0% (SD= 10.0%) with mean scores on individual steps ranging over 43.9% - 71.2%. For the unfamiliar voice, mean overall performance was 50.4% (SD= 11.1%), ranging over 45.4% - 56.9% across the individual modulation steps.

To evaluate the effect of the acoustic modulations on recognition of the three identities, I analysed the interaction between degree of modulation (i.e. modulation "step"), and familiarity using LMMs. In this confirmatory analysis, the outcome measure was the H_u score for recognition performance; familiarity and degree of modulation were included as fixed effects, including the interaction between familiarity and degree of modulation. Participant and speaker identity were included as random effects. However, after accounting for the variance explained by participants, speaker identity did not explain any additional variance and was thus removed from the models. Statistical significance was again established by comparing the full model including the interaction, fixed, and random effect to a reduced model that included all of the same fixed and random effects, but did not include the interaction.

Comparing the full model to the reduced model indicated a significant interaction between familiarity and the degree of modulation ($\chi^2(8) = 40.68$, $p < .0001$; see Figure 3b and Table 1 for the full model output).

Post-hoc pairwise comparisons (using *emmeans*) were run to assess the effect of increasing the degree of modulation on recognition of the three identities, and FDR-corrected for multiple comparisons. Performance for the personally-familiar voices was negatively affected by each additional step in both directions (unshifted vs -1 step: t ratio = 3.16, p = .0008; -1 step vs. -2 steps: t ratio = -4.70, p < .0001; unshifted vs. +1 step: t ratio = 4.01, p = .0001; +1 step vs. +2 steps: t ratio = 5.71, p < .0001). For the lab-trained voice, acoustic modulation produced a significant decrease in performance for two comparisons (unshifted vs. one step shift in both directions; negative (-1 step): t ratio = 2.65, p = .012; positive (+1 step): t ratio = 2.11, p = .046). For the unfamiliar voice condition, acoustic modulation did not produce a significant

difference in performance relative to the original voice. These results suggest that acoustic manipulations had a bigger effect on performance for the personally-familiar voice identity than for the lab-trained and unfamiliar identities. However, it should be noted that this effect is in part due to performance being overall much better for personally-familiar voices, such that there was also greater scope for performance to decrease.

Table 1. Model estimates and confidence intervals (CIs) from the full model for the couples group.[1]

		Hu score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.53	1.43 – 1.63	**<0.001**
Familiarity [Lab-trained Anna/Adam]	-0.72	-0.85 – -0.59	**<0.001**
Familiarity [Unfamiliar]	-1.04	-1.17 – -0.91	**<0.001**
Modulation step [-2]	-0.56	-0.70 – -0.43	**<0.001**
Modulation step [-1]	-0.24	-0.37 – -0.11	**<0.001**
Modulation step [+1]	-0.27	-0.41 – -0.14	**<0.001**
Modulation step [+2]	-0.67	-0.80 – -0.53	**<0.001**
Familiarity [Lab-trained Anna/Adam] * Modulation step [-2]	0.24	0.05 – 0.43	**0.012**
Familiarity [Unfamiliar] * Modulation step [-2]	0.45	0.26 – 0.64	**<0.001**
Familiarity [Lab-trained Anna/Adam] * Modulation step [-1]	0.06	-0.13 – 0.25	0.533
Familiarity [Unfamiliar] * Modulation step [-1]	0.18	-0.01 – 0.36	0.065
Familiarity [Lab-trained Anna/Adam] * Modulation step [+1]	0.13	-0.06 – 0.32	0.173
Familiarity [Unfamiliar] * Modulation step [+1]	0.21	0.02 – 0.39	**0.031**
Familiarity [Lab-trained Anna/Adam] * Modulation step [+2]	0.38	0.20 – 0.57	**<0.001**
Familiarity [Unfamiliar] * Modulation step [+2]	0.51	0.32 – 0.69	**<0.001**

[1] The reference categories are the 'personally familiar' voice for Familiarity, and the 'unshifted' condition for Modulation Step.

Next, I explored the effects of familiarity, via FDR-corrected pairwise comparisons at each modulation step (see Figure 3b). At all but one modulation step, performance was significantly different depending on familiarity with the speaker (personally-familiar > lab-trained, lab-trained > unfamiliar, personally-familiar > unfamiliar). For the most masculinised condition (i.e. step -2), there was no significant difference between the lab-trained voice and the unfamiliar voice (t ratio = 1.58, p = .141). Differences in recognition accuracy between the personally-familiar voice and the two other conditions were smaller at the largest modulation steps (i.e. -2 and +2) compared to the unshifted condition (lab-trained (step -2): E = 0.240, CI = [0.05, 0.43], (step +2): E = 0.383, CI = [0.20, 0.57], unfamiliar (step -2): E = 0.452, CI = [0.26, 0.64], (step +2): E = 0.507, CI = [0.32, 0.69]), again suggesting that acoustic manipulations had a larger effect on personally-familiar voice recognition. Table 1 also displays the differences in recognition accuracy between the personally-familiar voice and the two other conditions for a shift of one modulation step in either direction compared to the unshifted condition. Only the difference between the personally familiar voice and unfamiliar condition was significantly larger for the unshifted condition compared to a shift of one modulation step in the positive direction (E = 0.21, CI = [0.02, 0.39]).

Confusion matrices displaying the group averages of raw responses for each trial were constructed (collapsed across direction of acoustic modulation) to examine the types of categorisation errors made by listeners (see Figure 4) – these show that increasing distance from the original voice led to decreases in hits (i.e. labelling the partner as the partner) and increases in misses (i.e. labelling the partner as another identity) while false alarms (i.e. labelling another identity as the partner) remained very low and stable across conditions.

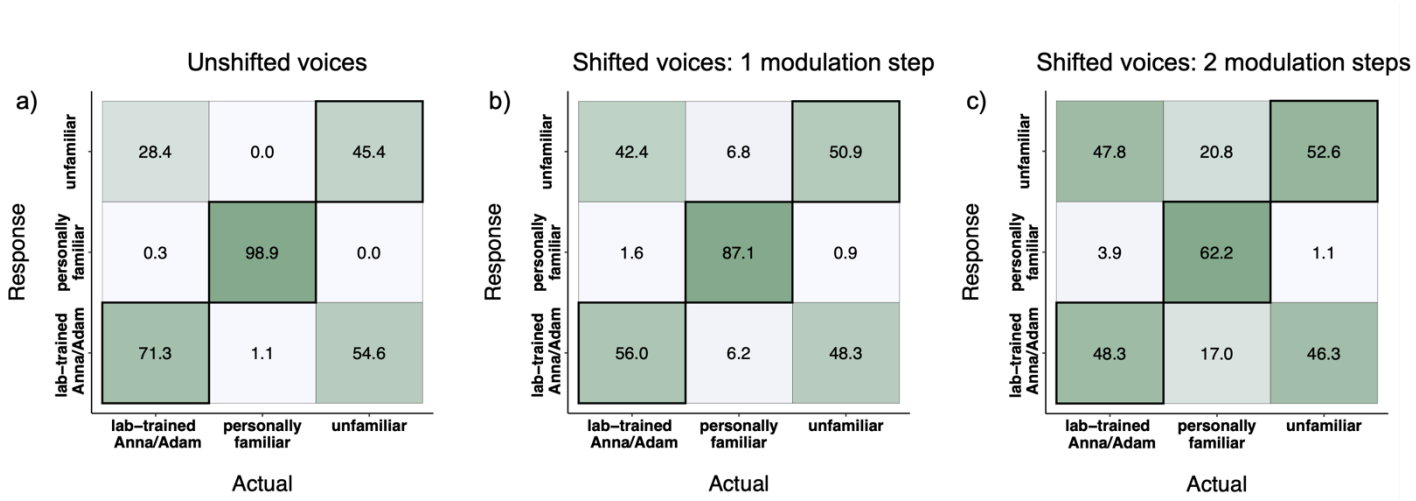

Figure 4. Confusion matrices displaying the couples group's responses in the modulation task (Experiment 2). Matrices are shown for each modulation step: a) Unshifted condition: participants' raw responses to the speaker's "original" voices; b) 1 modulation step: displays hits, misses, and false alarms for the three identities when these voices had been modulated by one step (collapsed across direction of acoustic modulation); c) 2 modulation steps: displays hits, misses, and false alarms for the three identities modulated by 2 steps (collapsed across direction of acoustic modulation).

2.4.3.2 Controls

Averaging across all modulation steps per speaker identity showed that the mean overall performance for the lab-trained "Beth"/"Ben" voice (corresponding to the romantic partners of the couples group) was 55.8% (SD = 20.5%), with mean scores on the individual modulation steps ranging over 44.3% - 65.2%. Mean overall performance for the lab-trained "Anna"/"Adam" voice was 47.3% (SD =11.6%) with mean scores on individual steps ranging over 25.9% - 64.7%. Lastly, for the unfamiliar voice, mean overall performance was 37.8% (SD = 13%), ranging over 27% - 52.9% across the individual modulation steps.

To assess the effect of acoustic modulation on recognition of the three identities, the interaction between modulation step and familiarity was analysed using LMMs as described for the couples group above. Comparing the full model to a reduced model that did not contain an interaction, I found a significant interaction between modulation step and familiarity ($\chi^2(8) = 32.49$, $p < .0001$; see Table 2 for full model output). As in the couples group, I assessed both the effect of modulation step on performance within each identity, and differences between

familiarity conditions (lab-trained voices and unfamiliar) within each modulation step. Post-hoc pairwise comparisons (using *emmeans*) were first run comparing performance between successive modulation steps (e.g. -2 steps vs. -1 step, -1 step vs. unshifted condition) for each identity separately, and FDR-corrected for multiple comparisons. The results showed that modulation step did not have an effect on performance for all three identities except for one comparison: for lab-trained "Anna"/"Adam", a shift of one step in the negative direction resulted in significantly lower performance than performance for the original unshifted "Anna"/"Adam" voice (t ratio = 3.10, p = .004).

Next, performance was compared for the three identities (lab-trained "Anna"/"Adam", lab-trained "Beth"/"Ben", unfamiliar), using FDR-corrected pairwise comparisons at each modulation step. Significantly better performance was observed for lab-trained "Beth"/"Ben" compared to lab-trained "Anna"/"Adam" for voice tokens shifted by 1 and 2 steps in the negative direction (see Figure 3c). Performance was also significantly better for lab-trained "Beth"/"Ben" compared to the unfamiliar voice at all modulation steps. Performance for lab-trained "Anna"/"Adam" voice was better than the unfamiliar voice for the unshifted condition, and for tokens shifted in the positive direction.

Confusion matrices displaying the group averages of raw responses for each trial were constructed to examine the types of categorisation errors made by controls (see Figure 5) – these show that performance was relatively similar across all modulation steps.

Table 2. Model estimates and confidence intervals (CIs) from the full model for the control group.[1]

		Hu score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.80	0.69 – 0.90	**<0.001**
Familiarity [Lab-trained Anna/Adam]	-0.14	-0.28 – -0.01	**0.038**
Familiarity [Unfamiliar]	-0.53	-0.66 – -0.39	**<0.001**
Modulation step [-2]	-0.16	-0.29 – -0.02	**0.023**
Modulation step [-1]	-0.02	-0.16 – 0.11	0.746
Modulation step [+1]	-0.09	-0.23 – 0.04	0.178
Modulation step [+2]	-0.23	-0.36 – -0.10	**0.001**

Familiarity [Lab-trained Anna/Adam] * Modulation step [-2]	-0.15	-0.34 – 0.04	0.118
Familiarity [Unfamiliar] * Modulation step [-2]	0.28	0.09 – 0.47	**0.004**
Familiarity [Lab-trained Anna/Adam] * Modulation step [-1]	-0.19	-0.38 – -0.00	**0.045**
Familiarity [Unfamiliar] * Modulation step [-1]	0.13	-0.06 – 0.32	0.182
Familiarity [Lab-trained Anna/Adam] * Modulation step [+1]	0.02	-0.17 – 0.21	0.817
Familiarity [Unfamiliar] * Modulation step [+1]	0.05	-0.14 – 0.24	0.577
Familiarity [Lab-trained Anna/Adam] * Modulation step [+2]	0.05	-0.14 – 0.24	0.581
Familiarity [Unfamiliar] * Modulation step [+2]	0.17	-0.02 – 0.36	0.087

[1] The reference categories are "Lab-trained Beth/Ben" for Familiarity, and the "unshifted" condition for Modulation step.

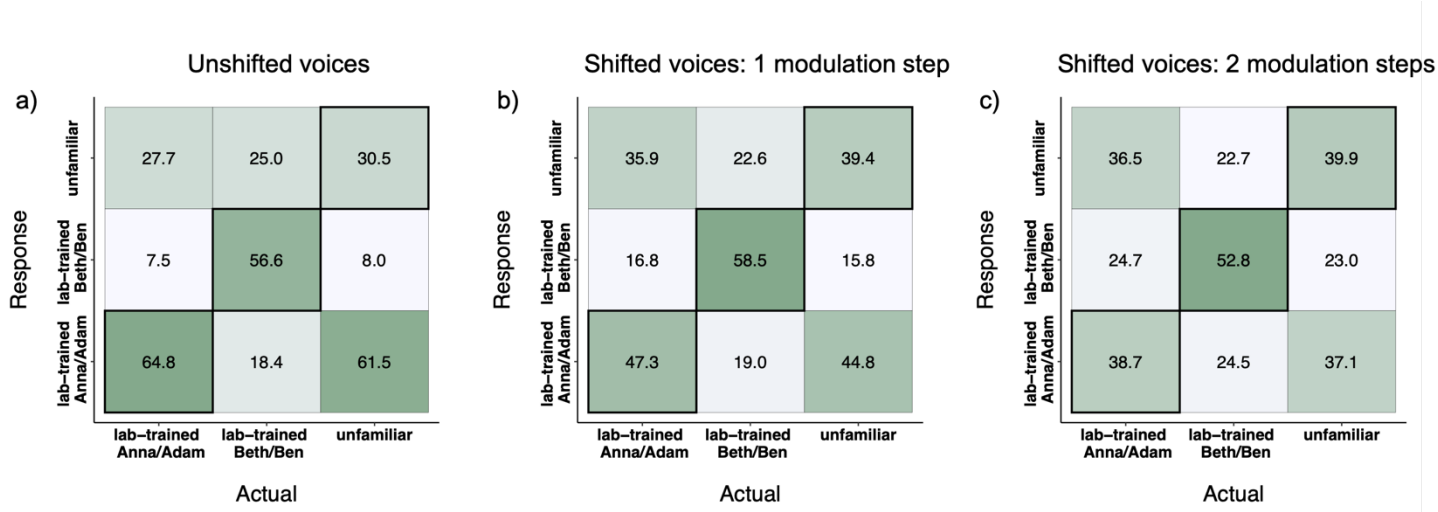

Figure 5. Confusion matrices displaying the control group's responses in the modulation task (Experiment 2). Matrices are shown for each modulation step: a) Unshifted condition: participants' raw responses to the speaker's "original" voices; b) 1 modulation step: displays hits, misses, and false alarms for the three identities when these voices had been modulated by one step (collapsed across direction of acoustic modulation); c) 2 modulation steps:

displays hits, misses, and false alarms for the three identities modulated by 2 steps (collapsed across direction of acoustic modulation).

2.4.4 Discussion

Experiment 2 explored the effect of acoustic modulation on voice recognition, comparing accuracy for personally familiar to lab-trained voices. Similarly to Experiment 1, Couples group listeners excelled at recognising their romantic partner's voice, despite acoustic manipulation. Performance for recognising the personally familiar voice was significantly better than that observed for lab-trained voices. Again, personal familiarity is associated with a more comprehensive knowledge of the speakers' vocal inventory, which facilitates recognition that is highly accurate in comparison to a lab-trained voice. It has been argued that when salient cues to recognition are absent or modified (e.g. GPR/VTL), we may be able to rely on alternative cues such as a person's speech rate or accent information in order to maintain accurate recognition of personally familiar voices (Maguinness, Roswandowitz, & von Kriegstein, 2018).

However, despite personally familiar voice recognition that was overall superior to lab-trained and unfamiliar voices, acoustic modulation did have an impact on recognition of all three voices. Notably, there was a steep, symmetrical drop in accuracy for personally familiar voices and a sharp "tuning function" as the extent of the modulation increased. A similar pattern was observed for lab-trained voices but this was considerably flatter. This may suggest that listeners were in fact more sensitive to deviations from the expected acoustic properties for personally familiar voices.

It should be noted, however, that a sharper tuning function may not necessarily signify a more detrimental effect of acoustic manipulation for recognising personally familiar voices. The sharper 'tuning' observed for personally familiar voices may reflect that there is greater scope for performance to decline as recognition for the unmodulated condition was extremely accurate for these voices (raw accuracy = 98.9%). Nonetheless, looking at the nature of the errors made by constructing confusion matrices revealed a unique pattern of response bias for personally familiar voices. In this condition, false alarms were low. That is, the lab-trained and unfamiliar voice excerpts were very rarely appointed as the romantic partner across all

modulation steps (5% of "personally familiar" responses for the largest deviations). In actuality, the decrease in performance with increasing modulation was due to a greater number of incorrect rejections of the personally familiar voice, such that participants increasingly attributed their partner's voice tokens as belonging to the lab-trained or unfamiliar voice. These 'incorrect rejection' errors became more frequent as acoustic deviance increased. In comparison, the lab-trained and unfamiliar conditions involved many mutual confusions and the magnitude of errors increased with increasing acoustic modulation. Therefore, listeners may have more robust representations of their partner's voice and be more sensitive to the dynamics of their vocal system (Lavan, Scott, & McGettigan, 2016; Fontaine, Love, & Latinus, 2017). Therefore, an acoustically manipulated voice that is pushed beyond what is anatomically achievable may be less likely to be perceived as familiar due to an incompatibility between the manipulated signal and the listener's stored representations for that particular voice. A similar pattern of results has been observed in a recent face morphing study (Chauhan & Gobbini, 2020). This experiment involved morphing two faces along a continuum, and participants had to decide which of the two faces they perceived. When two unfamiliar faces are morphed, at the midpoint of the continuum (i.e. faces that contain 50% of each identity) faces are perceived as either of the two identities equally as often. However, a personally familiar face when morphed with an unfamiliar face was more likely to be perceived as unfamiliar at the midpoint of the continuum. Instead, a morph needed to include at least 60% of the personally familiar face in order to be perceived as personally familiar more often. These findings were attributed to a categorical boundary shift towards the personally familiar face, due to sharpened tuning to the features that represent the familiar identity. Consequently, listeners displayed a conservative response bias and were able to reject face images that were incompatible with representations of personally familiar faces. Thus, the findings from Chauhan and Gobbini's study and the current experiment observed more conservative decisions about identity when perceiving personally familiar others, suggesting that there may be similar mechanisms involved in both. In Experiment 2, fine-tuned representations of personally familiar voices allowed for a rejection of voice excerpts that violated stored representations of their romantic partner (due to acoustic modulation), whilst preserving the ability to accurately reject tokens from other speakers.

2.5 Experiment 3: Speech perception from personally-familiar voices

2.5.1 Introduction

In the final experiment of this chapter, I aimed to explore whether personal familiarity with a voice can also produce other benefits in addition to identity perception, such as benefits for comprehending speech. Understanding the content of speech is highly important for communication, yet accurate speech recognition is not always easy to attain. For example, hearing voices against background noise, in the presence of multiple talkers, or if the listener has a hearing impairment can make understanding speech difficult. However, there is a growing body of literature that reports familiarity benefits for speech perception by comparing familiar to unfamiliar voices in various speech in noise tasks (Holmes, Domingo, & Johnsrude, 2018; Johnsrude et al., 2013; Kreitewolf, Mathias, & von Kriegstein, 2017; Newman & Evers, 2007; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). In these tasks, familiar voices are found to be more intelligible than unfamiliar voices, and these familiarity advantages have been observed for lab-trained voices (e.g. Nygaard et al., 1994; Nygaard & Pisoni, 1998; Kreitewolf et al., 2017), and personally familiar ones (e.g. Souza, Gehani, Wright, & McCloy, 2013; Holmes et al., 2018; Johnsrude et al., 2013; Holmes & Johnsrude, 2020).

For instance, several studies have found that newly familiar/lab-trained voices are more intelligible than unfamiliar voices when placed in white (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994), or signal-correlated noise (SCN; Kreitewolf, Mathias, & von Kriegstein, 2017). The magnitude of the observed familiarity benefits appears smaller for lab-trained voices (e.g. Nygaard, Sommers, & Pisoni, 1994: 5-10%; Nygaard & Pisoni, 1998: 3-15%) compared to those using personally familiar voices, such as the listeners' spouse or close friend (approximately 10-15%; Domingo, Holmes, & Johnsrude, 2020). Souza, Gehani, Wright, and McCloy (2013) found that older listeners with hearing loss performed better on speech recognition when target speech was spoken by a personally known talker (close friend or spouse) in both quiet and noise, but the magnitude of this benefit was greatest under adverse listening conditions.

Taking a slightly different approach, participants in one study completed both an explicit recognition task and an intelligibility task (Holmes, Domingo, & Johnsrude, 2018). Voice acoustics were manipulated in the explicit recognition phase, and participants indicated on each trial whether a speaker was their spouse or unfamiliar. In the intelligibility task, listeners were simultaneously presented with two different sentences produced by different talkers at two

fixed target-to-masker ratios (TMRs; -6dB and +3dB), and were required to report a target sentence (that began with a particular target name e.g. “Bob”) and ignore the distractor sentence. Both acoustically modulated and original versions of the spoken sentences were presented. A familiar voice benefit was observed for the speech intelligibility task, intriguingly even in the absence of explicit recognition of the participants’ spouse (trials in which vocal tract length was modulated). In this condition, other acoustic cues may have been retained (e.g. speaking rate/style), and it is conceivable that there could be used to facilitate speech intelligibility, even if explicit recognition was impaired due to modulation of specific acoustic properties. Thus, this intelligibility benefit for comprehending speech produced by familiar speakers appears to be robustly reported.

Therefore, in Experiment 3, I examined whether my group of participants would also show familiarity advantages for recognising their partner’s speech in noise. Due to the highly consistent previous research observing intelligibility benefits for familiar voices, I predicted that there would be a higher percentage of correctly reported sentences for stimuli spoken by the personally familiar voice, compared to sentences produced by an unfamiliar speaker.

2.5.2 Methods

2.5.2.1 Stimuli

In this task, I tested speech perception from the personally-familiar voice and a novel unfamiliar voice. The unfamiliar voice was distinct from the unfamiliar voice (“Someone else”) used in the familiarisation, and from the unfamiliar voice (“Clara”/“Charlie”) used in Experiments 1 and 2. All recorded CRM sentences were first RMS normalised. Four-talker babble (multi-talker babble is background noise made up of multiple talkers, in this case four talkers) was then added to each of the sentences from the personally-familiar and unfamiliar voices at a signal-to-noise ratio (SNR) of -6dB, such that the multi-talker babble and target sentences played simultaneously (i.e. there was no delay between the start of the masker and the start of the target sentences). Sample stimuli used in this task are publicly available via the OSF, and can be accessed via the following link: https://osf.io/g2jk6/. The babble noise was created from recordings in the EUROM database of English speech (Rosen, Souza, Ekelund, & Majeed, 2013; Chan et al., 1995), and comprised speakers of the same sex as the to-be-masked speaker – hence, male voices in this experiment were masked with male babble, and

female voices masked with female babble. The same four-talker babble (sex-matched to the speaker) was used on each trial; however, the starting point of the babble noise was randomised for each voice excerpt. Eighty sentence-in-noise stimuli (40 from each voice) were selected for use in the task.

2.5.2.2 Procedure

This experiment was always completed last in the testing session. Here, participants were instructed to listen to the CRM sentences produced by the target speakers (partner [lab-trained Beth/Ben for controls], unfamiliar), whilst ignoring the background noise (four-talker babble). Once each stimulus had played, participants were presented with a grid comprising four rows: each row contained the digits 1-8 in one of the four colour options (red, green, blue, white). Participants were instructed to select the colour and number combination they had perceived from the target sentence. For example, for the sentence stimulus "Ready Baron, go to blue three now", the participant should select the blue 3 from the grid. The 80 stimuli (40 sentences per voice) and eight vigilance trials were presented in a fully randomised order, and the task took around ten minutes to complete. Vigilance trials required participants to respond with a keypress (left or right arrow key) instead of selecting a text response option with their mouse.

2.5.2.3 Data Analysis

Correct answers were defined as trials where participants correctly identified both the colour and number in the target sentence. I did not inspect partially correct answers (e.g. correct colour with incorrect number). The binary correct/incorrect sentence report scores per trial were analysed using generalised linear mixed models (GLMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For GLMMs, odds ratios and confidence intervals are reported. An odds ratio of 1 means that no effect is present. The further an odds ratio deviates from 1, the larger the size of the effect. Confidence intervals that do not cross 1 indicate significant effects.

2.5.3 Results

Data from 5 couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 26 participants per group were retained for the statistical analyses.

2.5.3.1 Couples

In order to investigate the effect of familiarity on speech perception accuracy, a binomial GLMM was constructed. In this confirmatory analysis, the outcome measure was the binary correct/incorrect sentence report score on each trial. Familiarity was defined as a fixed effect; participant and voice identity were entered as random effects. Statistical significance was established by comparing the full model that included the fixed and random effects, to a reduced model. The comparison of the full model to the reduced model was not significant ($\chi^2(1) = .085$, $p = .771$), indicating that accuracy was similar for personally-familiar (mean = 79.3%, SD = 18.9%) versus unfamiliar (mean = 79.0%, SD = 17.4%, OR= 0.87, CI = [0.33, 2.26]) voices (see Figure 6). The odds ratios here compare the odds of being correct (vs incorrect) on unfamiliar voice trials relative to personally familiar voice trials. The OR indicates that participants were 1.15 (1/0.87) times more likely to get a correct answer on personally familiar voice trials compared to unfamiliar voice trials. However, as the 95% CI includes 1, this effect is not statistically significant. Against my predictions, I did not find a familiarity benefit in this task.

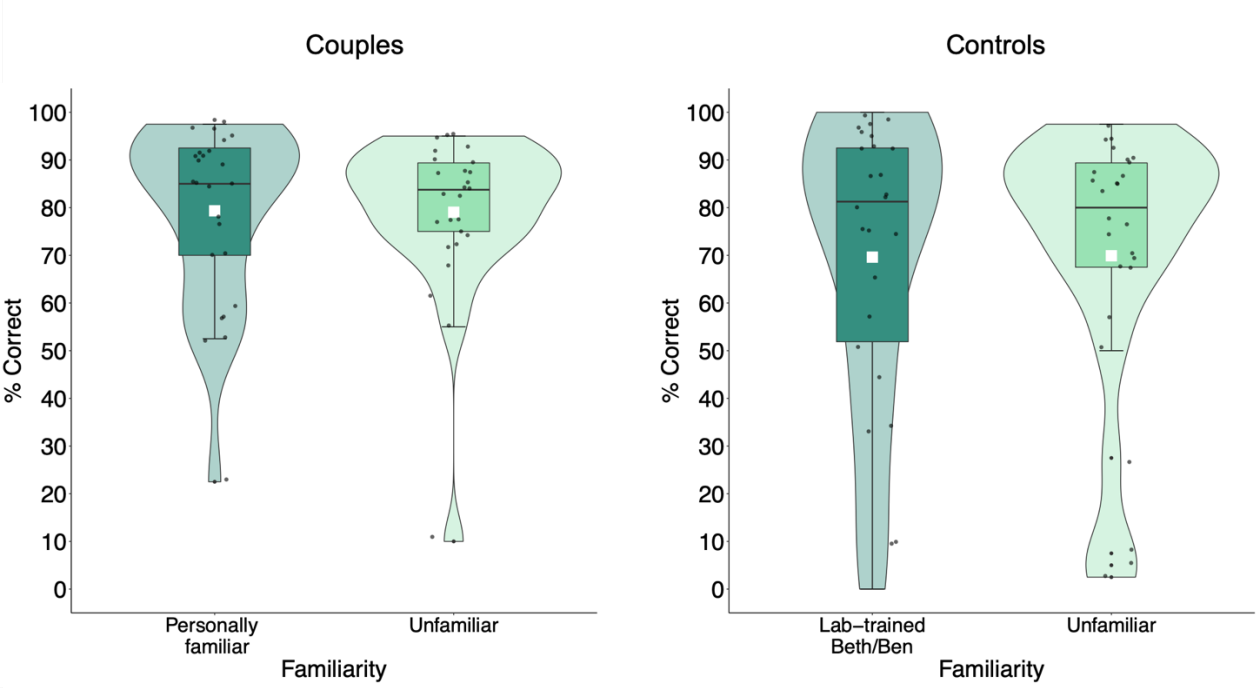

Figure 6. Box plots display median accuracy for the speech intelligibility task (Experiment 3) as a percentage for the personally-familiar and unfamiliar (couples) or lab-trained and unfamiliar (controls) identities. The boxes range from the first to third quartiles (25th and 75th

percentiles). Whiskers extend to no further than 1.5*interquartile range above and below the 1st and 3rd quartiles. Points represent individual participants' scores for each identity. White squares display the group means per condition.

2.5.3.2 Controls

If it is assumed that enhanced speech intelligibility in this study reflects relative familiarity with a voice, rather than variations in the acoustic clarity of some talkers, then any observed personal familiarity advantage for speech perception should be at least as large as that seen in the control group (for whom the familiar voice in this task is lab-trained).

A binomial GLMM was used to examine whether lab-trained familiarity (here, using the "Beth"/"Ben" voice only) had an effect on participants' accuracy for sentences in background four-talker babble. The full and reduced models were constructed in the same way as described for the couples group. Statistical significance was established by comparing the full model that included the fixed and random effects, to a reduced model that did not contain familiarity. I found that the comparison of the full model to the reduced model was not significant ($\chi^2(1) = .007$, $p = .933$; see Figure 6). Thus, there was no speech perception benefit for the lab-trained identity (mean = 69.6%, SD = 30.3%) compared to the unfamiliar voice (mean = 69.9%, SD =28.5%, OR = 1.04, CI = [0.46, 2.34]). The OR here indicates that participants were 1.04 times more likely to get a correct answer in the unfamiliar compared to the lab-trained voice condition, but again the 95% CI crosses 1 meaning that this effect is not statistically significant.

2.5.4 Discussion

In Experiment 3, I sought to examine potential effects of personal familiarity on speech perception. Although previous research has reported familiar talker benefits for understanding speech in noise, the results of this experiment did not suggest any benefit to personal familiarity when participants attempted to comprehend their partner's speech presented in background multi-talker babble. No significant differences in intelligibility were observed when sentences were produced by the participants' romantic partner compared to an unfamiliar voice.

These findings are contrary to previous studies which have reported found large intelligibility benefits (i.e. 10-15% difference in accuracy) for speech produced by a familiar talker against

a background masker, compared to when the target speaker was unfamiliar (e.g. Johnsrude et al., 2013; Nygaard, Sommers, & Pisoni, 1994). There may be various reasons as to why intelligibility benefits were not found in the current experiment. The type of masker used and the relative amplitude of the target voice (i.e. the SNR) may be important. Studies exist that have used the same type of masker (i.e. multi-talker babble; e.g. Souza et al., 2013) or the same SNR (e.g. Johnsrude et al., 2013; Domingo et al., 2019) as the current study. For instance, Domingo, Holmes, and Johnsrude (2020) used a range of target-to-masker ratios (TMRs), including -6dB used in the current study. At this TMR, accuracy for familiar speech in the condition most similar to my participants (i.e. young spouses) was found to be around 75% accurate when attempting to comprehend their partner's speech in the presence of a competing sentence spoken by an unfamiliar speaker. This was in comparison to around 60% accuracy on average when both the target and masker speakers were unfamiliar. In contrast, the current experiment found ~80% accurate responses on average for both personally familiar and unfamiliar targets using a four-talker babble masker. Thus, the type of masker used may be important and indeed recent research has shown that this influences the magnitude of the familiarity benefit for intelligibility. Specifically, the authors found that familiarity benefits were largest when the masker was linguistically similar to the target (Holmes & Johnsrude, 2020). Moreover, as overall accuracy for comprehending the unfamiliar voices was high in the current study (~80%) compared to the previous study outlined above, another possibility may be that the combination of SNR and type of masker used may have been less challenging for participants. Indeed, prior research reporting familiar voice benefits often finds moderate levels of accuracy for unfamiliar targets (i.e. ~40-65%; Johnsrude et al., 2013; Holmes, Domingo, & Johnsrude, 2018; Levi, Winters, & Pisoni, 2011). There is also some evidence that reveals the greatest benefits for speech recognition at intermediate levels of background noise in studies that use intelligibility-enhancing signals such as lip-reading visual cues (e.g. Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Despite this, familiarity benefits have been observed even when accuracy was high for familiar voices (Nygaard, Sommers, & Pisoni, 1994). Therefore, although it might be concluded that perhaps a ceiling effect was been reached in this experiment, 80% accuracy is still not perfect. Thus, the absence of a familiarity benefit is surprising when placing the results within the highly consistent preceding literature. It also contradicts the large familiar voice benefits observed in the previous experiments in this chapter, particularly as the same personally familiar voices and listeners were used in all three experiments.

2.6 General Discussion

The three experiments in this chapter aimed to explore voice recognition and speech intelligibility by comparing different types of familiar voices; namely personally familiar and lab-trained voices. In each of these experiments, challenging manipulations were included that were anticipated to impede perception. Such manipulations were introduced to examine whether the expected impairments to recognition/intelligibility differed as a function of familiarity, and enabled an exploration of the extent of human voice recognition capabilities. Experiment 1 tested voice recognition from conversational filler sounds (e.g. "um", "uh") that were short in duration and contained minimal linguistic information. I found that recognition of personally familiar voices was extremely accurate despite the challenging nature of the task. In contrast, the use of conversational fillers proved significantly more detrimental for recognising lab-trained voices. Similarly, Experiment 2 showed significantly better recognition of acoustically modulated and unmodulated sentences for personally familiar voices compared to lab-trained voices across all modulation conditions. Lastly, in addition to voice recognition, Experiment 3 instead explored familiarity advantages for speech intelligibility in noise (a multi-talker babble). However, here, no significant benefit to speech intelligibility was observed, contrary to prior research finding familiar talker benefits for understanding speech.

In the experiments exploring vocal identity perception (i.e. Experiments 1 and 2), listeners demonstrated highly accurate recognition of their romantic partner's voice, both from brief, conversational filler sounds, and when voice excerpts had been acoustically modulated. This was in contrast to recognition that was largely disrupted for lab-trained voices, despite both of these voices falling within the category of "familiar." Thus, the observed differences in accuracy for personally familiar and lab-trained voices demonstrates that familiarity can be defined in a number of ways, and this has implications for perception. Both personally familiar and lab-trained voices could be recognised with high accuracy during familiarisation, indicating that both were indeed "familiar", yet during the vocal identity tasks, listeners struggled to maintain accurate recognition for the lab-trained voice whilst recognition of the personally familiar voice remained largely unimpaired. This suggests that recognition is disrupted for lab-trained voices when the tasks require the listener to generalise beyond what was learned during training. Previous research using lab-trained and unfamiliar voices supports this. For instance, Winters, Levi, and Pisoni (2008) investigated the effects of the language of

speech on voice recognition for lab-trained voices. Listeners were trained to recognise a group of bilingual speakers in either English or German, and performed a recognition task. If trained to recognise the speakers in English, recognising the same speakers when they were speaking German was impaired, and thus the listeners failed to generalise their knowledge of the identities from English to German and vice versa. Similarly, voice sorting studies find that unfamiliar listeners' poorer performance is largely due to a reduced ability to generalise identity information across different instantiations of a single speaker, with these listeners often incorrectly fragmenting a single identity into many perceived identities (Lavan, Burton, et al., 2019). This difference in generalisability may be explained by the underlying voice representations that exist for different familiar voices. For lab-trained voices, the representations formed were constructed from a particular set of vocalisations during familiarisation – these were spontaneous speech excerpts in the current chapter. This means that these representations are likely to be relatively rigid and generalisation from such under-specified representations is challenging (Lavan, Knight, Hazan, & McGettigan, 2019b). In comparison, representations for personally familiar voices should be well-rounded and robust as these voices have been encountered in a wide variety of contexts. In Experiment 1, attempts to recognise brief conversational filler sounds may have been difficult for the lab-trained voice as the conversational speech excerpts used during training did not include any filler sounds, meaning that listeners had to generalise to perform Experiment 1. On the other hand, listeners may have remembered what their partner's filler sounds are like from previous conversations, and/or from hearing the exact stimuli during the recording session, such that generalisation was not necessary here. However, in Experiment 2, generalisation was essential as the acoustic manipulations pushed all of the voices outside the speakers' usual vocal repertoire, sometimes extending beyond what would be physically possible with the speaker's vocal apparatus (i.e. increasing/reducing the apparent vocal tract length). It is compelling that again, for personally familiar voices, listeners were much better able to generalise to these truly novel portrayals of their partner's voice. Therefore, perhaps a greater ability to generalise across different examples of a speaker in order to maintain accurate recognition is at the crux of what makes a robust or refined stored voice representation.

The results from Experiments 1 and 2 largely align well with recent theoretical models of vocal identity (Maguinness, Roswandowitz, & von Kriegstein, 2018; Lavner, Rosenhouse, & Gath, 2001). These models propose that incoming vocal signals are compared to a prototype (an average/very commonly experienced voice), deviant acoustic features are extracted, and these

deviations are compared to stored reference patterns or representations for familiar voices. If the distance between the deviant features and stored representations is smaller than a particular threshold, the voice is recognised (Maguinness, Roswandowitz, & von Kriegstein, 2018; see General Introduction chapter). Moreover, an account is given as to how unfamiliar voices may become familiar, arguing that representations for these voices are established through iterative exposure to these voices over time, until a robust representation is established. The lab-trained voices in this chapter were unfamiliar prior to the experiment, and the training provided enabled representations to be formed, although these proved not to be robust enough to contend with the challenges introduced at test. On the other hand, the romantic partner voices were naturally acquired, highly familiar others, and were associated with, in the model's terms, many more iterations through the perceptual processing loop. This was verified by participants that reported speaking to their romantic partner for 34.66 hours per week on average. Therefore, the notion that repeated exposure enables robust representations to be developed over time is supported by the current findings, whereby listeners could recognise these speakers extremely accurately from challenging, previously unheard vocalisations when personally familiar, but the same listeners struggled to contend with these challenges for voices they had less experience with.

Nonetheless, some outstanding questions remain, and future research is needed with both lab-trained and personally familiar voices to answer these questions. For personally familiar voices, this type of familiarity is associated with more robust representations that allow for accurate and flexible recognition. Yet a full appreciation of how this robustness and flexibility is encoded into voice representations is needed. Research by Lavan, Knight, and McGettigan (2019a) found some evidence for average-based representations of individual speakers. However, this principle alone does not explain how listeners can recognise highly familiar speakers from varying vocalisations, and how within-person representations support such flexible recognition. One recent proposal is that as a voice becomes increasingly familiar, representations may be expanded from a single point in representational space to a region in voice space (Stevenage, 2020). Perhaps deviant features from this within-person prototype may be extracted and compared to the incoming signal in order to recognise varying instances of a single speaker that fall within a circumscribed region in representational space, similarly to the ideas laid out by theoretical models. However, this raises questions about how the processes involved for between-speaker and within-speaker recognition may interact. For instance, if deviations from a between-person prototype are calculated, as well as deviations from a within-

person prototype to recognise variations within an individual speaker, how these processes interact with each other to afford accurate recognition is not yet fully understood. Moreover, it may be expected that individual voice spaces may overlap with each other in a between-person representational voice space, and therefore the question remains as to how accurate recognition is achieved. Thus, there is a need for further research to examine the nature of representations for highly familiar voices, such as what is contained within these representations, how they enable flexible recognition, and how such representations are formed.

For the lab-trained voices used in this chapter, the amount of training provided was relatively brief, but was arguably representative of the extent of training provided in previous vocal identity studies. Further research is needed to investigate the type and amount of exposure necessary for listeners to build up robust representations for lab-trained voices. Perhaps with extensive training, it may be possible to develop a familiarity with lab-trained voices that produces recognition rates similar to personally familiar voices. However, the amount and type of exposure listeners had to personally familiar and lab-trained voices is not the only difference between these familiar voices. For instance, listeners' knowledge of their partner's voice is also accompanied by stored memories of what that person looks like, as well as other physical, biographical, historical, and affective information, which is crucially built up via naturalistic, social interaction with the partner. There is some evidence from the voice learning literature that voice learning is facilitated by the presence of extra information, such as a face (Zäske, Mühl, & Schweinberger, 2015; Sheffert & Olson, 2004) and that memory for faces is improved with additional person information (Mattarozzi, Colonnello, Russo, & Todorov, 2018). Whether extensive training alone in the absence of this "extra" information is sufficient for robust voice representations to be formed is not yet clear, and therefore this is a question that should be explored in future investigations.

Moreover, lab-trained voices could be useful for exploring how decision-making about identity for voices changes over time when learning new vocal identities. In Experiment 2, I found that listeners were more flexible in recognising personally familiar voices with reduced cues to identity, but more conservative in rejecting tokens that did not align with stored representations. It is unknown whether this type of "tuning" happens in the learning of new voices. It may be that listeners initially accept a wider range of plausible voice tokens as representative of a newly learned identity, and become more restrictive over time through greater exposure and encoding of this new voice. This would be logical with reference to

successful social communication as it may be safer to incorrectly accept a heard voice stimulus as belonging to somebody you have newly encountered, than to reject and fail to recognise a token from that speaker. Recognition could be examined at various time points to investigate how decisions about identity may change as a function of increasing familiarity. These questions are addressed in Chapter 3. It should be noted that conclusions about "tuning" observed for personally familiar voices in Experiment 2 remain tentative as performance for unmodulated voice tokens for these voices was extremely accurate and thus this also meant there was more space for performance to decrease. Thus, research could first further investigate this tuning function for personally familiar voices, followed by an investigation of whether this transpires in the learning of new identities.

In Experiment 3, personally familiar voice benefits did not translate to an enhanced ability to comprehend speech in 4-talker babble. This finding is surprising, as intelligibility benefits have been replicated quite consistently in the existing literature. Therefore, outstanding questions also remain here, such what conditions produce the largest familiarity benefits for enhancing speech intelligibility, such as the combination of stimulus type, masker type (noise, one talker, multiple talkers), task type (e.g. open vs. closed set recognition), and target-to-masker ratio.

Although two types of familiar voices were used in this chapter, namely lab-trained and personally familiar speakers, it is important to note that even within these categories, the level of familiarity may vary substantially. For instance, personally familiar people are defined as those we have real world experience and interactions with (Sugiura, 2014). I selected romantic couples to represent personally familiar voices. However, the voices we interact with on a daily basis are mainly personally familiar or unfamiliar, and the degree of familiarity with different people is highly variable. This is exemplified in a study by Lavan and colleagues (2016) who used University lecturers' voices as stimuli and participants who were students that had/hadn't been taught by these individuals. This study found that familiarity gained listeners advantages in making identity judgements, however, when the task required them to generalise across different types of vocalisations or to less frequently encountered vocalisations, familiar listeners' performance was similar to an unfamiliar group of listeners. The experiments in this chapter and the above study both used personally familiar voices, but the ability for recognition and generalisation appears to depend on the amount and contexts of prior exposure to these voices. Familiarity then, may be better conceptualised as a continuum, whereby the lab-trained and personally familiar voices used in this chapter may represent two possible points along this

familiarity gradient. Defining familiarity in this way may affect how familiar voice perception is studied and understood and can enable us to update models of vocal identity accordingly.

Lastly, the generalisability of the results in this chapter is subject to certain limitations. For instance, the use of closed set tasks across all three experiments does not closely resemble the 'open set' nature of voice recognition in naturalistic settings. Moreover, participants in Experiment 2 were explicitly told that the three voices would sometimes sound different, and to attempt recognition despite this. It is unclear if participants would have performed differently if they were not informed of this (see Holmes, Domingo, & Johnsrude, 2018). It is important to note the influence that task instructions can have on the observed behaviour. However, in this case the use of explicit instructions and forced choice response options revealed participants' decision-making strategies and the types of errors that were made, which in turn was informative for understanding the underlying representations of these voices.

Overall, the findings in this chapter investigated the accuracy of voice identity perception, directly comparing voices of differing degrees of familiarity. I found that performance for newly-learned, lab-trained voices can be poor and prone to error, particularly when there is a need to generalise to novel contexts. As most voices that we engage with in day-to-day life are contained within specific contexts, one may conclude that voice recognition is often error-prone (e.g. earwitness testimony does not hold up as evidence in court). However, these experiments also revealed that vocal identity perception can be extremely reliable for voices at the upper end of the familiarity spectrum, highlighting that the extent of human voice recognition capabilities may have been underestimated.

3 Recognition of Lab-Trained Voices from Acoustically Modulated Speech: Effects of Voice Training

3.1 Experiment 4

Voice recognition is a crucial skill for successful communication, yet it is a challenging and complex process. The human vocal system is flexible, allowing for the production of unique vocalisations, and a speaker will never produce exactly the same utterance twice (Latinus & Belin, 2011). A speaker may purposefully or unconsciously alter their voice in response to various social (e.g. speaking to a child vs. making a telephone appointment) and environmental contexts/pressures (e.g. speaking in a noisy restaurant vs. in quiet; Lavan, Burton, et al., 2019). A speaker's voice can also be highly variable within a single interaction, such as transitioning from a neutral tone to speaking through laughter, to lowering the volume of the voice to a whisper. When a new voice is encountered, listeners must learn to recognise this voice quickly, and be able to maintain accurate recognition of the voice in all of its variations. How rapidly recognition of voices improves with increasing experience and how such experience affects recognition not only in optimal listening conditions, but in perceptually challenging contexts, is something the two experiments in the current chapter sought to explore.

3.1.1 Introduction

To successfully recognise a voice, a listener must be able to form a robust and flexible voice representation. A prominent account of how voices are learned and encoded proposes that voices are represented in a multidimensional voice space (Maguinness, Roswandowitz, & von Kriegstein, 2018). Upon hearing a voice, listeners extract voice identity features and compare these against a prototype, which is deemed to be either an average or very frequently encountered voice (see Chapter 1). The differences between the incoming signal and the prototype are calculated which can then be used to compare to "stored reference patterns", which are argued to be exclusive to each voice identity. If the distance between the incoming signal and stored reference pattern is smaller than a perceptual threshold, a listener should get a sense of familiarity or be able to identify the speaker. Further to this account, recent work has

found evidence that listeners may also form average-based representations of individual voice identities to be able to recognise that varying vocalisations produced by the same speaker 'go together' despite potentially sounding very different from each other (Lavan, Knight, & McGettigan, 2019a).

For unfamiliar voices to become familiar, a reference pattern needs to be established over time, leading to newly familiarised voices having relatively incomplete reference patterns. How representations are built up over time is underspecified in the prominent models of voice identity processing. One account mentions that reference patterns will be refined with continued exposure to the voice, building up more robust representations with each iteration through a "perceptual voice-identity processing loop" (Maguinness, Roswandowitz, & von Kriegstein, 2018). However, the factors important for this process to take place and the specifics of how long it takes for a familiar voice pattern to be acquired, as well as the acoustic features stored in such representations, is not fully known. Based on the current knowledge, we can assume that a listener will probably need to have prolonged, repeated, and variable exposure to a speaker's voice to be able to build up a representation (or stored reference pattern) robust enough to be able to tell this voice apart from other voices, and to also be able to recognise the voice in all or many of its variations (Lavan, Burton, et al., 2019).

In the previous chapter, I explored how well personally familiar voices of romantic partners could be recognised under various challenging listening conditions. I found that these voices could be recognised extremely accurately, even from short filler sounds with no meaningful linguistic content (Experiment 1) and from acoustically modulated voices (Experiment 2). For these voices, highly robust representations or stored reference patterns have been established, enabling flexible recognition under difficult listening conditions. In Experiment 2, I observed a sharp "tuning function" whereby increasing distance from the unmodulated personally familiar voice was associated with a steep, symmetrical drop in accuracy. This could, in part, have been observed due to overall better performance for personally familiar voices, leaving more room for performance to decrease. However, the configuration of errors for the personally familiar voice revealed an interesting pattern. That is, when errors were made, they involved listeners rejecting modulated voice excerpts that had actually been produced by the listeners' romantic partner, whereas hardly any errors were observed where participants incorrectly labelled one of the other identities as the partner. This may be interpreted as evidence for sharper tuning to the features of personally familiar voices, enabling listeners to retain highly

accurate recognition whilst rejecting tokens that do not fit (i.e. other voice identities) or no longer fit the stored representation of the personally familiar voice.

This more conservative categorisation has also been observed in a face perception experiment using a morphing paradigm (Chauhan, Kotlewska, Tang, & Gobbini, 2020). In their study, the researchers morphed a personally familiar face (friend/self) with an unfamiliar face, as well as morphing two unfamiliar faces for comparison. They created morph continua that ranged from one face to the other in increments of 10% (e.g. 10% face A, 90% face B to 90% face A, 10% face B). For morph continua constructed from two unfamiliar faces, it was found that the 50% morph between the two identities (i.e. the midpoint of the continuum), was judged to be each identity half of the time. In contrast, for morphs made up of a familiar and unfamiliar identity, participants were more likely to label the 50% morph image as unfamiliar. That is, observers' categorical decision boundaries were shifted towards the personally familiar face, such that a morph needed to contain relatively more of the familiar face to be labelled as such. Similarly to my findings in Chapter 2, it was concluded that repeated and prolonged exposure to familiar faces leads to the construction of flexible, enriched representations that are resilient to distortions, whilst simultaneously increasing sensitivity to features that are inconsistent with these stored representations (Chauhan et al., 2020).

These are two features that appear to set personally familiar voices/faces apart from the perception of other, less familiar, individuals (e.g. lab-trained voices). That is, accuracy is improved for recognising personally familiar individuals from highly variable exemplars (the ability to generalise is improved), in addition to having a larger "criterion" value in signal detection terms, or a conservative response bias – i.e. the perceiver needs to have stronger evidence before labelling an incoming signal as familiar.

For lab-trained voices, the nature and extent of training varies from experiment to experiment: from a single 20 to 30-minute training session (von Kriegstein & Giraud, 2006; Zhang, Li, Zhou, Zhang, & Shu, 2021) to multiple sessions across multiple days (Latinus & Belin, 2011; Latinus, Crabbe & Belin, 2011). The number of unique stimuli and the type of stimuli used in these experiments also varies, with some studies using few unique sentences per speaker (~10-20 e.g. Zhang et al., 2021; Winters et al., 2006), and others using considerably more (e.g. >100 stimuli, Holmes et al., 2021; von Kriegstein & Giraud, 2006). It is worth noting the differences that exist in defining "amount of exposure/training" as this can affect the conclusions drawn.

In the experiments in this chapter, the amount of exposure is defined as the duration of exposure to unique voice excerpts (i.e. longer exposures would include a greater number of unique items, rather than the same excerpts presented multiple times). Despite this variation, examining the contexts in which participants show accurate recognition compared to situations where recognition ability is impoverished can allow for inferences to be made as to the nature of the representations of these voices and how they differ from personally familiar ones. Generally, it has been found that voices can be recognised relatively well after only a short amount of training, provided the listening conditions are roughly the same at training and test (i.e. the listener does not have to make generalisations to variations in e.g. speaking style; Lavan, Knight, et al., 2019b). For instance, a recent voice training study explored recognition of new voices after 10, 20, and 60 minutes of training (Holmes, To, & Johnsrude, 2021). Participants displayed relatively accurate and similar rates of recognition for all three conditions (~73% accuracy), suggesting that even as little as 10 minutes of exposure may be sufficient to recognise voices moderately well. Similarly, earwitness testimony research has shown that recognition rates improve with increasing duration of speech excerpts, with poor recognition being observed at very short durations (e.g. 6s) and markedly improved performance at longer durations (e.g. 30s, 70s, 8 minutes; Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2004; Roebuck & Wilding, 1993; Yarmey et al., 2001). This suggests that increasing the amount of exposure or training can create stronger stored reference patterns of recently learned voices, such that recognition is improved (but there may be some limits to this, e.g. see Holmes et al., 2021). This could be akin to a greater number of iterations through the so-called "perceptual voice-identity processing loop" (Maguinness, Roswandowitz, & von Kriegstein, 2018).

However, recognition of trained-to-familiar voices has also been shown to be somewhat unstable, and identification of these voices is easily disrupted. For instance, there is some evidence from early earwitness testimony literature that recognition is poor if the target speaker deliberately disguises their voice at test (Reich & Duke, 1979), or if listeners are required to recognise speakers across languages (Winters, Levi, & Pisoni, 2008). Even relatively modest changes such as varying expressiveness of the voice (e.g. neutral to angry tone) have been found to negatively affect recognition for low-/moderate-familiar or trained-to-familiar voices (Saslove & Yarmey, 1980; Lavan, Burston, et al., 2019b). In the previous chapter, listeners were trained to recognise the lab-trained voices from training that involved 20 excerpts from each speaker (plus passive exposure to 24 novel excerpts). Recognition of these lab-trained voices was also poor in both identity tasks when listeners were required to generalise to

different vocalisations or to acoustically manipulated vocalisations. In Experiment 2, the "tuning function" for recognition of acoustically-modulated lab-trained voices was much flatter, with errors not following a particular pattern as was observed for personally familiar voices. This suggests that for a lab-trained voice, participants had a less robust ability to recognise when a modulated voice token was or was not produced by that particular speaker. Thus, the evidence suggests that the lab-trained familiarity established in previous studies has not been sufficient to support identity recognition in tasks that required listeners to generalise beyond what was learned at training. Representations of these voices appear to be relatively rigid, making generalisation from an underspecified reference pattern challenging (Lavan, Knight, et al., 2019b).

Increasing the amount of training from seconds to minutes appears to improve voice recognition performance (Kerstholt et al., 2004; Roebuck & Wilding, 1993; Yarmey et al., 2001; Schweinberger, Herholz, & Sommer, 1997), but beyond a certain point there may be no additional benefits, at least on the same task (Holmes et al., 2021). However, the effects of training and amount of exposure might be more evident in terms of how voice identity knowledge generalises to new tasks – for instance, the study by Holmes and colleagues (2021) found that the amount of voice training did provide benefits for speech intelligibility, despite no apparent differences being observed in voice recognition ability. Specifically, the largest benefits to speech recognition in noise were observed in the longest training condition (60 minutes of training) compared to the other training conditions. Additionally, the findings in Chapter 2 highlighted that voice recognition ability - on tasks that required generalisation - differed based on the extent and type of familiarity with a speaker. Whilst personally familiar voice recognition remained largely unimpaired/intact, recognition of lab-trained voices suffered when generalisation was required. Thus, there is a need to better understand the relationship between the amount of exposure during voice learning and the robustness of the resulting voice identity representations. In the face perception literature, one study carried out by Stevenage (1998) explored the effect of training on participants' ability to discriminate previously unfamiliar identical twins. Before training, many errors were made in accurately determining whether two images were of the same twin or different twins. After training, same-twin pairs (i.e. twin A- twin A photograph pair) looked significantly more similar to each other, and different-twin pairs of photographs (i.e. twin A – twin B pair) were rated as looking significantly more different. Thus, explicit training (including 70 trials in total) was associated with an improvement in identifying subtle differences between identical twin faces. This

provides some limited evidence that there may have been sharpening of the tuning function for previously unfamiliar faces due to training.

Whether increasing training can improve generalisation and create sharper tuning to the specific features of the learned voice is something the current chapter aimed to explore. Specifically, in two experiments, I examined the effect of the amount of training on participants' ability to recognise acoustically modulated versions of lab-trained voices. Participants either received shorter training (20 unique voice excerpts per identity), or longer training (80 unique voice excerpts per identity) to learn to recognise voices. Recognition for acoustically modulated and unmodulated excerpts was tested. The results of the previous chapter demonstrated that for personally voices that have been learned through repeated and prolonged exposure, representations of these voices are highly robust. As a result, these voices can be recognised accurately from widely variable exemplars, as well as listeners maintaining an improved ability to correctly reject a voice that does not fit with stored representations. Thus, in the current experiments, it was hypothesised that more training would result in improved recognition ability due to more robust stored reference patterns of these voices formed via increased exposure. Improved generalisation ability to acoustically modulated voices as well as sharper tuning to the specific features of the learned voices was also expected. Therefore, it was predicted that performance would be significantly better across all modulation steps in the longer training condition compared to the shorter training condition. Further, a significant interaction between amount of training and modulation step was expected. In particular, a sharper and symmetrical drop in performance with increasing distance from the unmodulated learned voices in the longer training condition was expected, compared to the shorter training condition, where a flatter function was expected.

3.1.2 Methods

3.1.2.1 Participants

All participants were recruited on the online recruitment platform Prolific.co (www.prolific.co) and completed the experiment on Gorilla.sc (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018). At the start of the experiments, participants set their volume to a comfortable listening level and were required to pass a headphone screening to ensure that they were

wearing headphones and able to hear the stimuli presented (Woods, Siegel, Traer, & McDermott, 2017). All participants were native English speakers, had normal or corrected-to-normal vision, and reported no hearing difficulties. None of the participants recruited had taken part in any of my prior research studies conducted on Prolific. On completion of the experiment, participants were paid at a rate of £7.50/hr of participation. Ethical approval was obtained via the UCL research ethics committee (approval code: SHaPS-2019-CM-030) and informed consent was given by all participants.

Seventy participants in total were recruited to take part in Experiment 4. Six participants were removed due to failing to show sufficient attention to the task (5 participants), or performing below chance level in the familiarisation phase (1 participant), leaving 64 participants (32 female, mean age = 28.38 years, SD = 6.44 years, range = 18-39 years) in the final data set. Half of the participants were assigned to the shorter 20 stimuli training condition, and the other half assigned to the longer 80 stimuli training condition.

3.1.2.2 Materials

Voice excerpts were extracted from audio recordings of individuals performing the Diapix task (Baker & Hazan, 2011) as well as read sentences taken from the freely-available LUCID corpus materials from those same speakers (Baker & Hazan, 2011). The Diapix task involves pairs of participants performing an interactive spot-the-difference task, and elicits spontaneous speech from the speakers.

In the modulated voices task described in Experiment 2, each control participant had their own version of the experiment involving 3 opposite-sex voice identities: two lab-trained voices, and an unfamiliar identity. I selected 4 of these voice combinations for use in Experiment 4 of the current chapter: 2 with female voices, and 2 with male voices. I chose two versions from the two sexes to provide some variation in the stimuli used across participants, and to reduce the potential effects of individual voice distinctiveness on learning. To match this to the experiments in Chapter 2, male participants learned two female voices, and vice versa.

The familiarisation of the two new voices involved a passive exposure phase and training phase. Spontaneous voice excerpts taken from the Diapix recordings were used for familiarisation. In the passive exposure phase, spontaneous voice clips were arranged into two

12-excerpt sequences per speaker, with each sound clip separated by 1s of silence. Therefore, each voice had 24 stimuli used for passive exposure. For the training phase, either 20 or 80 spontaneous excerpts were selected from three identities (2 to-be-learned, 1 unfamiliar). Participants were required to actively identify these spontaneous voice excerpts and were provided with feedback on each trial (see Design and Procedure below). There was no overlap in the excerpts used in the passive exposure phase and training phase.

For the voice modulation test phase, 50 read sentences from the LUCID corpus materials (e.g. "The beach stall sold bats and balls") were selected for each of the speakers (2 lab-trained, 1 unfamiliar; note that the unfamiliar speaker here was not the same unfamiliar speaker as the training phase). These sentences were the same modulated sentences used in the modulation task in Chapter 2 (see this chapter for details on voice modulation). 12 stimuli were used for each modulation step for the two lab-trained voices. As there were only 50 recorded sentences available, two randomly selected items from each modulated step and from the unshifted voice recordings were repeated once during the task. Six tokens per step were used for the unfamiliar voice.

3.1.2.3 Vigilance stimuli

A text-to-speech online tool (https://text2speech.us/) was used to generate voices reading "Please press the left key" and "Please press the right key." These were used in vigilance trials (8 per task; 4 of each instruction) to check participants' attention in the test phase of the task.

3.1.2.4 Design & Procedure

Participants were randomly assigned to one of the 2 opposite-sex versions of the experiment. Participants first set their volume and completed the headphone check.

3.1.2.4.1 Familiarisation training

In the familiarisation, participants were first passively exposed to two new named voices ("Anna" and "Beth" for male participants, "Adam" and "Ben" for female participants). Text on screen read "This is [NAME]. Listen carefully and try to memorise how this voice sounds." Participants firstly passively listened to the two 12-excerpt sequences of spontaneous speech for both newly introduced speakers to get an initial exposure to what these voices sounded like.

Following this, participants learned to accurately recognise these voices via a test phase. Recognition was tested using a forced choice paradigm. On each trial, a voice excerpt was presented, followed by three text response options: “Anna”/“Adam”, “Beth”/“Ben”, and “Someone Else.” Participants were required to select one of the response options with their mouse. Audio-visual feedback (correct/incorrect) was given on every trial to aid in the learning of the new voices. For participants in the shorter training condition, the forced choice test phase included 20 spontaneous speech excerpts from each of the two newly introduced speakers, as well as 20 voice clips from the distracter “someone else” voice (totalling 60 trials for the test phase), presented in a fully randomised order. For participants in the longer training condition, 80 spontaneous speech excerpts from each of the three voices were presented. The training in this condition was split into 4 training blocks of 60 stimuli (20 per voice identity; total of 240 stimuli), and participants had the opportunity to take breaks between each block. The shorter training corresponded to approximately 10 minutes of training, whereas the longer training corresponded to ~25 minutes of training.

3.1.2.4.2 Modulation Task

Following familiarisation, all participants completed the voice modulation task. Before this test phase began, participants were introduced to a novel speaker “Clara”/“Charlie” and were presented with one example excerpt from this speaker. Note that this was a different unfamiliar talker from the familiarisation phase. In this task, participants were presented with 150 modulated and unmodulated read sentences (60 each for the two learned speakers, 30 for the unfamiliar speaker) in a fully randomised order. Participants were informed that they would be hearing manipulated and original versions of the voices and were asked to attempt recognition regardless of these manipulations. On each trial, a sentence was presented, followed by three response options: “Anna”/“Adam”, “Beth”/“Ben” and “Clara”/“Charlie.” Participants were to decide on each trial which speaker they thought produced the sentence by clicking a response option with their mouse. Vigilance trials required participants to follow an instruction to respond with a left or right key press, instead of selecting a text response option with their mouse. Participants that failed to respond correctly on at least 75% or 6/8 of these trials were excluded and the participant replaced. This task took approximately 15 minutes to complete.

3.1.2.5 Data Analysis

For the familiarisation, overall mean accuracy scores as well as accuracy for each block of training were calculated. For the voice modulation task, to correct for any disproportionate usage of certain response categories, unbiased hit rates (Hu scores) were calculated for each of the three voices (Wagner, 1993). However, individual participant Hu scores for the two lab-trained voices were grouped into one as these voices represented one category. Hu scores were arcsine transformed (Wagner, 1993). Data were analysed using linear mixed models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For the LMMs, model estimates and associated confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero indicate significant effects.

3.1.3 Results

3.1.3.1 Familiarisation training

At the end of training, participants could recognise the two lab-trained voices reasonably well and to a similar extent in the group who received shorter training (mean = 77.03%, SD = 10.71), and the group who had longer training (mean = 81.1%, SD = 8.28; see Figure 7 for mean accuracy in each of the four training blocks).

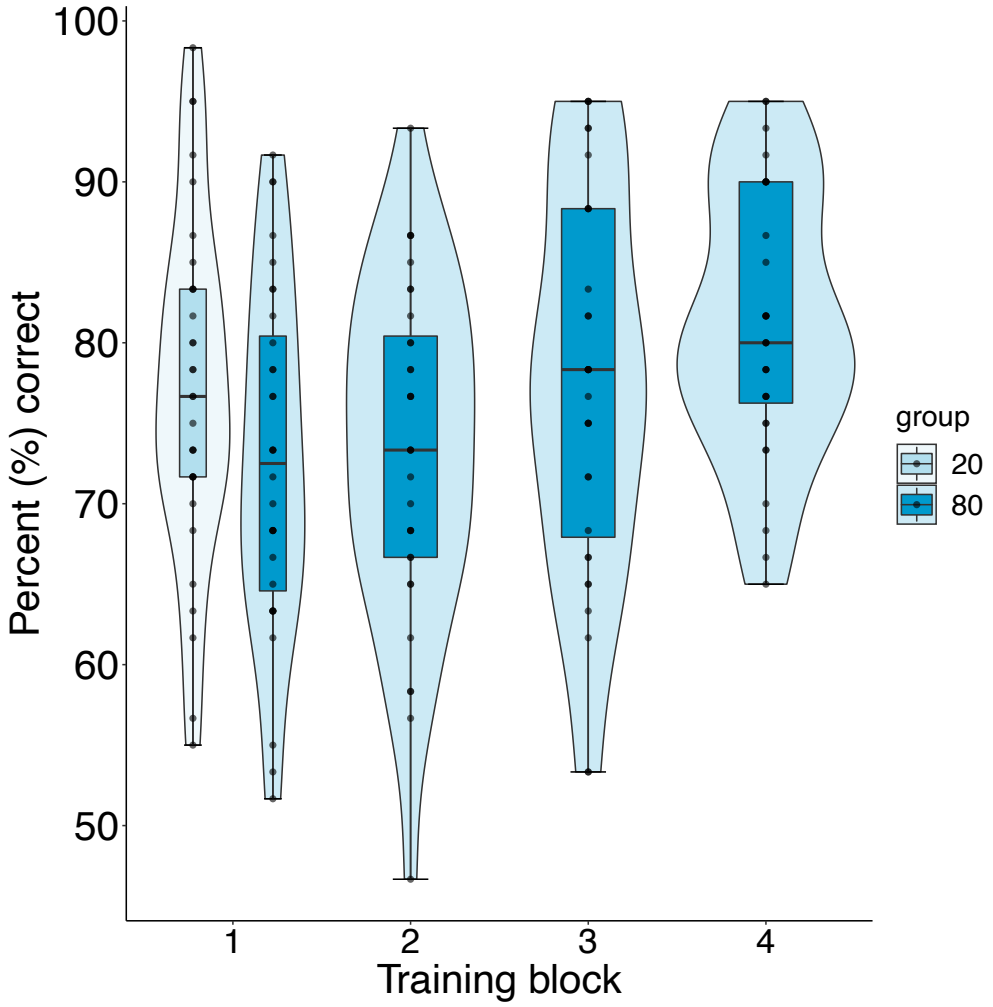

Figure 7. Box plots display median percent correct scores per training group (20 vs. 80 stimuli) and training block. Boxes range from the first to third quartiles and whiskers extend to no further than 1.5*interquartile range above and below the 1st and 3rd quartiles. Individual participants' mean performance is displayed as individual points. Note that the shorter training group (20 stimuli) only performed one block of training.

3.1.3.2 Modulation task

Averaging across all modulation steps for the lab-trained voices per training group showed that the mean overall performance for those in the shorter training condition (20 stimuli) was 44.9% (SD = 10.07%), with mean scores on the individual modulation steps ranging over 39.06% - 52.26%. Mean overall performance for those in the longer training condition (80 stimuli) was 44.76% (SD = 13.07%) with mean scores on the individual modulation steps ranging over 39.15% - 53.04%).

To compare the effect of acoustic modulations on the recognition of the lab-trained and unfamiliar voices in participants with shorter or longer training sessions, I analysed the three-way interaction between the degree of modulation, familiarity (lab-trained, unfamiliar), and amount of training (20 vs. 80 stimuli) using LMMs.

The model was set up as follows:

*lmer(hu score ~ familiarity*modulation_step*amount_of_training + sex + (1|participant) + (1|speaker identity))*

where the Hu score is the outcome measure for recognition performance. Familiarity, modulation step, amount of training, and sex of the participant were included as fixed effects; participant and speaker identity were included as random effects. Statistical significance was obtained by comparing the full model that contained the interactions, fixed effects, and random effects to a reduced model that contained all two-way interactions, fixed and random effects but did not contain the three-way interaction.

Using this method, I observed that the three-way interaction between familiarity, modulation step and amount of training was not significant ($X^2(4) = 3.85$, $p = .427$, see Table 3 for model estimates and confidence intervals). Significant two-way interactions between modulation step and amount of training were also not observed ($X^2(4) = 3.25$, $p = .517$), as well as the familiarity-by-amount of training interaction ($X^2(1) = 0.74$, $p = .390$). There was also no main effect of the amount of training on participants' accuracy ($X^2(1) = 0.001$, $p = .971$). Thus, no interactions or main effects involving the amount of training were observed, suggesting that the amount of training did not have a significantly different effect on recognition accuracy in the modulation task.

I did, however, observe a significant interaction between familiarity (lab-trained vs unfamiliar) and modulation step ($X^2(4) = 25.03$, $p < .0001$). Post-hoc pairwise comparisons (using *emmeans*; Lenth, 2019) were run to compare lab-trained to unfamiliar performance at each modulation step. I found that at all modulation steps, performance was not significantly better for the lab-trained identities compared to the unfamiliar voice (-2 steps: t ratio = 0.77, p = .510; -1 step: t ratio = 1.63, p = .194; unshifted: t ratio = 2.74, p = .052, + 1 step: t ratio = 2.34, p = .093; +2 steps: t ratio = 2.00, p = .130).

Therefore, I also assessed the effect of increasing the degree of modulation on recognition of the lab-trained and unfamiliar voices, using post-hoc pairwise comparisons (FDR corrected). Performance was compared between successive modulation steps (e.g. -2 steps vs. -1 step, -1 step vs. unshifted condition). For the lab-trained voices, performance was negatively affected by acoustic modulation in both directions (all ps < .05; see Figure 8), whereas for the unfamiliar voice, performance was equivalent across all modulation steps. Thus, within modulation step, there were no significant differences in voice recognition between the lab-trained and unfamiliar identities, however, within identity, increasing the distance from the unshifted condition led to significant decreases in performance, but only for the lab-trained condition.

Table 3. Model estimates and confidence intervals (CIs) from the full model containing the three-way interaction, all two-way interactions, and fixed effects.[1]

		Hu Score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.79	0.66 – 0.92	**< .001**
Familiarity [Unfamiliar]	-0.26	-0.46 – -0.07	**.008**
Modulation Step [-2]	-0.20	-0.28 – -0.12	**<.001**
Modulation Step [-1]	-0.12	-0.19 – -0.04	**.002**
Modulation Step [1]	-0.08	-0.16 – -0.01	**.027**
Modulation Step [2]	-0.14	-0.22 – -0.06	**<.001**
Amount of Training [80]	0.07	-0.02 – 0.16	.107
Sex [F]	-0.21	-0.37 – -0.05	**.009**
Familiarity [Unfamiliar]*Modulation Step [-2]	0.15	0.01 – 0.28	**.032**
Familiarity [Unfamiliar]*Modulation Step [-1]	0.07	-0.07 – 0.20	.326
Familiarity [Unfamiliar]*Modulation Step [1]	0.03	-0.10 – 0.16	.654
Familiarity [Unfamiliar]*Modulation Step [2]	0.06	-0.08 – 0.19	.412
Familiarity [Unfamiliar]*Amount of Training [80]	-0.10	-0.23 – -0.03	.135
Modulation Step [-2]*Amount of Training [80]	-0.11	-0.22 – 0.01	.063
Modulation Step [-1]*Amount of Training [80]	-0.07	-0.17 – 0.04	.216

Modulation Step [1]*Amount of Training [80]	-0.05	-0.16 – 0.05	.348
Modulation Step [2]*Amount of Training [80]	-0.10	-0.20 – 0.01	.078
Familiarity [Unfamiliar]*Modulation Step [-2]*Amount of Training [80]	0.16	-0.03 – 0.35	.097
Familiarity [Unfamiliar]*Modulation Step [-1] *Amount of Training [80]	0.13	-0.06 – 0.31	.182
Familiarity [Unfamiliar]*Modulation Step [1]*Amount of Training [80]	0.03	-0.16 – 0.22	.764
Familiarity [Unfamiliar]*Modulation Step [2]*Amount of Training[80]	0.06	-0.13 – 0.25	0.567

1 The reference categories are the 'Lab-Trained' condition for Familiarity, the 'Unshifted' condition for Modulation Step, and the 'Shorter Training' (20 stimuli) condition for the Amount of Training.

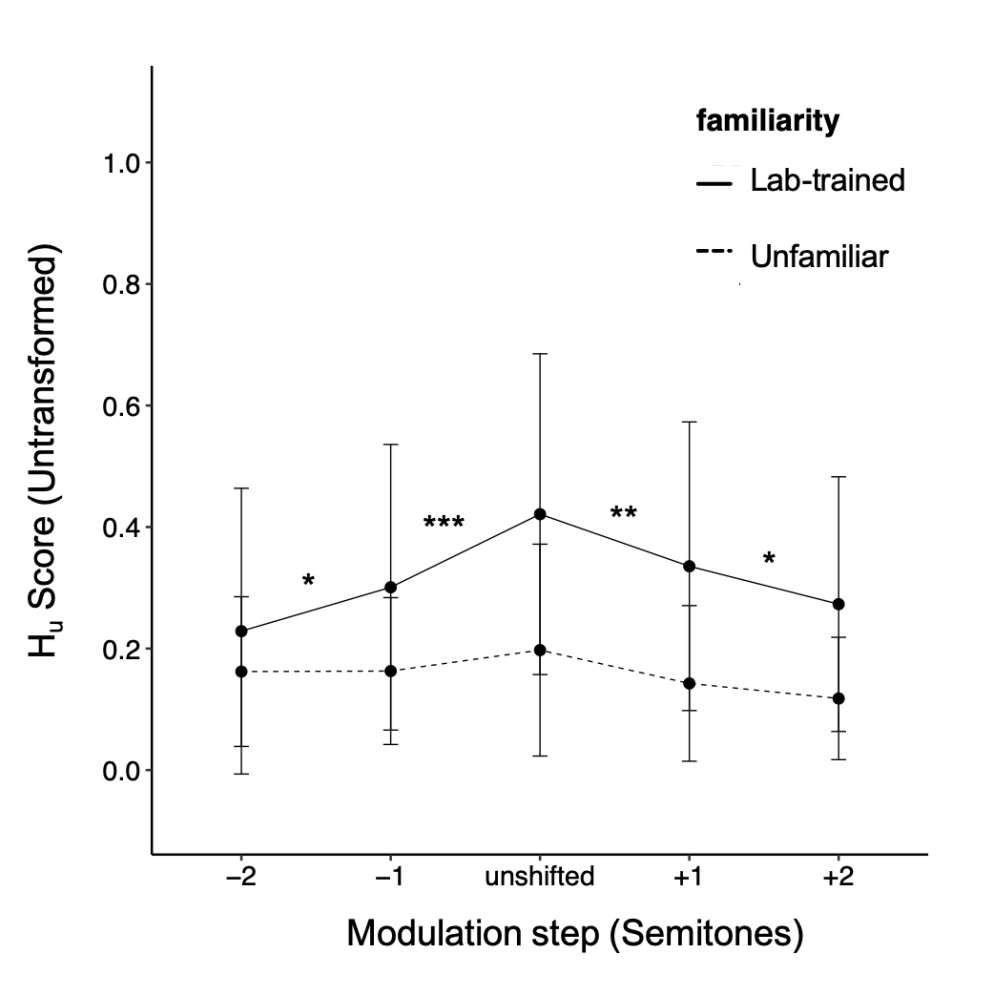

Figure 8. Mean Hu scores (untransformed) are displayed per familiarity condition (lab-trained and unfamiliar voices) and modulation step (x-axis). Error bars display standard deviations around the mean. Asterisks denote significance of pairwise comparisons between successive modulation steps. *** $p < .0001$, ** $p < .001$, * $p < .05$.

3.1.4 Interim Discussion

In this experiment, participants were split into two groups, receiving either shorter training (involving 20 voice excerpts per voice), or longer training (80 voice excerpts per voice) to learn to recognise two novel speakers. I found that both groups were able to recognise the speakers well at the end of training (shorter training: 77.03%; longer training: 81.1%). During the modulation task, listeners were required to attempt to recognise the speakers (2 lab-trained, 1 unfamiliar) from acoustically modulated and unmodulated voice recordings. Contrary to my predictions, I did not find any effect of the amount of training on performance in the voice modulation task, nor any interaction of training with any other factor in the experiment. I did

however, find that as the amount of modulation increased, there was a detrimental effect on accuracy for the lab-trained voices, but for unfamiliar voices, modulation had no effect on performance as performance was consistently poor across all modulation steps. But at each modulation step, recognition of the lab-trained and unfamiliar voices did not significantly differ. That is, within a single modulation step, recognition of the lab-trained voices was not significantly more accurate than the unfamiliar voice. This suggests that despite good accuracy during training, this did not translate to significant recognition benefits in the voice modulation task.

The finding that the amount of training did not reveal any differences in performance during the voice modulation task may suggest that the difference between the shorter and longer training was not large enough, or that the type of training did not amount to benefits in recognition. Performance for the lab-trained unmodulated voice tokens during the modulation task was reduced significantly from training (shorter training M = 77.03%, longer training M = 76.08%) to test (shorter training M = 52.26%, longer training M = 53.04%) in both groups. Secondly, evidence for a difference in recognition between the trained and unfamiliar voices at each modulation step was not observed. These two findings suggest that the task may have proved too difficult for both groups to perform. This is contrary to the findings in Experiment 2 in the previous chapter that did find significantly better performance for lab-trained voices compared to unfamiliar voices in both the Couples and Controls group. The challenge to perception after training was designed to be the acoustic voice modulation, however from training to test, another element also differed. During training, participants heard excerpts of spontaneous, conversational speech, whereas at test, they were presented with read sentences. Read sentences were originally chosen for the modulation task because these were easier to acoustically modulate without creating perceptible distortions. However, this change in speaking style from training to test may have been disruptive to voice recognition in and of itself. Earwitness research has shown that small changes to the voice, such as varying expressiveness of speech, or recognition from neutral to an angry tone of voice, is enough to disrupt recognition quite significantly in low to moderate familiar or lab-trained listeners (Saslove & Yarmey, 1980). Thus, participants had to contend both with a change in speaking style and acoustic modulation, and this may have concealed any potential effects of training. Overall, these considerations highlight that even after more substantial training based on 80 excerpts, voice representations may still be relatively unstable and susceptible to small variations in a speaker's voice.

To address the issue of changes in speaking style between familiarisation and test, a second experiment was run. In this experiment, read sentences were used during both the training and modulation task to keep speaking style consistent. The unfamiliar voice was also removed, and instead all participants were required to learn to recognise the same three voices. In Experiment 4, including an unfamiliar voice was useful to measure learning of both lab-trained voices, otherwise participants could have selectively learned only one lab-trained voice and "recognised" the other by elimination. However, recognition of the novel unfamiliar voice *per se* was never a key question of interest. A final modification for Experiment 5 was that the initial exposure to the target identities comprised two 7-excerpt sequences of speech from each speaker, with one of these sequences being presented twice (21 excerpts, 14 unique stimuli) rather than the two 12-excerpt sequences (24 stimuli), used in Experiment 4, due to a shortage of available recorded sentences to be used across all stages of this experiment.

3.2 Experiment 5

3.2.1 Methods

3.2.1.1 Participants

In Experiment 5, sixty-nine participants were recruited. Nine of these participants were removed for either failing to show sufficient attention to the tasks (7 participants) or performing below chance (+95% confidence interval) during the familiarisation training (2 participants). This left sixty participants in the final sample in total (34 female, mean age = 26.15 years, SD = 6.21 years, age range = 18-40 years). Half of the participants were assigned to the shorter 20 stimuli training condition, and the other half assigned to the longer 80 stimuli training condition. None of the participants had taken part in Experiment 4, and all participants learned to recognise the same three female voices.

3.2.1.2 Materials

Read sentences (e.g. "The beach sold bats and balls") were extracted from a single set of three female speakers taken from the freely-available LUCID corpus (Baker & Hazan, 2011), to be learned by all participants. In Experiment 5, read sentences were used across both training and

test phases so that speaking style would be held constant and participants would only have to contend with acoustic modulation at test.

As in Experiment 4, familiarisation of the three new voices involved a passive exposure phase and training phase. For passive exposure to the identities, read sentences were arranged into two 7-excerpt sequences per speaker, with each sound clip separated by 1s of silence. One of the 7-excerpt sequences was presented twice, such that the participant heard 21 (14 unique) voice excerpts. Repeats of stimuli were placed in the passive exposure phase because there were 144 recorded sentences available per speaker. Thus, to ensure no sentences were repeated from training to test, and so that participants would have a similar amount of exposure to each voice in Experiments 4 and 5, repeats were included here. For the training phase, 80 sentences were selected per the to-be-learned identities (either 20 or 80 of these were used depending on the training condition).

For the voice modulation test phase, due to constraints on available stimuli, 50 sentences were chosen (10 per the five modulation conditions), as well as each voice modulation condition having 2 stimuli that were repeats of sentences in other modulation conditions (10 repeats total). This meant that there were 60 stimuli (12 excerpts x 5 conditions) per voice, and a total of 180 stimuli in this task.

The same vigilance stimuli as in Experiment 4 were used in Experiment 5.

3.2.1.3 Procedure

In Experiment 5, there were two versions of this task (longer or shorter training) and participants were randomly assigned to one of these versions. The procedure was the same as in Experiment 4, with the only differences being that passive exposure involved three 7-excerpt sequences per the three to-be-learned speakers, and the text response options in both the familiarisation training and the modulation task were “Anna”, “Beth”, and “Clara”, reflecting the three lab-trained voices.

3.2.1.4 Data Analysis

The data for the familiarisation training and modulation task were analysed in the same way as for Experiment 4.

3.2.2 Results

3.2.2.1 Familiarisation training

At the end of training, participants could recognise the three new speakers reasonably well in both the shorter training group (mean = 73.83%, SD = 12.02), and longer training group (mean = 79.39%, SD = 14.49; see Figure 9 for mean accuracy across training group and block).

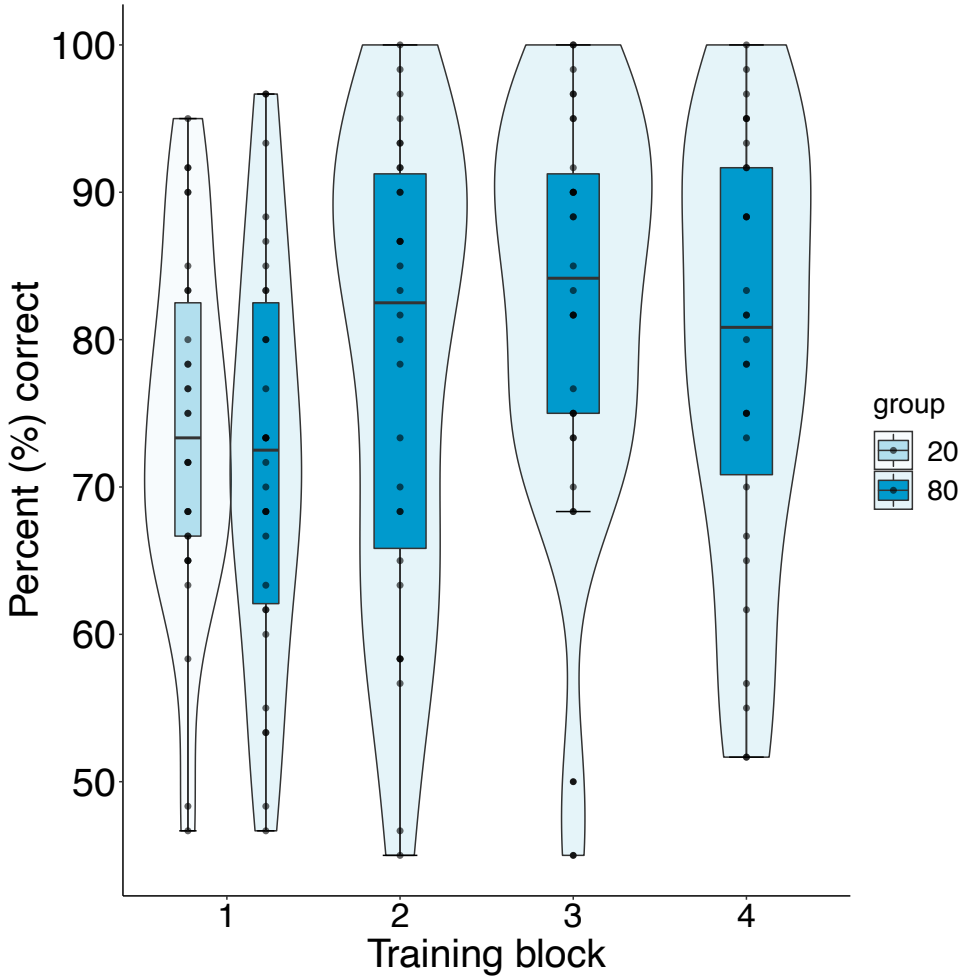

Figure 9. Box plots display median percent correct scores per training group and training block. Boxes range from the first to third quartiles and whiskers extend to no further than 1.5*interquartile range above and below the 1st and 3rd quartiles. Individual participants' mean performance is displayed as individual points. Note that the shorter training group only performed one block of training.

3.2.2.2 Modulation task

Averaging across all modulation steps per training group showed that the mean overall performance for listeners that received shorter training (20 stimuli) was 42.19% (SD = 4.68%), with mean scores on the individual modulation steps ranging over 37.78% - 49.17%. The mean overall performance for those in the longer training condition (80 stimuli) was 49.44% (SD = 7.48%), with mean scores on the individual modulation steps ranging over 42.50% - 60.56%.

To compare the effect of acoustic modulation and amount of training on the recognition of the three voices, I analysed the interaction between amount of training (20 vs. 80 stimuli) and acoustic modulation (5 modulation steps) using LMMs.

The model was set up as follows, where the Hu score (unbiased hit rate) is the outcome measure for recognition performance:

*lmer(hu score ~ Amount of Training*Modulation Step + Sex + (1|Participant) + (1|Speaker Identity))*

Modulation step, amount of training, and sex of the participant were included as fixed effects; participant and speaker identity were included as random effects. Statistical significance was obtained by comparing the full model that contained the interaction, fixed effects, and random effects to a reduced model that contained all of the same fixed and random effects but did not contain the interaction.

Using this method, it was found that there was no significant interaction between the amount of training a participant received and modulation step on recognition performance ($\chi^2(4) = 7.10$, $p = .130$). There was, however, a main effect of modulation step ($\chi^2(4) = 80.0$, $p < .0001$) on performance across both groups. Post-hoc pairwise comparisons (FDR-corrected) comparing performance at successive modulation steps (i.e. unshifted vs. 1 step, 1 step vs. 2 steps) highlighted that accuracy decreased with increasing distance from the original unshifted voice (all ps $< .05$). I also observed a main effect of amount of training ($\chi^2(1) = 7.02$, $p = .008$). Participants that received longer training (80 stimuli per voice) displayed significantly higher

accuracy (mean Hu score = 0.54, SD = 0.25) compared to those that received shorter training (20 stimuli per voice; mean Hu score = 0.46, SD = 0.24; see Figure 10).

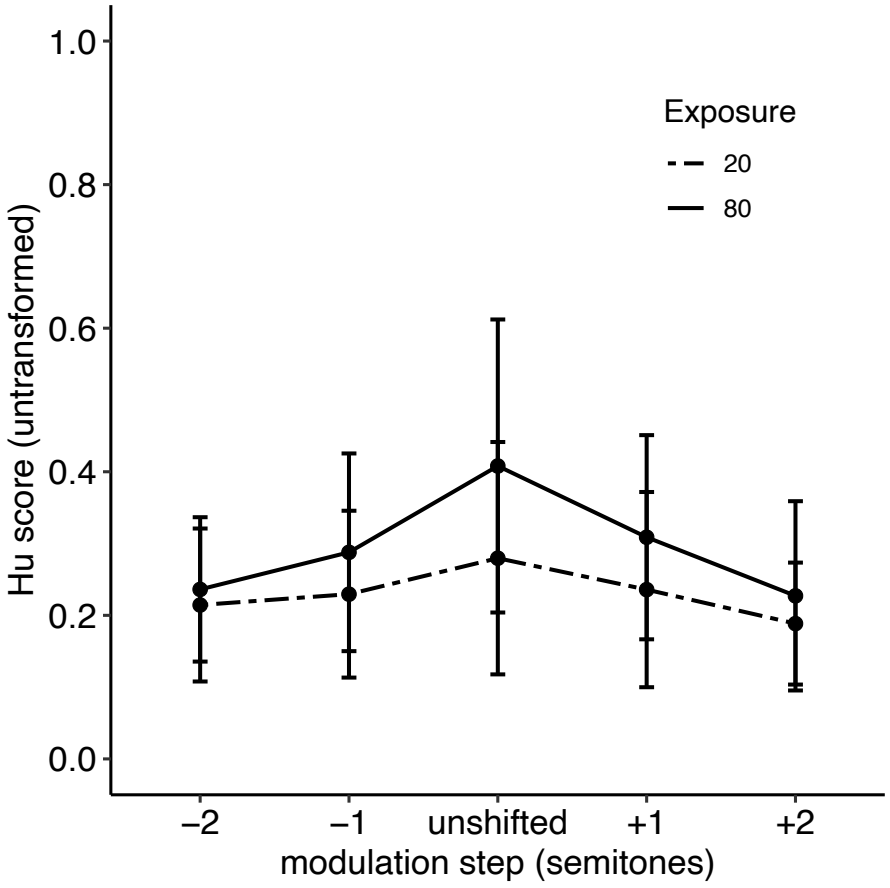

Figure 10. Mean Hu scores (untransformed) are displayed per exposure condition and modulation step (x-axis). Error bars display standard deviations around the mean.

3.2.3 Exploratory analyses

To gain some insight into the types of categorisation errors participants were making or potential biases in responding, I constructed confusion matrices per modulation step collapsed across both groups (see Figure 11). These confusion matrices revealed an interesting pattern of results. For sentences that had been shifted in the negative direction (steps: -1, -2; i.e. lower GPR, longer VTL), participants appeared to show a tendency towards responding "Anna". Conversely, for sentences shifted in the positive direction (i.e. higher GPR, shorter VTL), participants showed a tendency towards responding "Clara", with "Beth" responses being more evenly spread across the different modulation steps.

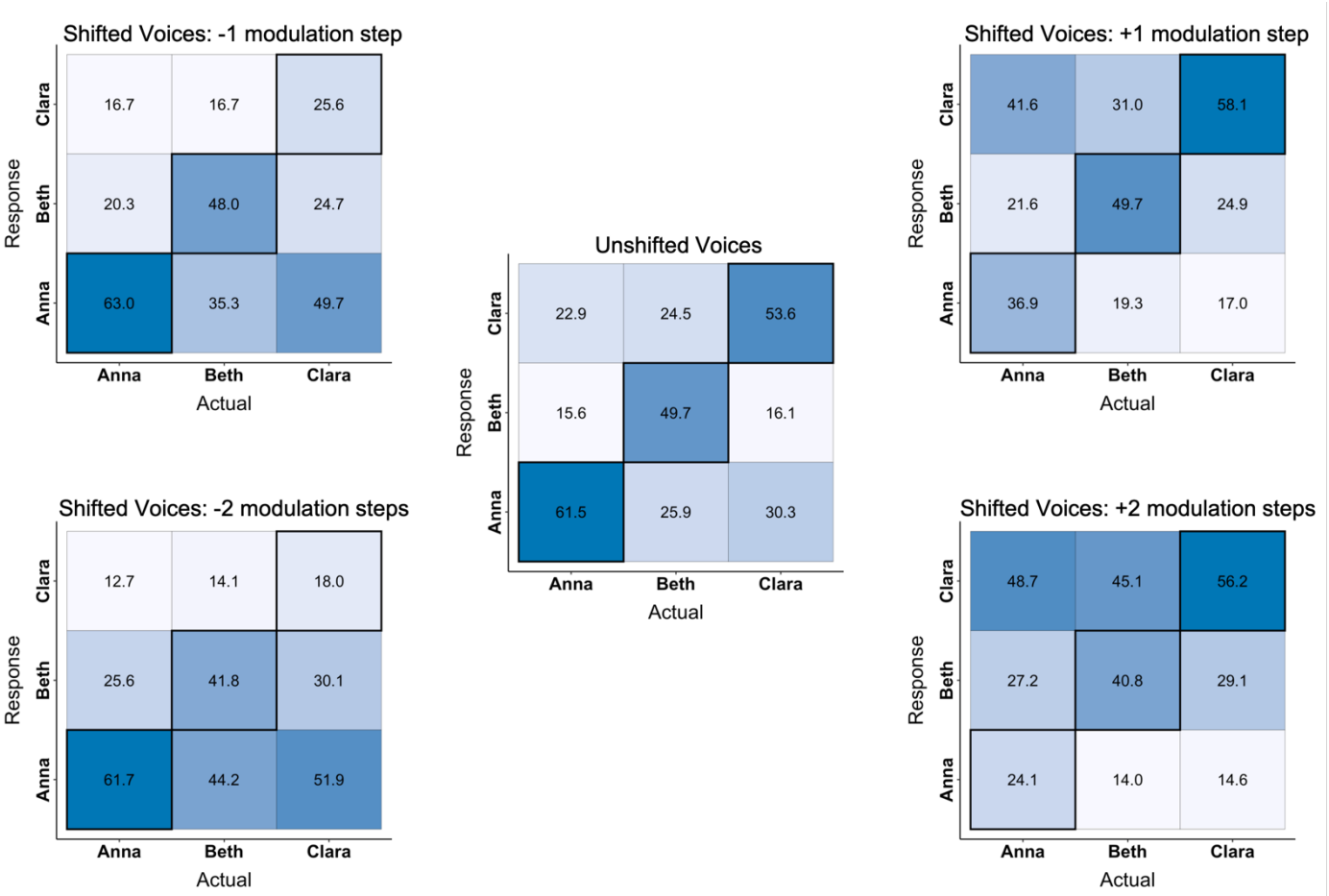

Figure 11. Confusion matrices displaying participants' responses for the recognition of acoustically modulated voices (both training groups). Matrices are shown for each modulation step: hits, misses, and false alarms for the three identities when these voices had been modulated by 1 step in both the negative (top left) and positive direction (top right); when these voices had been modulated by 2 steps in each direction (bottom left and right), and for unshifted tokens (centre).

Moreover, Cohen's Kappa was calculated at each modulation step for each training group (20 vs. 80). Kappa compares observed accuracy with expected accuracy (random chance; Cohen, 1960). The Kappa value indicated that performance was better the closer the voice excerpts were to the original unshifted versions of the speakers' voices, and that accuracy was better in the longer training group (80) across all modulation steps (see Table 4). Performance for the longer training condition ranged from moderate to slight, whereas performance ranged from fair to slight in the shorter training group (20). However, accuracy was above the no information rate, or chance performance across all conditions (all ps < .005).

Table 4. Displays Cohen's Kappa (K) and significance (p) of the comparison between observed and expected accuracy (random chance) in each modulation step and by training group.

Modulation Step	Group: 20		Group: 80	
	K	*p*	K	*p*
-2	.075	< .001	.139	<.0001
-1	.126	<.0001	.240	<.0001
Unshifted	.238	<.0001	.410	<.0001
+1	.161	<.0001	.286	<.0001
+2	.068	.001	.144	<.0001

Performance using the Kappa statistic can be classed as 0-0.20 (slight), 0.21-0.40 (fair), 0.41 – 0.60 (moderate), 0.61-0.80 (substantial), or 0.81-1 (almost perfect).

3.2.3.1 Examining the acoustic properties of the lab-trained voices

The patterns revealed by the confusion matrices raised the possibility that listeners may have been using a particular acoustic strategy to classify the three voices, in which the feminised voices (+1 and +2 modulation steps) were more likely to be labelled "Clara" and the masculinised voices (-1 and -2 modulation steps) more likely to be labelled as "Anna". To investigate this possibility, I first calculated the average F0 of each of the three lab-trained voices participants would have heard in the exposure and familiarisation training. Both GPR and VTL have been found to be important indicators of identity to listeners (Gaudrain, Li, Ban, & Patterson, 2009). Thus, both F0 and apparent vocal tract length (aVTL) were calculated to explore whether these might explain the apparent biases in participants' responses. It was revealed that the original "Anna" voice tokens used in the initial exposure and familiarisation

training had a mean F0 of 230.8Hz (range = 178.6 – 310.2Hz) and aVTL of 12.58cm (range = 10.93-14.26), "Beth" tokens had a mean F0 of 243.6Hz (range = 195.9 – 364.2 Hz) and aVTL of 12.65cm (range = 11.68-13.77), and "Clara" tokens had a mean F0 of 263.6Hz (range = 171.9 – 370.0Hz), aVTL =12.64cm (range = 11.28 – 13.93). Thus, "Anna" tokens had the lowest mean F0, and "Clara" tokens had the highest, which mirrored the apparent biases observed in the constructed confusion matrices. In contrast, mean aVTL was more similar across the three voices.

To calculate the acoustic measures, including both F0 and aVTL, I used VoiceLab (Feinberg & Cook, 2020), which is an automated voice analysis software. Apparent VTL was estimated in VoiceLab using Fitch's method below

$$\frac{\sum_{i=1}^{n}(2n-1)\frac{f_i}{4c}}{n}$$

which involves using the formant dispersion to estimate vocal tract length, where n is the number of formants measured (the first four formants were used here), f_i is the frequency (in Hz) of formant i, and c is the speed of sound in air (Reby & McComb, 2003; Fitch, 1997). If the larynx is lowered (vocal tract lengthened), all formant frequencies should lower, leading to a decrease in the spacing between formants. This has been found to correlate with vocal tract length and body size (Lammert & Narayanan, 2015; Fitch, 1997).

3.2.3.2 Estimating the effects of F0 and aVTL on voice identity categorisation:

In previous research, F0 and aVTL have been noted to be important acoustic measures for recognition of voices (Gaudrain et al., 2009), and it may be that this is heightened for voices that are newly familiar or not well known. As two acoustic parameters were modulated in this task, it is worth examining whether participants show reliance on these measures despite them no longer being stable indicators of identity. Binomial generalised linear mixed models (GLMMs) were constructed to explore whether there were effects of F0 and aVTL on participant responses for the three voices, and how the amount of training may affect this relationship. Separate models were constructed for each of the three voices. Inference is based on iteratively comparing full and reduced models following Type III sums of squares. Odds ratios (ORs) and confidence intervals (CIs) are reported. An odds ratio of 1 means that no effect is present. The further an odds ratio deviates from 1, the larger the size of the effect. Confidence intervals that do not cross one indicate significant effects. As F0 was measured in Hz and

apparent vocal tract length in cm, these data were transformed into z-scores to keep the acoustic measures on the same scale.

Taking the "Anna" voice as an example, the model was set up as follows:

*glmer(Anna Response ~ F0_z*aVTL_z*Amount of Training+ Sex + (1|Participant) + (1|Correct Answer))*

where "Anna Response" is a binary outcome variable with 1 meaning that the participant selected "Anna" as the speaker on that particular trial, and 0 meaning the participant did not select "Anna" as the person they thought was speaking. F0_z was the F0 of each voice token in the modulation task (transformed into a z-score), aVTL_z was the apparent vocal tract length of each voice token (transformed into a z-score), and "amount of training" was the training group the participant was in; shorter training (20 stimuli) or longer training (80 stimuli). Participant and the actual correct answer on each trial were included as random effects. Separate models were constructed for each of the three voices ("Anna", "Beth", and "Clara").

Results

3.2.3.2.1 "Anna" responses:

Comparing a full model that contained the three-way interaction between F0, aVTL, and amount of training to a model that contained all two-way interactions, fixed and random effects but not the three-way interaction showed that there was not a significant interaction between F0, aVTL, and amount of training on participants' likelihood of making an "Anna" response (E=0.02, OR=1.02, CI =0.91 – 1.16, SE=0.06, Z=0.37, p = .713). This means that the interaction effect between F0 and aVTL on making an "Anna" response on any given trial did not differ significantly depending on the amount of training participants received. I did, however, find a significant two-way interaction between the amount of training and F0 (E = 0.32, OR =1.38, CI = 1.23 – 1.55, SE = 0.06, Z = 5.53, p < .001; see Figure 12). In both groups, there was a negative relationship between F0 and the likelihood of making an "Anna" response, however, this was 1.38 times greater in the group that received less training (20-exposure) compared to those that received more training (80-exposure). Specifically, participants that received less training were 2.63 times less likely to respond "Anna" with every one-point increase in F0, whereas those that received more training were 1.92 times less likely to make

an “Anna” response as F0 increased. There was no significant interaction between aVTL and amount of training (E = 0.005, OR =1.01, CI= 0.87 – 1.16, SE = 0.07, Z =0.06, p = 0.94), nor a significant interaction between F0 and aVTL on “Anna” responses (E =0.005, OR =1.01, CI = 0.92 – 1.10, SE = 0.05, Z = 0.12, p = 0.905). There was, however, a significant main effect of aVTL on “Anna” responses (E = -0.20, OR = 0.82, CI = 0.74 – 0.91, SE = 0.06, Z = -3.58, p < .001). This was also a negative relationship. With each one-point increase in aVTL, listeners in both groups were 1.21 times less likely to respond “Anna”.

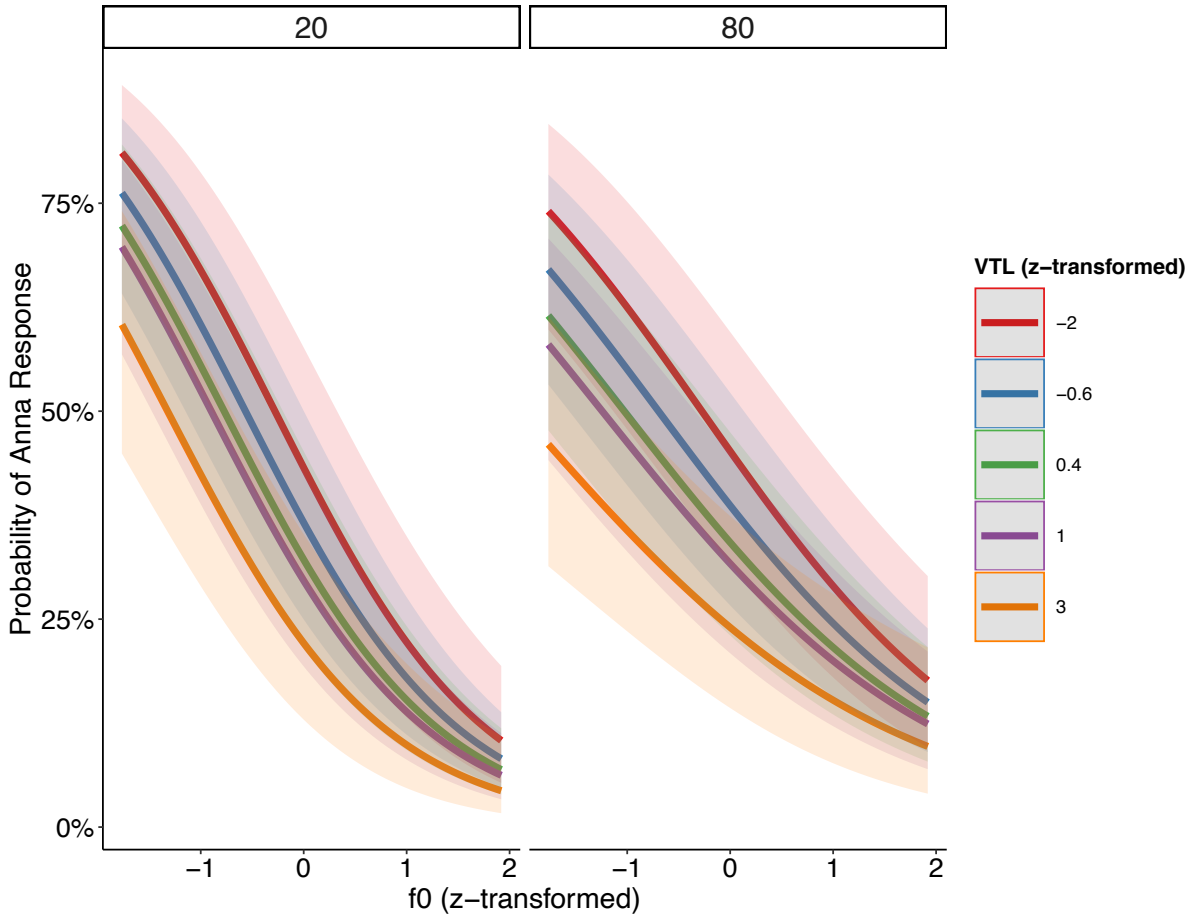

Figure 12. Figure shows the relationship between fundamental frequency (F0) and apparent vocal tract length (aVTL) on the probability of making an “Anna” response. The relationship for each training group is displayed separately: left: participants in the 20-exposure condition, right: participants in the 80-exposure condition.

3.2.3.2.2 “Clara” responses:

Using the same method as for the "Anna" responses, there was a significant three-way interaction between F0, aVTL, and amount of training received on "Clara" responses (E = -0.16, OR = 0.85, CI = 0.75-0.97, SE = 0.06, Z = -2.49, p = .01). This result means that increases in vocal tract length are 1.17 times more likely to amplify effects of increasing F0 in making a "Clara" response in participants that received less training (20-exposure condition) compared to those that received more training. To break this down, the interaction between F0 and aVTL was significant for participants who received more training (80-exposure; (E = -0.09, OR = 0.91, CI = 0.84-0.99, SE = 0.04, Z = -2.13, p = .03) and highlighted that the effect of F0 on making a "Clara" response was reduced by 9% (1-0.91) with increasing vocal tract length. In contrast, for participants that received less training (20-exposure condition), vocal tract length had no significant influence on F0 on choosing "Clara" (E = 0.07, OR =1.07, CI = 0.98-1.18, SE = 0.05, Z = 1.44, p = .149; see Figure 13). In other words, participants that received more training began to rely on cues from vocal tract length to assist their decision-making as F0 increased, reducing their sole reliance on cues from F0. This remained secondary to F0, however, as vocal tract length independent of F0 (main effect or across exposures) never showed a significant effect (E = -0.14, OR = 0.87, CI = 0.76 – 1.00, SE = 0.07, Z = -1.95, p > .05).

There was also a significant interaction between F0 and the amount of training participants received (E = 0.15, OR = 1.17, CI = 1.05-1.30, SE = 0.06, Z = 2.80, p = .005; see Figure 13). The direction of the effect was positive (in contrast to the relationship between F0 and "Anna" responses), and revealed that as F0 increased, participants were 1.17 times more likely to make a "Clara" response for participants that received less training (20-exposure) compared to those that received more training (80-exposure). Participants that received less training were 2.30 times more likely to respond "Clara" as F0 increased, whereas those that received more training were 1.96 times more likely to make a "Clara" response as F0 increased. In other words, F0 had a greater influence on "Clara" decisions for those that had received less training.

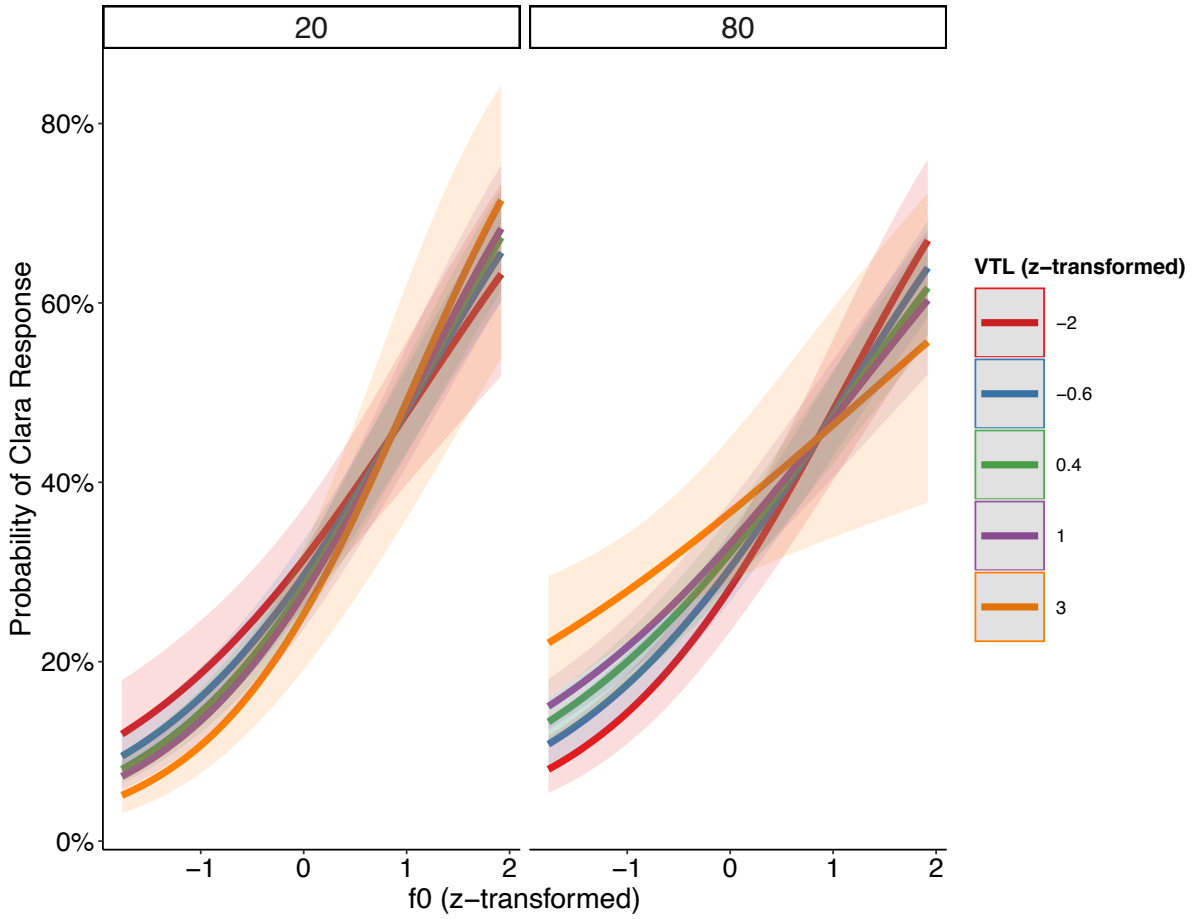

Figure 13. Figure shows the relationship between fundamental frequency (F0) and apparent vocal tract length (aVTL) on the probability of making a "Clara" response. The relationship for each training group is displayed separately: left: responses from participants in the 20-exposure condition, right: responses from participants in the 80-exposure condition.

3.2.3.2.3 "Beth" responses

For "Beth" responses, I did not observe a significant three-way interaction between F0, aVTL, and amount of training (E = -0.03, OR =0.97, CI = 0.86-1.09, SE = 0.06, Z = -0.56, p = .574; see Figure 14). There was a significant interaction between F0 and amount of training received, however (E = 0.16, OR = 1.18, CI = 1.06 – 1.31, SE =0.05, Z = 3.00, p = .003). Specifically, the influence of F0 on making a Beth response was 1.18 times greater for those who had less training, compared to those that had more training. Looking at the effect of F0 separately in the two groups, I observed that F0 was not a significant predictor of Beth responses in the group that received more training (E = -0.06, OR = 0.94, CI = 0.87-1.02, SE = 0.04, Z = -1.46, p = .145), however in participants that received less training, participants were 1.11 times more likely to respond Beth with each 1 unit increase in F0 (E = 0.10, OR = 1.11, CI = 1.03 – 1.20,

SE = 0.04, Z = 2.69, p = .007). A significant interaction was also observed between F0 and aVTL for participants that received longer training (E = 0.10, OR = 1.10, CI = 1.01-1.20, SE = 0.04, Z = 2.14, p =.032), with the influence of F0 on making a "Beth" response being 1.10 times greater with a longer vocal tract (stronger relationship between F0 and Beth responses at longer aVTLs). I did not observe this interaction for participants in the shorter training group (E = 0.06, OR = 1.06, CI = 0.98 – 1.15, SE = 0.04, Z = 1.43, p = .152). Again, I did not observe an interaction between VTL and amount of training (E = 0.05, OR = 1.05, CI = 0.91 – 1.22, SE = 0.07, Z = 0.71, p = .479).

There was a significant main effect of aVTL on making a "Beth" response (E = 0.16, OR = 1.18, CI = 1.07-1.30, SE = 0.05, Z = 3.20, p < .01). For participants that received shorter training (20-exposure), participants were 1.18 times more likely to respond "Beth" with each one unit increase in aVTL, whereas those that received longer training (80-exposure) were 1.12 times more likely to respond "Beth", but this effect of aVTL on "Beth" responses did not significantly differ between the two training groups.

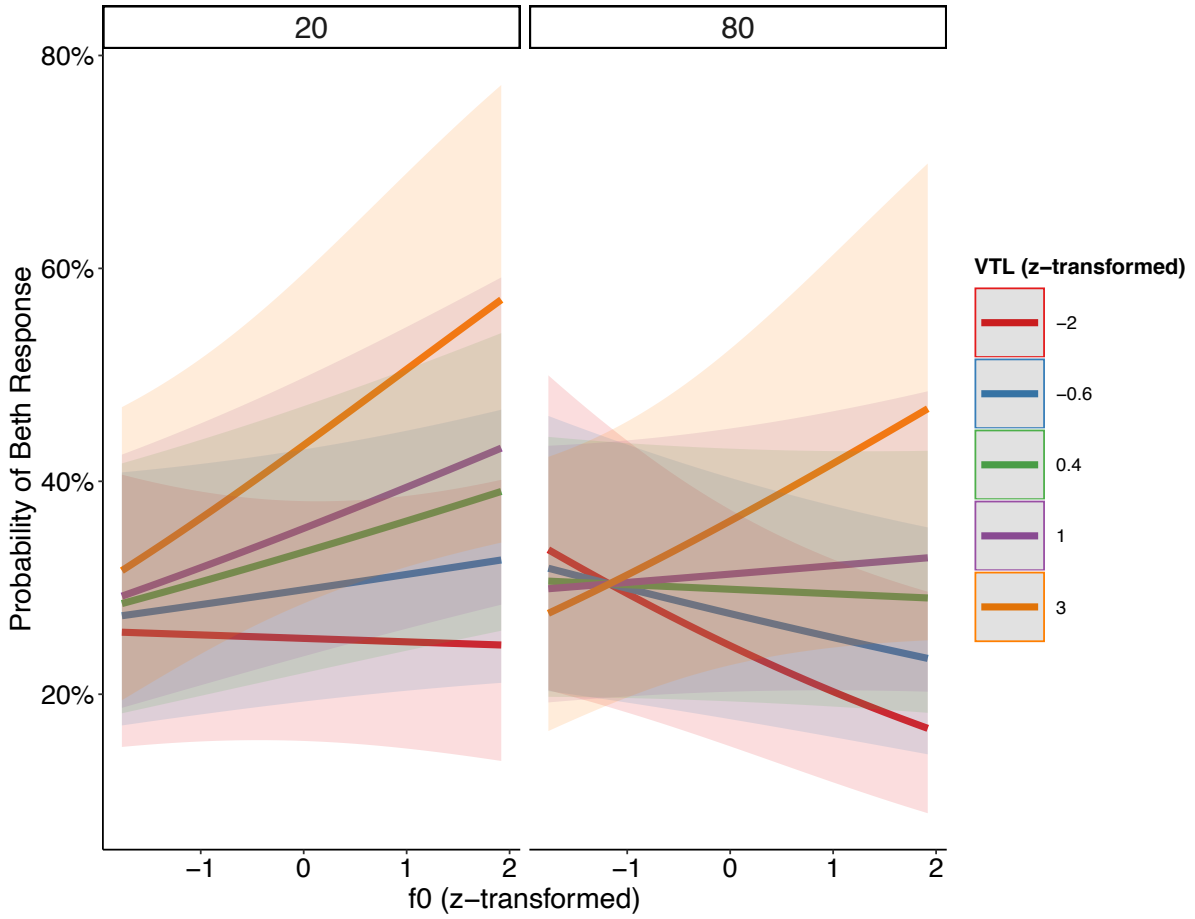

Figure 14. Figure shows the relationship between fundamental frequency (F0) and apparent vocal tract length (aVTL) on the probability of making a "Beth" response. The relationship for each training group is displayed separately: left: participants in the 20-exposure condition, right: participants in the 80-exposure condition.

3.2.4 Interim Discussion

This experiment again explored the impact of the amount of voice recognition training on listeners' ability to recognize voices in a perceptually challenging recognition task. However, here, speaking style was held consistent from training to test such that listeners should only have to contend with acoustic modulation to the vocal stimuli, rather than having to also contend with differences in speaking style – the latter of which has been found to disrupt recognition for low familiar listeners. At the end of training, participants were able to recognize the three speakers with good accuracy on average (shorter training = 73.83%, longer training = 79.39%). In the voice modulation task, I found that participants who received longer training showed overall significantly better recognition accuracy in the voice modulation task compared to those that received shorter training. I also found that in both groups, as distance increased

from the unshifted voice, performance was negatively impacted. That is, with each increasing modulation step away from the unshifted voice, there was a significant decrease in performance. However, an interaction between the amount of training and modulation step was not observed. These results suggest that more training was sufficient to improve listeners' ability to recognise the speakers through generalisation to novel excerpts, but that this longer training did not lead to a steeper tuning function or a more conservative response bias.

Based on the observed confusion data from this experiment, I also examined the acoustic properties of the lab-trained voices, and whether there were any relationships between these properties and the ways in which listeners were categorising the vocal stimuli in the voice modulation task. Broadly, I observed that there were significant relationships between F0 and the likelihood of categorising excerpts as particular identities. In particular, for "Anna" responses, this was a negative relationship (i.e. as F0 increased, the likelihood of responding "Anna" decreased), whereas this was a positive relationship for "Clara" voices. The relationship between F0 and responses were consistently stronger for those that received shorter training, compared to those that received longer training, suggesting an increased reliance on this particular acoustic cue with less training. The relationship between apparent vocal tract length (aVTL) and participants' responses was less clear cut. There was a negative relationship between aVTL and "Anna" responses, a positive relationship for "Beth" responses, and no significant effect of aVTL on making a "Clara" response. I did not, however, find any interactions between aVTL and training group. These results are discussed further in the general discussion.

3.3 General Discussion

In the two experiments described in this chapter, I examined how voice recognition from lab-trained voices is affected by the amount of training given to listeners. In particular, recognition was probed using acoustically modulated voice excerpts to examine the effect of training on generalisation ability and whether "tuning" to the specific features of the learned voices was sharpened as a result of training. In Experiment 4, listeners could be trained to recognise two lab-trained voices with good accuracy after encountering as little as 20 voice tokens per identity during training. At test, no significant differences based on the amount of training participants received were observed, and no clear overall recognition benefits for lab-trained voices. It was

concluded that any potential differences due to training may have been obscured because of the change in speaking style of voice excerpts from training to test. Thus, in Experiment 5, speaking style was kept consistent. Here, those with more training showed better performance in the voice modulation task overall, as well as there being an effect of acoustic modulation, with performance decreasing with increasing modulation step as was expected. No interaction was observed between the amount of training and modulation step, and therefore a sharpening/narrowing of the tuning to the learned voices was not detected.

In Experiment 5, listeners who had completed 80 trials of familiarisation training per voice showed significantly greater voice recognition accuracy at test than listeners with only 20 trials of training per voice. Prominent models of voice identity processing suggest that for unfamiliar voices to become familiar, reference patterns must be established (Maguinness, Roswandowitz, & von Kriegstein, 2018). Maguinness, Roswandowitz, and von Kriegstein (2018) proposed that reference patterns are refined with continued exposure to a voice. Thus, in the current experiment increasing the amount of training perhaps allowed for better established reference patterns to be formed over time, allowing for improved recognition performance in a challenging listening context. This model, however, does not acknowledge the importance of the type of exposure that may be important for building stable voice representations. Indeed, in Experiment 4, increased training did not result in an improvement in performance. This was due to a change in speaking style from training to test that may have offset any positive training effects on recognition.

Although significant improvements were observed with more training in Experiment 5, performance was still far from perfect and sharper "tuning" to the specific features of the learned voices was not observed in the form of an interaction between modulation step and training group. The voice modulation task requires listeners to generalise beyond what was learned during training, and I would argue that on the whole, both groups failed to do this well. Average accuracy across all modulation steps for both groups was relatively poor (20-exposure: 42.19%; 80-exposure: 49.44%). Thus, although increasing the amount of training led to improvements in the ability to generalise, it is unlikely that hearing 80 stimuli during training was enough to build up a representation that is robust enough to handle perceptual challenges, as was observed with personally familiar voices in the previous chapter. The type of experience with a voice is also highly likely to be important for building up representations that are robust and flexible (Lavan, Knight, et al., 2019b). For instance, a study by Lavan,

Knight and colleagues (2019b) trained participants to recognise voices, manipulating the types of vocal stimuli heard during training. High-variability training involved including voice recordings from a number of recording sessions, speaking styles, and settings. Low-variability training sets comprised recordings in one speaking style. In an old/new recognition task, high-variability training proved advantageous but only when listeners were required to generalise to novel test stimuli that differed in speaking style/variability.

In the experiments in this chapter, participants were trained to recognise the speakers from speech produced in a single speaking style (natural conversational excerpts in Experiment 4 and read sentences in Experiment 5). This may have meant that mental representations or reference patterns for these voices were rather rigid, making generalisation difficult for both groups alike. In the previous chapter, performance in the voice modulation task for the lab-trained voices in the Controls group was 51.6% accurate on average (collapsed across both lab-trained voices). This is similar to averaged performance in the voice modulation task in Experiment 4 (shorter training M = 44.90%, longer training M = 44.76%) and Experiment 5 (shorter training M = 42.19%, longer training M = 49.44%). This is in comparison to average performance. Therefore, even though speaking style was kept consistent in Experiment 5, this did not appear to improve performance in terms of overall percentage correct in comparison to Experiment 4. This suggests that perhaps the need for generalisation was not the only factor affecting performance in the voice modulation task. It may be that the inclusion of variably modulated voice excerpts from multiple speakers may have caused a high degree of confusion for recognising lab-trained voices generally, where these voices are expected to have weaker stored representations. Unstable representations may be more prone to interference from modulated voice excerpts, which could lead to listeners forgetting what the original voices sounded like. These results are in comparison to average performance for the personally familiar voices that were on average recognised with 79.5% accuracy across all modulation steps. Taken together, the results from the current and previous chapter highlight that even with more training, gains to recognition are small and significant, however these gains are not comparable to those observed for the recognition of voices that we know personally.

In contrast, though performance was overall poor for both groups in Experiment 5, and despite not observing sharper tuning, the significant improvement in recognition ability from shorter to longer training is compelling. This is especially pertinent as the duration of the longer training condition was approximately 15 minutes longer compared to the shorter training

condition (~10 mins vs. ~25 mins). This may suggest that the acquisition and establishment of voice representations may be a relatively rapid process. Perhaps basic stored representations are established fairly quickly, but it is the fine-tuning of these representations that could potentially be a slower process. Lavan, Knight, and colleagues (2019b) argued that once an initial voice representation has been formed, exposure to high variability might further aid in the formation of more robust representations. In my exploratory analyses investigating potential relationships between acoustic measures (F0 and aVTL) and likelihood of choosing certain response options revealed that F0 was the strongest predictor of participants' identity responses. For "Anna", a clear negative relationship between F0 and likelihood of choosing "Anna" was observed, meaning that as F0 was experimentally increased, participants were less likely to respond "Anna". For "Clara", the opposite was observed. As F0 increased, the likelihood of responding "Clara" also increased. These results are consistent with the relative differences between the three original versions of these speakers' voices, with "Anna" being perceptually the "lowest" sounding, and "Clara" the "highest".

Uncovering the combination of cues implicated in voice recognition is a complex and challenging task. A number of previous studies have sought to identify a fixed set of acoustic cues involved in this process (e.g. Lavner, Gath, & Rosenhouse, 2000; Gaudrain et al., 2009; Murry & Singh, 1980; Schweinberger, 2001). For instance, one study manipulated various acoustic properties (e.g. pitch, glottal waveform, and formants) of recordings of a single vowel /a/, testing listeners' recognition abilities for personally familiar voices (other members of a Kibbutz in which participants lived; Lavner, Gath, & Rosenhouse, 2000). This study found that on average, vocal tract features were more important than glottal source features (e.g. mean F0). However, the manipulations affected recognition of each voice differently, suggesting that the cues used for recognition vary depending on the speaker. Similarly, other research using professional impersonators also identified vocal tract features as being the most prominent for voice recognition (López et al., 2013). Using unfamiliar voices in a discrimination task, Gaudrain and colleagues (2009) studied the relationship between two acoustic features (glottal pulse rate and vocal tract length) and perceived speaker similarity. Consonant-Vowel voice recordings (e.g. /ba/) from a single speaker were acoustically modulated using the STRAIGHT software (Kawahara & Irino, 2004) and listeners made judgements about whether two voice excerpts could have been plausibly produced by the same speaker. The researchers again found that vocal tract parameters could be altered to a lesser extent (compared to GPR) before

participants no longer perceived these excerpts to be produced by the same speaker, suggesting that vocal tract properties may be more diagnostic of differences in vocal identity.

In Experiment 5, one acoustic feature – average F0 – was found to be most influential for distinguishing the three identities, whereas aVTL was not. However, this does not necessarily contradict the previous research finding that vocal tract features are most important for voice recognition. In the current experiment, average vocal tract length was similar across all three voices, and thus not particularly diagnostic in this case. Additionally, aVTL was calculated from the formant frequencies which will differ based on the particular sentences being uttered. This means that this approximation of vocal tract length is noisy and therefore results should be interpreted cautiously. Average F0 on the other hand varied across the three speakers in a way that was seemingly perceptible for listeners. Interestingly, the range in F0 across voice excerpts within each speaker was quite large, and there was considerable overlap in these ranges across speakers. The finding that listeners were relying on mean F0 suggests that listeners may have been able to automatically extract the average F0 information from each voice and subsequently use this in their decision-making in the modulation task. This supports previous research that found that listeners unconsciously create average-based representations of individual voice identities (Lavan, Knight, & McGettigan, 2019a).

Conclusions about the relative importance of certain acoustic features cannot readily be drawn from my findings as this experiment did not set out to systematically test this. Instead, what was observed was that average F0 may have been useful during training to distinguish the three identities. This strategy appears to have been subsequently applied to the voice modulation task. Importantly, across all identities, this influence of F0 on the likelihood of making either an “Anna”, “Beth”, or “Clara” response was greater for participants that received less training, compared to those that received more training. When perceptually salient cues to recognition such as F0 and aVTL are unavailable or modulated, listeners may rely on other available cues such as speech rate or accent information (Maguinness, Roswandowitz, & von Kriegstein, 2018). In the voice modulation task, apparent GPR and VTL were altered and so these cues were no longer reliable indicators of identity. However, both groups were still using this information. For accurate performance in this task, listeners needed to rely not on these acoustic features, but either needed to make use of residual diagnostic voice information such as differences in pronunciation or speech rate, or to partially generalise from stored representations to these novel voice stimuli. With more training, listeners were better able to

reduce their reliance on these cues, which in turn seemingly contributed to their greater recognition accuracy in the voice modulation task.

It should be noted that participants were not explicitly told which cues had been modulated during test, yet those in the longer training condition demonstrated less reliance on these cues. This supports the idea that with increasing familiarity, the ability for flexible recognition improves. This result also sheds some light onto how recognition of voices may change over time with increasing familiarity, and allows for inferences to be made about the potential speed with which representations are established and built upon. In order to understand more about how a reliance on salient acoustic cues might diminish with more training/experience with a voice, future work is needed. This research could manipulate both the amount of training listeners receive whilst also systematically varying how diagnostic acoustic cues are for recognition during training, and how this affects categorisation of these same voices at test. This would allow for an exploration of which cues listeners rely on preferentially, as well as whether increasing the amount of experience one has with a voice can subsequently reduce the reliance on these acoustic features.

However, even if, in the context of an experiment, listeners rely more on certain acoustic cues with less experience with a voice, this may not translate into a clear understanding of real-life voice identification. The type and number of diagnostic or salient cues to speaker identity are likely to vary considerably, depending on the task demands and the listening context, such as the voices they are being perceived alongside. For instance, in Experiment 5, F0 was more diagnostic than aVTL due to the natural properties of the voices in comparison with each other during training. Moreover, many studies find that vocal tract parameters are the most indicative of voice identity, but there is no universally accepted set of acoustic features that could be used to recognise any speaker. Many studies find that different combinations of acoustic features are relevant for recognising different speakers (Lavner, Rosenhouse, & Gath, 2001). Thus, it has been proposed that familiar voices may be encoded as a "Gestalt-like" complex pattern (Fontaine, Love, & Latinus, 2017). In this way, it is largely unknown whether listeners use different features in different listening contexts, as most studies necessarily use a fixed set of highly constrained stimuli in order to explore the effects of their manipulations. This highlights the complexity of investigating the acoustic cues important for voice identity processing and how the importance of these cues changes as a function of experience.

Overall, these findings showed that more experience with voices via an increased amount of training led to significant improvements in recognition ability in a perceptually challenging task (Experiment 5), but that this positive effect of training is lost with even small changes to the type of stimuli included from training to test (Experiment 4). I also demonstrated that although longer training resulted in better performance, generalisation ability was still poor, suggesting that representations may be relatively rigid and unstable and that many more exposures are necessary for more robust representations to be formed. But, the increase in accuracy in the longer training condition (corresponding to 15 minutes more training), suggested that the acquisition and establishment of voice representations may be a relatively rapid process. This was supported by a reduced reliance on basic acoustic cues such as F0 for listeners who received more training, and lays the groundwork for future research to explore how voice representations may be built and refined over time as a listener becomes more familiar with a voice.

4 Can Voices be Rewarding to Hear? Measuring Effort for Personally Valued Familiar Voices

4.1 Experiment 6

Experimental Chapters 2 and 3 explored questions surrounding the recognition and representation of personally familiar voice identities compared with lab-trained identities, and the effects of the amount of training on recognition ability. Personally familiar voices, however, are not simply familiar stimuli, but can also be indices of people that are important to the listener. Upon hearing a known voice, especially one that holds personal relevance, the process of recognition may be expected to involve additional information about that person, including socio-affective qualities such as episodic memories, emotional responses, and biographical knowledge. In order to understand more about how familiar and valued voices are processed, it is important to consider this aspect of voice processing. This chapter and the following chapter (Chapter 5) will investigate whether familiar voices can function as social rewards, by virtue of who these voices represent. The current chapter details a behavioural experiment that measures motivation to hear voices of differing value using an incentive delay task. Chapter 5 explores the neural underpinnings of these effects.

4.1.1 Introduction

Rewards are desired, appetitive, and positive outcomes of motivated behaviour that have the capacity to increase the frequency of such behaviour (Matyjek, Meliss, Dziobek, & Murayama, 2020). Classically, reward processing has been broken down into three psychological components: motivation, affect/emotion, and learning. Motivation refers to an appetitive phase of reward processing and involves ‘wanting’ or incentive salience. This component has the capacity to affect decision-making and induce approach behaviour via incentive salience. Affect refers to the consummatory phase of reward processing which comprises implicit ‘liking’ and the elicitation of conscious pleasure or hedonic impact. The third component refers to the ability of rewards to produce associative or cognitive learning (Berridge & Robinson,

2003). Specifically, if a stimulus is rewarding, humans (and non-human animals) will learn to repeat those behaviours that resulted in a reward, in an attempt to receive that rewarding experience again (Schultz, 2015).

Rewards have traditionally been classed as either primary or secondary reinforcers, although the assignment of rewards to these categories is not always straightforward (Matyjek et al., 2020). Primary rewards are those that are essential for survival and the maintenance of homeostasis; they have an innate value (e.g. food, sex, shelter). Secondary rewards on the other hand do not have a direct biological/evolutionary link, but rather gain value via associations with primary rewards (e.g. money, power; Sescousse, Caldú, Segura, Dreher, 2013). As human beings are inherently social, rewards are not experienced in isolation. Rather, they often operate within a social context, and are intertwined with the social interactions and relationships one has with others (Fareri & Delgado, 2014). Social rewards are broadly defined as positive experiences involving other people, and can include verbal and non-verbal behaviours and feelings, such as receiving praise, a thumbs up, or a smile from an attractive person (Bhanji & Delgado, 2014).

Studies investigating social rewards have used various types of social interactions or stimuli to examine this type of reward. Within this, the majority of research has been interested in the neural underpinnings of such rewards, with some studies making comparisons between social and other types of rewards e.g. monetary (Rademacher et al., 2010). It has been argued that individuals engage with social stimuli or perform social behaviours because of the value or subjective reward associated with them (Tamir & Hughes, 2018). For instance, if a person learns that performing a prosocial behaviour gains them a friend, or that a smiling person treats them in a kinder way than a person with an angry facial expression, the value associated with the behaviour or stimuli motivates the individual to engage in behaviour that increases the chances of receiving rewarding outcomes. Two partially overlapping approaches common for studying social rewards are 1) research that explores socially rewarding actions or interactions (e.g. receiving a thumbs up, social approval or positive feedback) that an individual could have with, or receive from, others (e.g. Anderson, 2016), and 2) research that looks at the rewarding nature of stimuli that represent socially valued or salient others, such as viewing an attractive face (Cloutier, Heatherton, Whalen, & Kelley, 2008; Aharon, Etcoff, Ariely, Chabris, O'Connor, & Breiter, 2001; Kobayashi, Watanabe, & Nakamura, 2020), a photograph of a loved one (Aron, Fisher, Mashek, Strong, Li, & Brown, 2005; Acevedo, Aron, Fisher, &

Brown, 2012), or hearing the soothing voice of a parent (Seltzer, Prosoki, Ziegler, & Pollak, 2012). The current chapter focuses on the latter. That is, can stimuli that represent a familiar, valued individual, such as the voice or face, be socially rewarding, and in turn motivate behaviour?

Neuroimaging research has revealed that familiar and socially relevant faces are more rewarding to view compared to the faces of strangers (Matyjek et al., 2020). For instance, in an fMRI study by Acevedo and colleagues (2012), participants passively viewed photographs of their romantic partner, as well as control images of a highly familiar acquaintance, close friend, and a low familiar person (a person known significantly fewer years than the highly familiar acquaintance). Activation in the dopaminergic reward system (including the ventral tegmental area and dorsal striatum; see General Introduction) was observed that was specific to viewing the romantic partner's face, as well as some activation in regions implicated in attachment and pair-bonding (Acevedo et al., 2012). For voices, the existing literature is more limited, yet a few studies allude to socially relevant voices functioning as rewarding stimuli. For instance, Seltzer and colleagues (2010) used mother-child pairs to explore whether verbal or tactile contact from the child's mother could have a positive biological effect on the child after engaging in a stressful task. The authors found the largest reductions in cortisol (a biomarker of stress) as well as increases in levels of oxytocin, when the children were comforted via physical and verbal means, as well as when solely comforted through hearing their mother's voice. Further, children who were comforted by their mothers solely via an instant messaging service did not show these stress-relieving responses, suggesting that speech content alone could not explain the observed effects. Conversely, those who received either full contact (including visual, verbal, tactile etc.) or verbal contact alone (comforting their child via telephone) did (Seltzer et al., 2012). Therefore, in both the auditory and visual domains, it appears that there is something about the voice/face of a highly significant other that is sufficient to induce meaningful biological changes in participants, whether this is observed through differences in neural activity or hormonal changes.

The previous research outlined above showing that voices of valued others can be physiologically and biologically impactful, does not necessarily or directly provide evidence that voices can be rewarding stimuli. Instead, these studies illustrate that individual known voice identities can be important, affective signals, emphasising an aspect of the human voice that goes beyond basic recognition. In order to demonstrate whether individual voices can be

rewarding to listen to, a more direct test is needed. Therefore, the current chapter focuses on a defining feature of rewards: that is, their ability to motivate behaviour. According to one conceptualisation (Halahakoon, Kieslich, O'Driscoll, Nair, Lewis, & Roiser, 2018), effort-based decision-making for rewards involves the following cognitive processes: First, the individual generates possible rewarding actions. This is followed by decision-making, which involves weighing up possible actions in terms of the ratio of costs to benefits of performing such actions. Once options have been evaluated, an action is selected. There is then an anticipatory phase where the individual anticipates receiving a reward and this is associated with physiological arousal. Next, the individual engages in motivated action and effort to obtain the reward. Receiving the reward produces a hedonic effect in the individual; this is termed the consummatory phase. Lastly, there is an element of learning where the individual learns from the outcome of their actions to guide future decision-making (Halahakoon et al., 2018).

Thus, many studies interested in examining motivation or 'wanting' for rewards use effort-based tasks such as those that test the speed or frequency of responses, or physical effort, with the notion that larger rewards will promote increased effort to obtain them. For instance, studies by Aharon and colleagues (2001) and Jaensch and colleagues (2014) used a keypress task whereby participants could work to increase or decrease viewing times of photographs of attractive or average faces by virtue of the number of keypresses made. Two other tasks that measure the speed of participants' responses include the Monetary Incentive Delay (MID; Knutson, Westdorp, Kaiser, & Hommer, 2000) task and the Cued Reinforcement Reaction Time (CRRT; Chase, Michael, Bullmore, Sahakian, & Robbins, 2010) task. These tasks focus on the actual action taken to obtain rewards, rather than simply the intention or choice to make an action (Halahakoon et al., 2018). The MID task is widely used and variations have been used to probe other types of reward such as social rewards (termed the Social Incentive Delay (SID) task; Martins et al., 2020). In this task, participants are required to make a button press as fast as possible in response to a target (e.g. white square). However, preceding this target are cues (arbitrary symbols) that are associated with the reward to be gained if responding to the target fast enough. Traditionally, circle cues with differing numbers of horizontal lines indicate different levels of reward, whereas a triangle indicates no reward or a neutral stimulus. If a participant responds too slowly, no reward is received. Thus, the idea is that the larger the anticipated reward (as signalled by the cue) the faster an individual should respond to ensure that that reward will be obtained.

The SID task has been used to examine various types of social rewards, from viewing positive, smiling faces (Barman et al., 2015; Cremers, Veer, Spinhoven, Rombouts, & Roelofs, 2015; Delmonte, Balsters, McGrath, Fitzgerald, Brennan, Fagan, & Gallagher, 2012), to receiving praise (e.g. "Good job!"; Dutra, Cunningham, Kober, & Gruber, 2015; Kollman, Scholz, Linke, Kirsch, & Wessa, 2017), written feedback (fast/slow; Kirsch et al., 2003) or other social approval (e.g. thumbs up, nodding; Gossen et al., 2014). For faces as social rewards, a few studies have used the SID task to investigate the rewarding quality of friendly faces and the neural underpinnings of this (Spreckelmeyer et al., 2009). For instance, a number of studies have used smiling faces as rewarding outcomes in the SID task, with increasing reward value signalled by the smile intensity (slight smile to wide grin). In these studies, reaction times are consistently inversely proportional to the reward magnitude; that is, the larger the reward level, the lower the reaction times (i.e. participants respond quicker), with the slowest reaction times for a baseline/no outcome condition. These behavioural results are accompanied by increased neural activity in regions of the reward circuitry often implicated in reward anticipation and consumption (Spreckelmeyer et al., 2009; Rademacher et al., 2010; Rademacher, Salama, Gründer, & Spreckelmeyer, 2014). In the vocal domain, a few studies exist that use positive verbal feedback as incentives (e.g. Kirsch et al., 2003; Stark et al., 2011; Kollman et al., 2017), though the emphasis in these studies is on the speech content rather than probing the potentially rewarding nature of a voice owned by/ belonging to a valued individual. In a similar way, the research using face stimuli focuses on the rewarding nature of a positive facial expression. Thus, examining the ability for familiar, valued others to serve as rewarding stimuli in the sense that they can motivate behaviour has not, to my knowledge, been studied for voices or faces using incentive delay tasks.

The current chapter thus presents a study using the social incentive delay task to explore whether a valued voice (i.e. the listener's musical idol) could function as a rewarding stimulus and consequently induce increased motivation to hear examples of such voices. Familiar faces and voices are ubiquitous in daily life, from encountering friends, partners, relatives, or celebrities etc. One quality that can separate these familiar individuals is social relevance. Friends and family are more socially relevant than colleagues, and one's manager more relevant than a friendly face at the gym. Thus, we should seek out or approach individuals who are socially relevant to us, and it would therefore be expected that voices or faces belonging to socially relevant individuals should motivate behaviour more so than less socially relevant

individuals. It should be noted that in past research comparing familiar to unfamiliar faces/voices, well-known celebrities have often been selected as stimuli with the intention of being maximally familiar to a large group of subjects. Famous people in these contexts may not be considered to be particularly socially relevant to participants who were selected on the basis that they were merely familiar with the chosen celebrities. However, under different circumstances, a famous voice could be highly socially salient – for example, if the participants are "super-fans" of that celebrity. Based on the previous research that shows increased effort for attractive/positively valenced faces, and the limited voice perception research highlighting positive biological effects upon hearing a valued voice, it is expected that socially relevant voices, due to the people that they represent, should be capable of motivating increased effort to hear them. Thus, in the current experiment, I predicted that participants would display significantly faster reaction times to the voice of their musical idol compared to a less salient voice (an unfamiliar athlete), and a neutral stimulus (pure tone). Moreover, as the human voice is important for communication (via speech) as well as carrying important demographic information (e.g. listeners can judge age, sex, and trait information from a voice; Belin, Fecteau, & Bedard, 2004; McAleer, Todorov, & Belin, 2014), it is expected that hearing a voice, compared to other less salient sounds, may be more behaviourally relevant. Therefore, I also predicted to observe significantly faster reaction times to an unfamiliar voice compared to a neutral stimulus (pure tone).

4.1.2 Methods

4.1.2.1 Participants

119 participants (96 female, 22 male, 1 "prefer not to say"; mean age: 24.74 years; SD = 5.40; range: 18-40 years) were recruited to take part in the current study. Participants that failed to respond quickly enough (i.e. before the target disappeared) in the task on more than 1/3 of trials were excluded, to ensure that participants had sufficient exposure to the three potential reward outcomes. 19 participants were excluded on this basis, leaving 100 participants (80 female, 19 male, 1 "prefer not to say"; mean age: 24.5 years; SD: 10.31 years; range: 18-40 years) included in the subsequent data analysis. Participants were recruited via social media (Twitter, Facebook, Instagram, Reddit), with recruitment advertisements posted on "fan pages." The sample included participants that were "superfans" of one of four popular singers pre-selected by the experimenter. These were: Beyoncé (29 participants), Taylor Swift (37 participants),

Justin Bieber (10 participants), and Harry Styles (24 participants). Four different celebrities were chosen to aid reaching the recruitment target. All participants had normal or corrected-to-normal vision and did not report any hearing difficulties. On completion of the tasks in this experiment, participants had the option to enter into a prize draw for Amazon vouchers. Four vouchers, totalling £100 in value, were randomly allocated to four participants. Note that in the preliminary stimulus ratings collected, each participant there was paid at a rate of £7.50 per hour. Ethical approval was obtained via the UCL research ethics committee (approval code: SHaPS-2019-CM-030) and informed consent given by all participants.

To determine a suitable sample size for the current experiment, an a priori power analysis was run using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). First, results from a previous study by Rademacher and colleagues (2010) using the SID task were used to determine the size of observed effects in that study. Effect size in this study was calculated to be d = 0.36 (df = 31, critical t-value (two-tailed) = 2.04; Cohen's d = 2.04/sqrt(32)). Therefore, the effect size in that study was 0.36 or above. Using G*Power to test the difference between two dependent group means (paired) using a two-tailed test, with the effect size d = 0.36, and an alpha of .05 showed that a total sample of 84 participants was required to achieve a power of 0.90. As this experiment was conducted online, I aimed to recruit a larger sample (approximately 100 participants), to factor in potential noise associated with online testing.

4.1.2.2 Materials

The voice clips used in the current study were taken from interviews with the chosen speakers, sourced via YouTube (https://www.youtube.com/). Interviews conducted in quiet (i.e. no background noise) were selected, and those with high audio quality were prioritised. Videos were saved, voice clips were extracted using Praat, and saved as WAV files. Criteria for the chosen voice excerpts were that they contained 1.5-2s of fluent and meaningful speech that was non-identifying. That is, speech content that would give clues to the speaker's identity or vocation were not included. Extracted voice clips were then RMS normed for amplitude and converted to MP3 format for use on Gorilla.sc. To create stimulus sets that were matched in terms of pleasantness, two rating tasks were conducted with independent groups of listeners.

4.1.2.2.1 Stimulus ratings 1

First, ratings were collected from an independent group of listeners to ensure that none of the celebrity voices were extreme outliers in terms of overall pleasantness, and to match the celebrity voices to an unknown voice of similar pleasantness. For the unknown voices, three athletes were chosen per celebrity (12 athletes total). The chosen athletes were matched to the celebrities by broad regional accent and presumed gender. 124 independent raters were recruited via Prolific.co (www.prolific.co), and each participant was randomly assigned to rate voice clips from 1 of the four celebrities (Justin Bieber: N = 32; Taylor Swift: N = 31; Beyoncé: N = 30; Harry Styles: N = 31) and their three matched athletes. 15 voice clips per speaker were included, so there were 60 stimuli in total. Participants rated each voice clip on "Pleasantness/Attractiveness", "Valence", and "Arousal" on a 9-point scale, where 1 represented very unattractive or unpleasant/negative/low arousal, and 9 represented very attractive or pleasant/positive/high arousal. For Pleasantness/Attractiveness, participants were asked: "How pleasant/attractive, or unpleasant/unattractive does the voice sound to you?" For Valence, participants were asked to rate how positive or negative the voice sounded. For Arousal, participants were asked: "How aroused does this sound to you? Low arousal: the sound is very drowsy and not energetic; High arousal: the sound is wakeful and energetic." Participants rated each voice clip on the three traits by selecting a number with their mouse, and the order of the voice clips was fully randomised. Listeners were not told who the voices belonged to. Six catch trials were also included that required participants to select a specific number from 1-9 as specified by written text (e.g. "Please select the number 2") to ensure sufficient attention was paid to the task. Participants were compensated at a rate of £7.50 per hour.

Mean pleasantness, valence, and arousal ratings were calculated for each of the 12 voices, and one athlete voice was selected to match each of the four celebrity voices, based on these ratings. Ratings were similar and around the middle of the nine-point scale on average for pleasantness (mean = 5.48, range = 4.75 – 6.06) and valence (mean = 4.52, range = 4.17 – 6.06), with a slightly wider variation for arousal (mean = 5.19, range = 3.29 – 6.17; see Table 5 for ratings for all tested voices).

Table 5. Displays average ratings for Pleasantness, Valence, and Arousal for each of the included celebrities and all of the matched athletes tested. The chosen celebrity-athlete pairs are highlighted in **bold**. Standard deviations are reported in parentheses ().

Speaker	Pleasantness		Valence		Arousal	
	M	*SD*	*M*	*SD*	*M*	*SD*
Taylor Swift	**6.06**	**(1.03)**	**5.46**	**(0.76)**	**5.47**	**(1.04)**
Julie Ertz	**5.79**	**(0.71)**	**6.01**	**(0.77)**	**6.17**	**(0.88)**
Allie Long	5.58	(0.92)	5.22	(0.69)	5.14	(0.80)
Emily Sonnett	5.49	(1.02)	5.22	(0.69)	5.15	(0.84)
Beyoncé	**5.74**	**(0.98)**	**5.27**	**(0.79)**	**5.11**	**(0.76)**
Simone Manuel	**5.67**	**(0.97)**	**5.55**	**(0.73)**	**5.13**	**(0.74)**
Dominique Dawes	5.79	(0.94)	5.84	(0.67)	6.06	(0.80)
Laila Ali	5.32	(1.22)	5.55	(0.92)	5.48	(1.08)
Justin Bieber	**5.01**	**(1.34)**	**5.72**	**(0.84)**	**5.80**	**(0.94)**
Sean Monahan	**5.12**	**(1.64)**	**4.17**	**(1.03)**	**3.29**	**(1.18)**
Aaron Ekblad	5.52	(1.11)	5.55	(0.80)	5.03	(0.87)
Tyson Jost	5.53	(1.33)	5.95	(0.67)	6.03	(0.87)
Harry Styles	**5.76**	**(1.14)**	**5.05**	**(0.84)**	**4.35**	**(1.02)**
Max Whitlock	**5.33**	**(1.36)**	**6.06**	**(1.07)**	**5.77**	**(1.39)**
Nile Wilson	5.25	(1.07)	5.50	(0.91)	5.40	(1.33)
Dan Crowley	4.76	(1.16)	4.68	(1.01)	3.80	(1.23)

4.1.2.2.2 Stimulus Ratings 2

A second, short stimulus ratings task was given to listeners to ensure that the 24 chosen stimuli to be used in the social incentive delay task were matched on overall pleasantness. A second group of 120 participants were recruited on Prolific.co to rate 30 stimuli from one celebrity, and 30 from their matched athlete voice (60 stimuli total). Participants rated each clip for pleasantness, valence, and arousal, with the same method as the first ratings task. Means and standard deviations were calculated for the ratings for each of the voice clips, and for the overall ratings of each voice per trait rated. Based on these ratings, 24 voice clips per voice were

selected to be used as stimuli for the social incentive delay task (see Table 6 for descriptives). Excerpts that were rated around the middle of the pleasantness rating scale were chosen - any stimuli rated extremely high or low on any scale were discarded.

Table 6. Displays mean ratings for pleasantness, valence, and arousal for each of the included celebrities and the matched athlete speakers, across the chosen 24 voice stimuli. Standard deviations are reported in parentheses.

Speaker	Pleasantness		Valence		Arousal	
	M	SD	M	SD	M	SD
Beyoncé	5.52	(1.29)	5.48	(0.99)	5.40	(1.06)
Simone Manuel	5.58	(1.33)	5.95	(0.85)	5.84	(0.96)
Taylor Swift	5.61	(1.00)	5.47	(0.83)	5.16	(0.86)
Julie Ertz	5.49	(1.10)	5.93	(0.90)	5.97	(1.04)
Harry Styles	5.35	(1.23)	4.65	(0.83)	4.02	(1.11)
Max Whitlock	4.96	(1.40)	5.87	(0.89)	5.99	(1.09)
Justin Bieber	5.99	(1.34)	5.84	(1.00)	5.63	(1.05)
Sean Monahan	5.75	(1.54)	4.88	(1.01)	3.75	(1.25)

The final pairs of celebrities and their matched athlete voices were as follows: Beyoncé & Simone Manuel, Taylor Swift & Julie Ertz, Harry Styles & Max Whitlock, and Justin Bieber & Sean Monahan. 192 voice clips in total were included in this experiment (24 clips x 4 celebrities + 24 clips x 4 athletes), as well as one pure tone (200Hz) generated using Audacity (https://audacityteam.org/). Voice excerpts were 2.0s in duration on average (range: 1.63 – 2.47 seconds), and the pure tone had a duration of 1.5s. The final set of stimuli were selected on the basis that they were rated "average" in terms of pleasantness, verified via ratings from an independent group of raters (see Stimulus Ratings 2 above).

4.1.2.3 Procedure:

4.1.2.3.1 Social Incentive Delay (SID) task

The current experiment uses the social incentive delay (SID) task (Spreckelmeyer et al., 2009), an adaptation of the monetary incentive delay (MID) task (introduced by Knutson, Westdorp, Kaiser, & Hommer, 2000). This task aims to probe participants' motivation to gain positive social outcomes. In the MID/SID task, participants are shown one of a number of cues that correspond to a particular incentive or outcome, where the incentive's appearance is contingent on the participant responding quickly enough to a target (traditionally adjusted to participants' own reaction times in a practice task). Usually, the potential reward outcomes increase in order of magnitude as signified by the number of horizontal lines in a circle cue. For instance, in the MID task, a circle with three horizontal lines may indicate that the participant would receive +£2 on responding quickly enough to the target, whereas a circle with one line may reflect a gain of 50p. A triangle indicates a neutral outcome e.g. a 'gain' of £0.

The current experiment was conducted on Gorilla.sc (Anwyl-Irvine, Massonié, Flitton, Kirkham, Evershed, 2019), and consisted of a practice phase and the SID task. In the practice phase, participants were instructed to press the "space" key on their keyboard as fast as possible whenever an orange circle (target) appeared on the screen. They were also informed that they would see various symbols appearing on the screen before the orange target appeared, that they did not need to do anything until the orange target appeared, but to keep their eyes focused on the images in the centre of their screen. These were identical to the symbols participants would view in the main task (circle cues with varying numbers of horizontal lines/triangle, sized at 270x266 pixels) and were included to mirror the three cues participants would view in the SID task as well as to get a more accurate measure of reaction times in the context of the SID task. In the practice, a cue was shown on the screen, followed by a fixation cross for a variable duration (between 500-1000ms), followed by the orange target (sized at 270x266 pixels) which stayed on the screen until the space key was pressed. Reaction times to the target were stored on a trial-by-trial basis, and were calculated from the onset of the target on-screen. From this, each participants' individual mean reaction time was saved to be used as the threshold in the SID task; i.e. the duration for which the target would stay on the screen. An image of a loudspeaker icon was presented after the participant made a button press, with a short reminder

that in the "real thing" participants would expect to hear sounds. There were 45 practice trials in total (15 trials x 3 cues).

Next, participants completed the main SID task. As in the practice, participants were presented with a cue for 250ms, followed by a fixation with variable delay (500-1000ms), and then a target circle. Duration of the target was set to each individual participant's mean reaction time in the practice phase. This meant that the task was tailored to each participant such that gaining the rewards remained challenging but not entirely unattainable, thus motivating the participant to continue exerting effort. Participants were required to press the space key as fast as possible before the target disappeared. In the main SID task, there were 72 trials in total (24 trials x 3 cues). The cues either signalled potential reward (n = 48; denoted by circles), or no outcome (n = 24; denoted by a triangle). Immediately prior to performing the main SID task, participants were explicitly informed about which voice/sound each cue represented. The three cues were defined as follows: circle with three horizontal lines = musical idol voice, circle with one line = athlete voice, triangle = pure tone. If participants responded quickly enough to the target, they would hear either a voice clip, or the pure tone, depending on the cue that preceded the target for that trial. If participants responded too slowly on a trial, they would hear the pure tone and see the text "Too Slow!" in red (see Figure 15). Participants that responded too slowly on more than 1/3 of trials (24 trials) were not included in the subsequent analyses.

On completion of the SID task, participants filled out a questionnaire that included a test of the cue-voice pairings, ratings of pleasantness of the 3 voice conditions overall, a question asking whether they recognised the athlete voice, and a multiple-choice quiz containing 10 questions about their chosen celebrity (e.g. "What is Beyoncé's middle name?" see Appendix B). For the cue-voice pairings, each of the symbols viewed in the experiment were shown on screen one at a time and participants were required to select with their mouse the voice/sound that was associated with it. Participants that did not respond correctly on this part of the questionnaire were not included in the analysis. Ratings of pleasantness were collected for each condition as a whole, rather than getting participants to rate each stimulus heard in the experiment. The question asking about recognition of the athlete voice was a "Yes" or "No" question, and participants that selected "Yes" were asked to name this person. For the multiple-choice quiz, participants were explicitly asked not to search for the answers to the quiz on the internet. After completing all ten questions, participants were invited to admit whether they had cheated on

any of the questions, and to state which questions they had cheated on. This quiz was used to certify the fan status of the listeners, but was not used to exclude any participants.

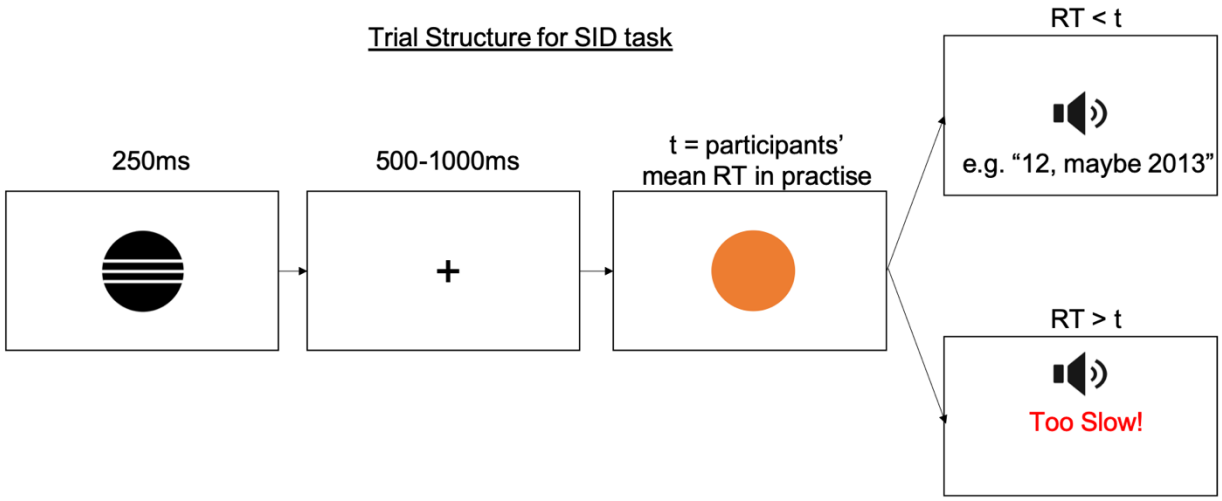

Figure 15. Trial Structure for the social incentive delay task. Participants first viewed one of three symbols that signalled to the participant which voice/sound they would hear upon successfully reacting to the target stimulus (orange circle). If participants responses were faster than the duration of the target on screen (t; individually set based on participants' RTs in a practice task), they would hear either a voice clip, or the pure tone, depending on the cue that preceded the target for that trial. If they responded too slowly (i.e. RT > t), they would hear the pure tone, and see the message "Too Slow!" on the screen. RT = Reaction Time.

4.1.2.4 Data Analysis

Data were analysed using linear mixed effects models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). Model estimates and associated confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero are significant. To assess the impact of the three possible outcome conditions (musical idol voice, athlete voice, pure tone) on participants' reaction times (RTs), an LMM was run with participant RT in ms as the outcome variable, the reward outcomes as a fixed effect with three levels (musical idol, unfamiliar athlete, pure tone), and participant as a random effect. Statistical significance was established via likelihood ratio tests comparing the full model that contained all fixed and random effects to a reduced model where the relevant effect (i.e. reward outcome) was dropped.

4.1.3 Results

4.1.3.1 Quiz Scores, Pleasantness Ratings, and Recognition of the Unfamiliar Athlete Voice:

4.1.3.1.1 Quiz Scores:

The 10-item quiz was scored for each participant as a percentage. Questions in which participants admitted to looking up the answers online were scored as incorrect, regardless of the answer. The majority of participants scored 7/10 or above (85.7% of participants). The quiz questions were not standardised across the four celebrities and thus it could have been possible that the set of questions for each celebrity were not matched for difficulty. Thus, no participants were excluded on this basis. Nevertheless the data were analysed with and without poorly scoring participants (i.e. those scoring <7/10), and this did not change the results.

4.1.3.1.2 Pleasantness Ratings:

Participants' pleasantness ratings were used to examine whether the three obtainable outcomes (musical idol, athlete, pure tone) were rated differently on average. A one-way within-subjects ANOVA was run with the outcome type as the independent variable (3 levels: idol voice, athlete voice, pure tone), and participant rating as the dependent variable. As expected, there was a significant difference in pleasantness ratings of the three outcome types ($F(1.88, 221.33) = 433.11$, $p < .0001$). The participants' musical idol was rated as being the most pleasant (mean = 8.75, SD = 0.54), followed by the athlete voice (mean = 6.85, SD = 1.64), with the pure tone rated as least pleasant (mean = 3.57, SD = 1.99). Pairwise comparisons using the *emmeans* package showed that the differences in pleasantness ratings between conditions were significant for all possible pairs of conditions (musical idol-athlete: t ratio = 10.67, $p < .0001$; musical idol-pure tone: t ratio = 29.09, $p < .0001$, athlete-pure tone: t ratio = 18.42, $p < .0001$; see Figure 16).

4.1.3.1.3 Recognition of the Athlete Voice:

Participants were also asked whether they recognised the unfamiliar athlete voice. If they selected "Yes", they were asked to name this person. One participant recognised Simone Manuel by her profession (i.e. reported that she was a swimmer), but not by name. No other participants correctly identified any of the athlete voices. Thus, the athlete voices were truly unfamiliar for all but one participant.

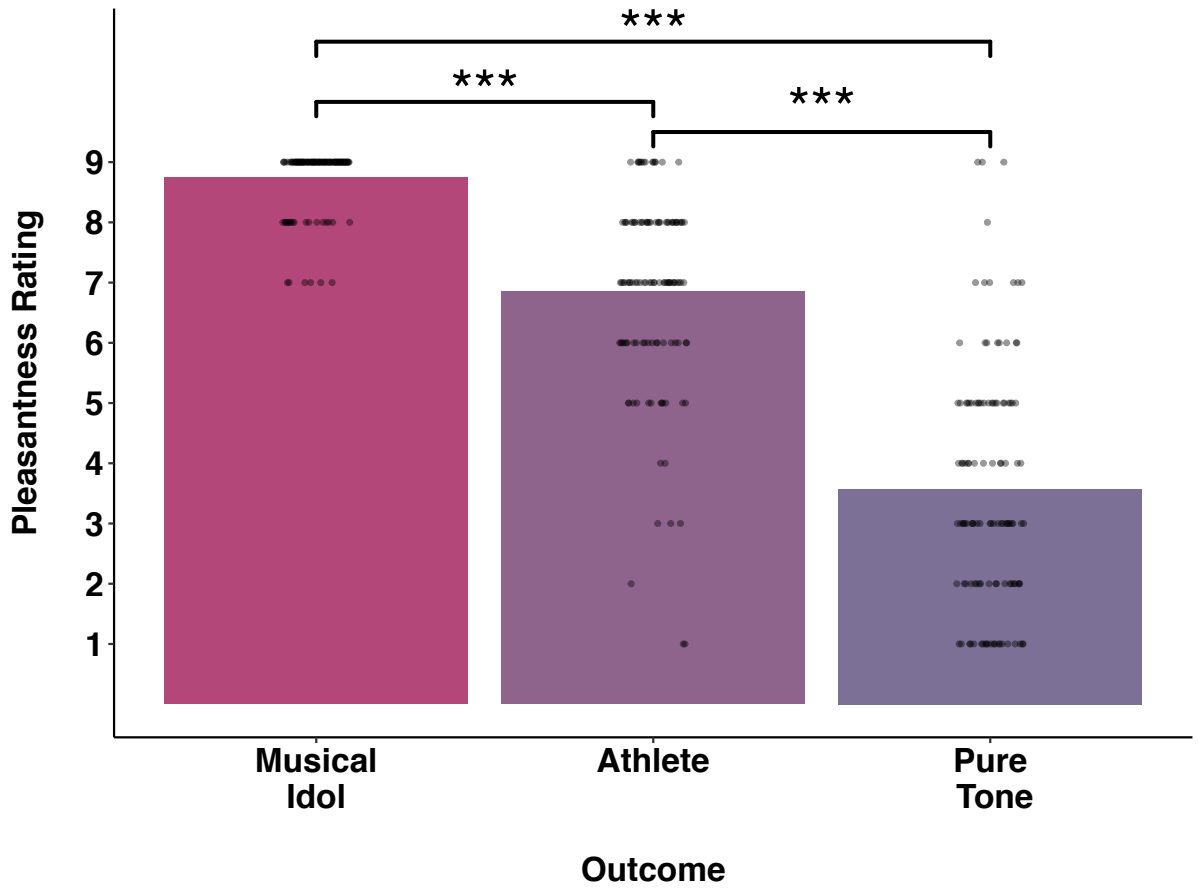

Figure 16. Bars display mean pleasantness ratings for each outcome condition. Ratings ranged from 1 (very unpleasant) to 9 (very pleasant). Individual participants' ratings are displayed as individual points. Asterisks (*) denote significance of pairwise differences in mean pleasantness ratings between conditions. *** $p < .0001$.

4.1.3.2 Social Incentive Delay:

4.1.3.2.1 Confirmatory analysis

Two participants incorrectly matched the cues to the outcomes in the questionnaire task, and were thus removed from the data analysis, leaving a total of 98 participants in the analysis. It

has previously been observed that genuine reaction times have a minimum value of 100ms, and a cut-off between 100-200ms has been suggested to filter out trials that are not likely to reflect true reaction times (Whelan, 2008). Therefore, reaction times that were faster than 150ms were also removed on a trial-by-trial basis. 1,102 trials across 98 participants were excluded on this basis (an average of 11 trials per participant/18% of total trials).

For linear mixed effects models (LMMs), the same assumptions as for regression analysis apply here, except the assumption of independence, as the data in LMMs are grouped in some way and thus this assumption is violated (Davies & Meteyard, 2020). To test that the residuals were normally distributed, a Q-Q plot and histogram were constructed using the residuals. This highlighted that the data were slightly positively skewed (see Figure 17, top panel). Thus, the outcome variable was log-transformed, which was found to improve the normality of the residuals (See Figure 17, bottom panel). The assumptions of homoscedasticity and linearity were tested by constructing a residual plot. This suggested that the linearity assumption was met. Log-transforming the outcome variable improved the homoscedasticity assumption. Therefore, the constructed LMMs included log-transformed reaction times.

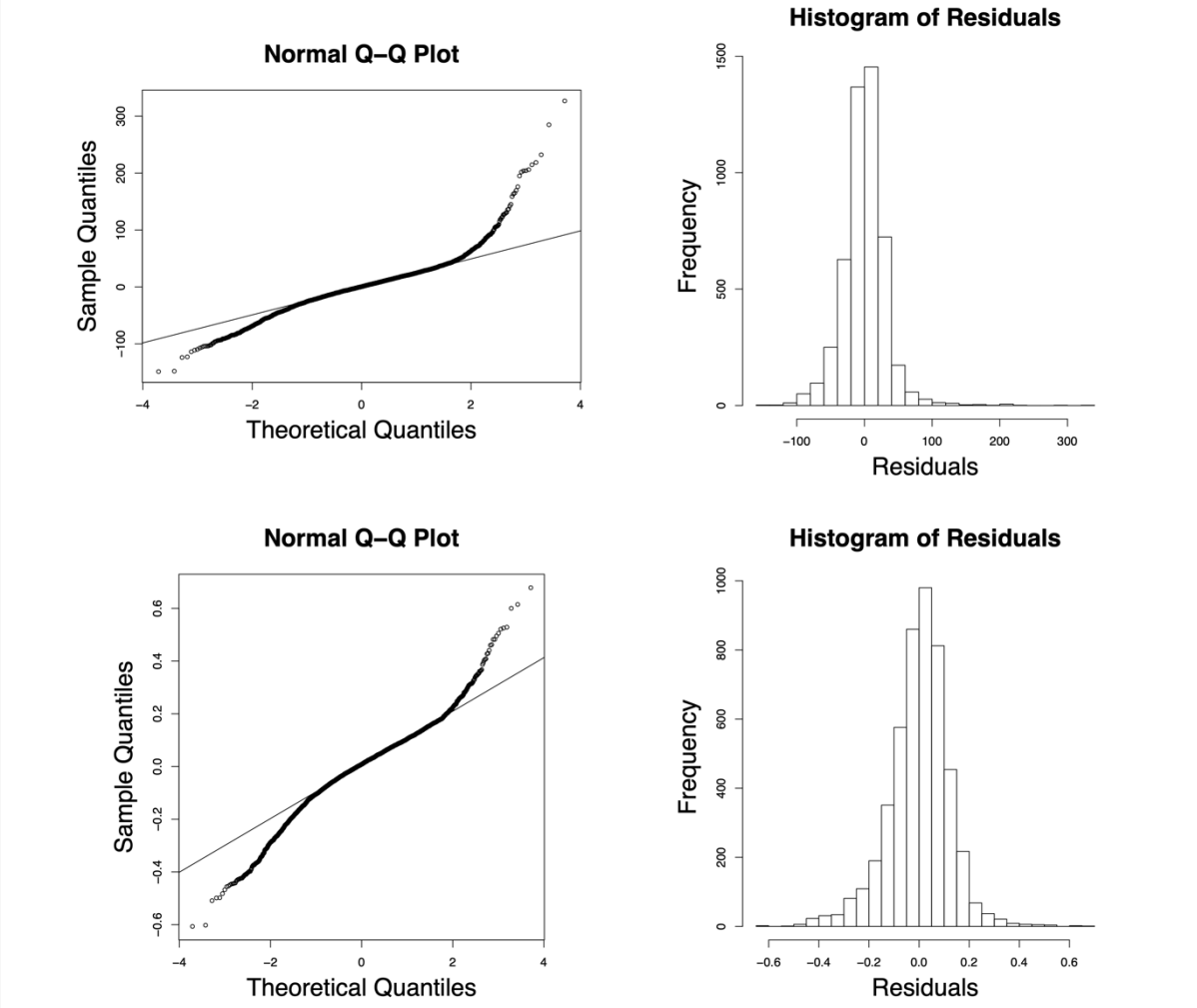

Figure 17. Q-Q Plots and Histograms of residuals for the raw (untransformed) data (top row), and the log-transformed data (bottom row).

Comparing the full model including all fixed and random effects to the reduced model that had all of the same effects except outcome type showed that the type of outcome had a significant effect on participants' reaction times ($\chi^2(2) = 35.21$, $p < .0001$). Post-hoc comparisons (using the *emmeans* package in R) showed that participants were significantly faster at responding to the target when the cue indicated that they would hear their musical idol (raw mean = 241.18 ms, SD = 40.15 ms) compared to both cues that were linked to the unfamiliar athlete voice (raw mean = 246.86 ms, SD = 46.55ms, E = 0.02, CI = 0.01 – 0.03), and the pure tone (raw mean = 246.87 ms, SD = 47.59 ms, E = 0.02, CI = 0.01 – 0.03). However, no significant difference in reaction times to the target was observed between the unfamiliar athlete voice condition (raw

mean = 246.86 ms, SD = 46.55 ms) and the pure tone condition (raw mean = 246.87 ms, SD = 47.59 ms, E = 0, CI = -0.01 – 0.01; see Figure 18).

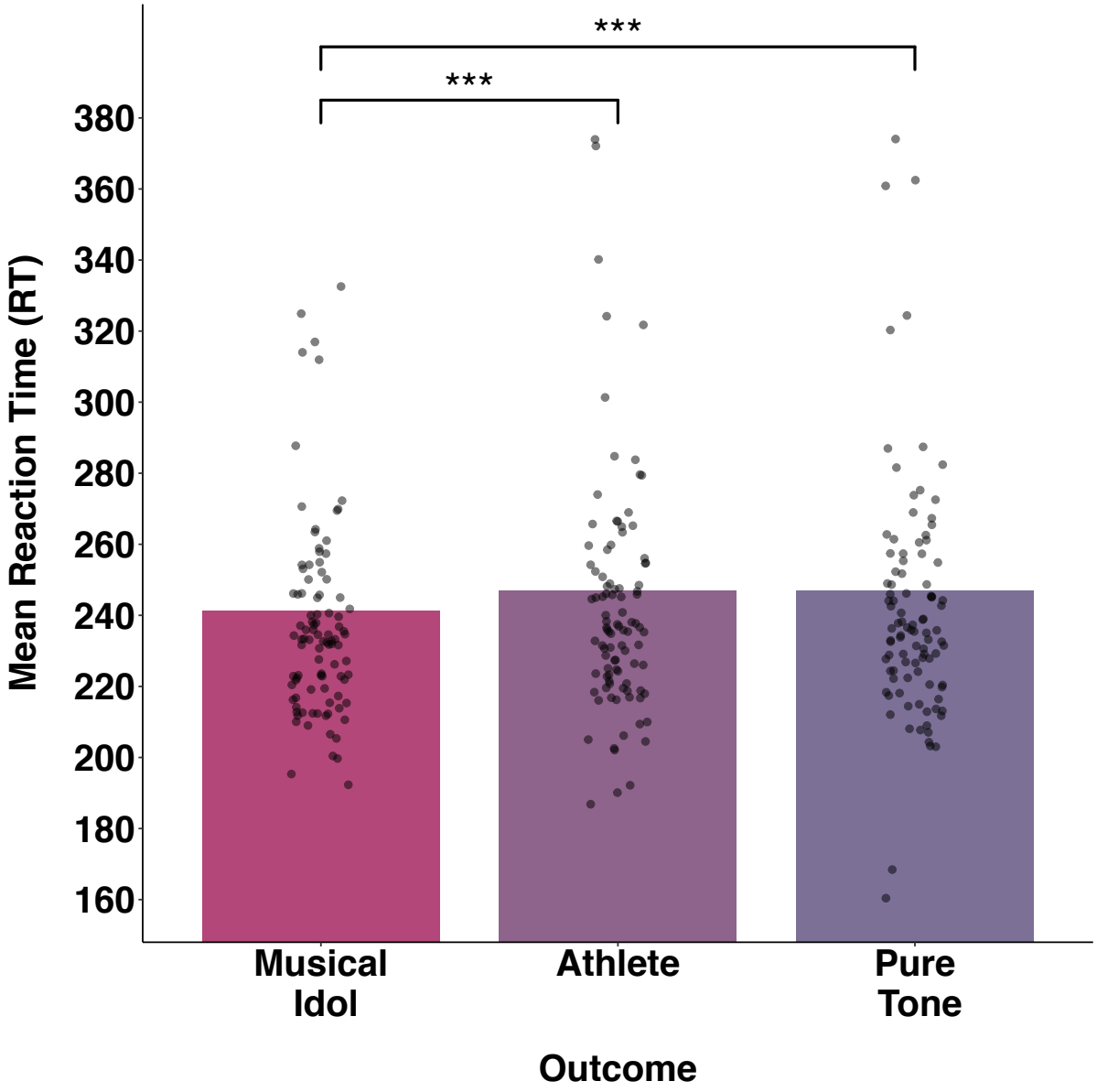

Figure 18. Bars display mean reaction times to the target in each outcome condition. Individual participants' mean reaction times are displayed as individual points. Asterisks (*) denote significance of pairwise comparisons for reaction times between conditions. *** $p < .0001$.

4.1.3.2.2 Exploratory analyses

4.1.3.2.3 Missed Targets

Missed targets were defined as trials for which participants did not respond with a button press to the target quicker than the predetermined threshold (set using the participants' mean RT in

a practice task). To determine whether there was a difference in the number of missed targets depending on the reward outcome that followed a successful response to the target, a binomial generalised linear mixed effects model (GLMM) was run with the binary variable hit/miss on each trial as the outcome measure. Outcome (musical idol voice, athlete, pure tone) was defined as a fixed effect, and participant was entered as a random effect. Statistical significance was established by comparing the full model that contained the fixed and random effect, to a reduced model that only contained the random effect. The comparison of the full to reduced model revealed that there was no statistically significant difference in the number of missed targets ($\chi^2(2) = 5.86$, $p = .053$) between the three outcomes (see Figure 19).

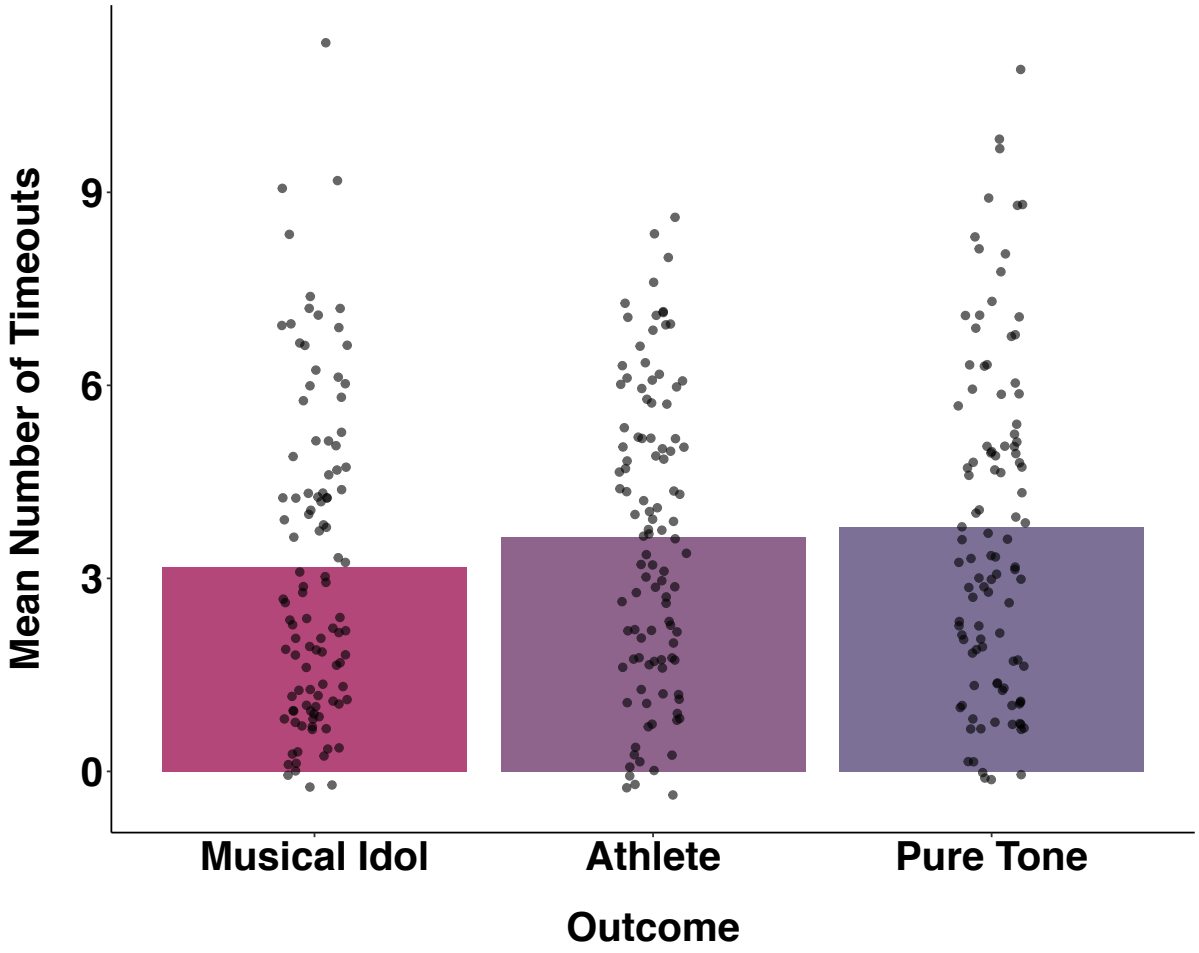

Figure 19. Bars display the mean number of missed targets across all participants for each outcome condition. Individual participants' number of timeouts are displayed as individual points.

4.1.3.2.4 Testing for Familiarity Effects on Reaction Times

To account for the possibility that faster observed reaction times in the SID task to the musical idol were due to these voices being familiar, reaction times (RTs) to each of the reward outcomes and neutral stimulus in the first ten and last ten trials of the experiment were compared. For each outcome condition (musical idol, unfamiliar athlete, pure tone), mean reaction times were calculated for the first ten and last ten trials per participant and these were compared using paired-samples t-tests (Bonferroni corrected for multiple comparisons). These analyses revealed that there were no significant differences in reaction times in the first 10 and last 10 trials for all outcome conditions (musical idol: $t(94) = -0.23$, $p = 0.818$; first 10: M = 237.7ms, last 10: M = 238.1ms; athlete: $t(94) = -0.12$, $p = .903$; first 10: M = 242.9ms, last 10: M = 243.2ms; pure tone $t(94) = -0.09$, $p = 0.929$; first 10: M = 244.0ms, last 10: M = 244.3ms). This suggests that motivation to work to hear the voices was likely not due to familiarity with the musical idol's voice.

4.1.4 Discussion

In the current experiment, the Social Incentive Delay (SID) task was used to explore whether vocal stimuli could function as social rewards, and subsequently motivate effort in listeners in order to receive such vocal rewards. In particular, the motivational salience of a valued voice (the speaking voice of the participants' musical idol) was compared to an unfamiliar speaker and a neutral stimulus (pure tone), with the expectation that the personal relevance, positive valence and affect associated with the listeners' idol would mean that these voices would be socially rewarding stimuli and thus engaged increased effort – through faster reaction times in the SID task – to hear them. Firstly, pleasantness ratings showed that listeners rated their musical idol as being the most pleasant, followed by the athlete voice, and lastly the pure tone as the least pleasant stimulus. In the SID task, it was found that listeners displayed the fastest reaction times to cues signalling the participants' musical idol, and these reaction times were significantly faster than cues signalling an unfamiliar athlete voice and a pure tone. However, participants were equally slow to cues signalling the unfamiliar athlete voice and the pure tone, contrary to my predictions. In addition, the number of missed targets (i.e. "timeouts") in each condition were compared and there were no significant differences in the frequency of timeouts between the three conditions.

Despite the prominence of voices in social interactions and the large amount of socially relevant information contained in a voice, relatively little is known about the social and emotional value of voices and their ability to function as social rewards. Tasks such as the SID task falls into the category of effort-based decision-making tasks, whereby the amount of effort is proportional to the magnitude of the expected reward/outcome (Halahakoon et al., 2020). These tasks are used as a measure of appetitive behaviour, and have been associated with "wanting", or implicit incentive salience (Husain & Roiser, 2018). In the current experiment, listeners were motivated to respond significantly faster to a target when the preceding cue represented their musical idol in this task. Thus, the current study is one of the first to demonstrate that specific vocal identities can be used as motivationally salient stimuli. Listeners' rapid reaction times to their musical idol highlights that these voices are socially relevant and there is a desire to work harder to achieve hearing them, compared to other, less salient stimuli. This is particularly interesting as the stimuli used were short voice excerpts of neutral speech taken from interviews, which alone would not afford the listener much information in terms of the context of the conversation or the subject matter.

This experiment highlights an important distinction in the study of the social value of voices: that not all voices appear to be salient and capable of motivating behaviour, as listeners only worked harder to hear voices of a valued other (i.e. their musical idol). The voice of an unfamiliar person (athlete) did not appear to be motivationally salient and participants did not work harder to hear this voice compared to a pure tone, suggesting that who the voice belongs to is an important aspect of whether an individual will work harder to hear it. It is worth mentioning that these results highlight how the sound of a familiar valued voice can be rewarding over and above speech content or emotional valence, however this is not to say that vocal stimuli can only serve as social reinforcers if they are familiar and valued. Under different circumstances, listeners have been found to work harder for the opportunity to receive verbal praise (e.g. "Good job!") compared to neutral or negative verbal feedback (e.g. "Unfortunately too slow!"; Kollman et al., 2017) and in such contexts, the identity of the speaker has been less relevant. The current experiment however, specifically explores how the voice and who it represents, over and above speech content and vocal emotion, can serve as a socially rewarding stimulus in and of itself. This extends the ideas set out by previous research that explored the ability for loved voices to produce physiological benefits in the listener. For instance, research by Seltzer and colleagues (2012) found reductions in cortisol (a stress hormone) and increases in oxytocin in children upon hearing their mother's voice, after completing a stressful task.

This positive effect was not found when receiving a reassuring text message from their mother. Thus, the valued voice has been shown to be capable of inducing meaningful biological changes in listeners, with the current study displaying the ability for these voices to serve as rewarding incentives, positively affecting listeners' effort to hear them.

The two vocal identities used in the current experiment differed not only in terms of social relevance to the listener, but also in terms of familiarity. The participants' musical idol is familiar, and the athlete voice is unfamiliar. It is important to ascertain whether the observed effects were due to social relevance or familiarity. Socially relevant individuals are those who bear personal importance to the individual, with personal subjective meaning. These individuals are often associated with emotional responses and affective knowledge (Matyjek et al., 2020). Familiar individuals, on the other hand, are those one recognises and has varying degrees of knowledge about. All socially relevant people are familiar, but not all familiar individuals are socially relevant (Matyjek et al., 2020). It is assumed in the current experiment that observed effects were due to social relevance rather than familiarity, as one may be highly familiar with a voice, for instance the voice of a radio show host, but this individual may not be personally relevant or salient to the listener and thus we may not expect a listener to be motivated to work harder to hear these types of voices. It is worth noting here that famous voices make up a unique category, as most of the time, these are individuals that listeners have never met and do not have a conventional relationship with. Rather, these are parasocial relationships. However, a "super fan" listener may possess a great deal of biographical knowledge of their musical idol, have strong emotional responses towards them, and highly value these individuals, which are key features associated with a socially relevant individual. In order to examine whether there were any familiarity effects on participants' responses, reaction times to the target during the first ten trials and last ten trials of each outcome condition were compared. Over the course of the experiment, listeners would become increasingly familiarised with the unfamiliar athlete voice and the pure tone, and thus by comparing reaction times at the start to the end would help to get a sense of whether there were any familiarity effects on participants' reaction times and their motivation to hear these stimuli. It was found that for all conditions, there were no significant differences in reaction times in the first ten to the last ten trials. This suggests that faster reaction times for the listeners' musical idol was more likely due to the social relevance of these voices rather than an effect of familiarity.

Another important finding of the current experiment was that no significant differences in reaction times were observed between the unfamiliar athlete voice and the pure tone conditions, contrary to expectations. It was expected that as the voice of an unfamiliar person is a social stimulus, regardless of its personal relevance, listeners would work harder to hear this voice in comparison to a pure tone that is not a social stimulus. Previous research has highlighted that the mere presence of a social stimulus can have effects on behaviour and physiological responses. For instance, the presence of a pair of eyes can increase generosity and prosocial actions (e.g. Bateson, Nettle, & Roberts, 2006). Voices convey a wealth of demographic information such as information about a person's age, health, sex and mood (Belin et al., 2004), as well as aspects of the individual's personality or traits through the sound of their voice and content of their speech (Mitchell & Ross, 2013). In the current experiment, a different voice excerpt was presented on each trial for the musical idol voice and athlete voice, whereas in the pure tone condition, the same tone was played on each successful button press to the target. Therefore, at the very least, it was expected that the athlete voice condition would be more interesting with the prospect of hearing novel voice excerpts on a trial-by-trial basis.

It is unclear as to why no differences in reaction time between the unfamiliar voice and the pure tone were found, however upon speculation, there may be various reasons for this finding. One possibility may be that as the musical idol was a highly salient voice for participants, this may have led to a narrow focus on this voice, with a disregard for all other conditions. Indeed, a small number of participants not included in the analysis solely made button presses for their musical idol, and did not exert any effort to make a button press for the other conditions despite being instructed to respond to the target on every trial. Moreover, rewarding stimuli have been found to capture attention involuntarily. For instance, Asutay & Västfjäll (2016) trained participants to learn to associate two sounds with a large and a small reward. After reward training, it was found that participants performed poorer on a task when the sound previously associated with the large reward was a distractor stimulus, compared to when the sound associated with the small reward was the distractor. This highlights that stimuli that have gained motivational salience can bias attention involuntarily. Perhaps the prospect of hearing the musical idol's voice biased attention and effort to this condition, disregarding the other conditions.

This highlights a related possibility, that the context of the available outcomes in relation to each other may be important. In previous SID and MID tasks, the possible reward outcomes

usually follow a graded pattern. That is, the magnitude of the rewards appears to increase in predictable increments. For instance, in a study by Spreckelmeyer and colleagues (2009), three types of happy faces were presented with increasing levels of intensity; these were a smiling face with closed mouth, smiling face with open mouth, and a smiling 'exuberant' face. These faces belonged to the same pool of identities, were all unfamiliar, and the only aspect that changed was the magnitude of the smile. Thus, these stimuli were highly controlled and the relative distances between these conditions in terms of the magnitude of the expected reward are comparatively distinct and predictable. In the current experiment, as the aim was to explore the incentive salience of familiar socially relevant voices, the relative distance between the musical idol voice and the other conditions may not be as highly controlled, predictable, or equally spaced in terms of reward magnitude. That is, the distance in reward magnitude between the unfamiliar voice and the pure tone may have been smaller than the distance between the musical idol voice and the other two conditions, and this may have prevented the observation of a graded effect as viewed in previous research. Moreover, differences in reaction times between incentive conditions in SID studies have been found to be larger for faces than those observed in the current experiment, and this may be reflective of smaller reported effects in voices compared to faces observed more generally (Hanley & Damjanovic, 2009; Barsics, 2014).

That said, reaction times between the unfamiliar athlete voice and the pure tone were near identical, which highlights a mismatch between subjective ratings of pleasantness, and quantified measures of reward valuation in the SID for these two conditions. Namely, participants' average pleasantness rating of the athlete voice was 6.85 on a 9-point scale (where 9 indicated highly pleasant), and the pure tone was rated lower on average (mean = 3.57), yet there were no significant differences in reaction times in these two conditions. Thus, subjective ratings of pleasantness did not translate into differences in effort in the SID task. This may be explained in part by the dissociation between two separate psychological components of reward: implicit 'wanting', and liking. 'Wanting' refers to implicit incentive salience, is a motivational rather than affective component of reward, and produces approach or appetitive behaviour. Liking refers to the hedonic impact of receiving a reward, how pleasant the stimulus is. Ordinarily, what is liked is also 'wanted', however dissociations exist. For instance, research by Koranyi, Brückner, Jäckel, Grigutsch, & Rothermund (2020) used two versions of the implicit association test (IAT) that were devised to test 'wanting' and 'liking' for coffee respectively. Participants were either heavy or low-consumers of coffee. It was found that

heavy coffee drinkers showed increased "wanting" for coffee without showing increased "liking" compared to low-consumers of coffee. In a more closely related example, a study by Aharon and colleagues (2001) used a keypress task to measure how hard participants would work to increase/decrease viewing time of attractive/average faces. They also collected attractiveness ratings. Heterosexual male subjects rated attractive male and female faces as more attractive than average faces, but only worked harder to prolong viewing times for the attractive female faces. Thus, in both studies, dissociations were found between implicit "wanting" and implicit or explicit liking. However in the first study, increased "wanting" but not "liking" was observed for coffee, whereas for the latter study, participants displayed increased liking but not "wanting" for attractive same-sex faces. I propose a similar mechanism was at play in the current experiment, whereby the athlete voices were subjectively more pleasant to listen to in comparison to a pure tone (increased liking), but listeners were not motivated to work harder to hear these voices (no difference in "wanting").

In the voice perception literature, studies have used unfamiliar voices to explore how vocal stimuli can affect trusting behaviour, finding that aspects of voice such as F0, accent, and expressiveness (e.g. "smiling voice") all had significant influence on trusting behaviours in an investment game (Torre, Goslin, & White, 2015, 2016, 2020). However, in these studies, speech content was always relevant to the game (e.g. "If you invest, we will both succeed") and so effects of voice may have been confounded with linguistic content. One study that controlled for speech content by using neutral sentences (i.e. "Get ready, it's me") investigated the interaction between social traits in the voice, and trusting behaviour in an investment game (Knight, Lavan, Torre, & McGettigan, 2021). The authors found that participants invested more in generous partners (signalling higher levels of trust) compared to mean partners, but they found no effect of ratings of social traits on trusting behaviours in the investment game. That is, a voice rated as happier and more trustworthy sounding did not receive higher investments than a voice rated more neutrally on these traits. This again highlights the dissociation between pleasantness and motivation. Subjective ratings of positive traits were not sufficient to socially motivate a listener in the absence of familiarity or positive social actions, such as displaying generous behaviour or saying trustworthy things. The results of the current study highlight that voices can motivate behaviour regardless of linguistic content, but only if these voices are personally valued.

Therefore, perhaps for voices to function as social incentives over and above speech content and context, these voices need to belong to socially relevant or valued individuals, as it is who these voices represent and the value associated with them that listeners work to hear. For other types of voices to be able to motivate behaviour, such as unfamiliar voices, perhaps the social context or speech content matters more in this case, rather than the identity of the speaker. For instance, recent fMRI research found that when participants either had a live conversation or heard speech directed to them, activation was observed in reward and mentalising networks, however, when hearing pre-recorded or other-directed speech, activity in these networks was not observed (Rice & Redcay, 2016; Warnell, Sadikova, & Redcay, 2017). Therefore, in the current experiment, perhaps the unfamiliar voice alone was not rewarding in the context of isolated excerpts of speech from interviews. It would be interesting in future research to explore whether the opportunity to either have a live conversation or hear speech directed towards the listener would be more socially motivating in the context of the SID paradigm. Although not the aim of the current experiment, it would help to determine under what conditions/circumstances voices more effectively function as social incentives. Moreover, the use of famous voices in the current study meant that creating highly controlled stimulus sets across conditions was difficult. The speech content contained in the voice clips was different for each voice, even though criteria were in place to ensure content was neutral and non-identifying. Nevertheless, for future research, it would be of interest to use voices of personally known individuals, such as a loved one's voice or highly familiar acquaintance. This would allow for tighter control over the speech content of the voice excerpts, whilst also allowing for variation in emotional tone or social relevance (e.g. comparing a loved one reading sentences vs. addressing the listener directly).

Overall, this experiment demonstrated that valued or socially relevant voices can serve as motivational incentives. By studying effort displayed to hear a personally relevant voice, this allowed for the exploration of the rewarding nature of vocal stimuli using a key facet of rewards; that is, their ability to motivate behaviour. I found that listeners exerted more effort to hear the voice of their musical idol, compared to an unfamiliar voice and neutral stimulus, suggesting that particular voices can be rewarding to hear. This effect was observed across four different celebrity voices. In addition, I showed that not all voices are motivationally salient, as listeners only displayed increased effort for their musical idol, and did not work harder to hear an unfamiliar athlete voice compared to a pure tone. Lastly, a dissociation between subjective ratings of pleasantness and quantifiable measures of reward valuation was observed,

whereby unfamiliar voices were rated as more pleasant to listen to compared to a pure tone, but this did not translate to increased motivation to hear the former. This is akin to other voice perception research that investigated the interaction between social traits such as trustworthiness in voices, and trusting behaviour in an investment game. Voices rated as happier and more trustworthy were not rewarded with higher monetary investments in an economic game (Knight, Lavan, Torre, & McGettigan, 2021). This emphasises that voices may be positively valenced signals, but to motivate behaviour over and above speech content, these voices may need to be socially salient, establishing another aspect of familiar voices, namely their emotional value and the effects of this on motivation and behaviour, that has previously been largely overlooked in voice perception research.

5 The Neural Underpinnings of Hearing Personally Valued Familiar Voices

5.1 Experiment 7

In the previous chapter, the findings indicated that listeners displayed more effort in the social incentive delay (SID) task to hear a socially relevant voice – their musical idol - in comparison to an unknown voice, and a pure tone. This result suggests that the voices of socially relevant persons possess rewarding and motivational qualities. Thus, in the following chapter, I aimed to explore whether voices in the SID task would also show engagement of reward and motivation systems in the brain in a functional neuroimaging study, by using voices that differed in their degree of familiarity and personal relevance to the listener.

5.1.1 Introduction

Voices are highly socially relevant signals central to communication, with the voices of familiar, personally important others possessing a "special" status to the listener (Sidtis & Kreiman, 2012). The salience of a mother's voice to infants, for instance, may be present even before birth, and is crucial for bonding, learning, and survival of the infant (Hepper et al., 1993; Purhonen, Kilpeläinen-Lees, Valkonen-Korhonen, Karhu, & Lehtonen, 2004). Neuroimaging research into the perception of voices has studied the processing of speech, vocal emotion, and identity, however, investigating the neural systems implicated in hearing a personally important, valued voice beyond basic recognition is yet to be explored in voices. In faces, passively viewing a loved one's face has been associated with neural activity largely positioned within the brain's reward network (Acevedo et al., 2012; Aron et al., 2004; Xu et al., 2011; Bartels & Zeki, 2000; see Introduction chapter). For instance, the caudate, ventral tegmental area (VTA), cingulate cortices (ACC, PCC) and OFC implicated in reward processing have been observed in these studies, as well as regions implicated in attachment or pair bonding. In voices, a collection of studies centred on examining voice processing in children with or without Autism compared brain responses to hearing a mother's voice to a female control voice (Abrams et al., 2013, 2016, 2019). In one of these experiments, neurotypical children showed greater activity in primary auditory regions and temporal voice areas (e.g. voice-selective STS) when hearing their mother's voice compared to an unfamiliar female speaker. They also

showed greater activity in reward circuitry (NAcc, OFC), salience network (aIns, cingulate), and the amygdala (often implicated in affective processing) for this same contrast (Abrams et al., 2016). Interestingly, the strength of connectivity between voice-selective regions and reward, affective, and salience networks when listening to their mother's voice was associated with social communication abilities in these children. Therefore, this study suggests that familiar voices may not solely be processed in regions involved in the processing of voices, but may also implicate additional systems associated with socio-affective processing due to their personal importance to the listener. However, the use of mother-child dyads taps into a very specific relationship that is important for the child's development and survival. It is still not yet clear whether other personally relevant voices in general would engage the brain similarly.

Behaviourally, the experiment in the previous chapter was one of the first demonstrations of the motivational salience of voices, over and above speech content. It illustrated that listeners exerted more effort in order to hear a personally relevant musical idol voice. Thus, in this chapter, I aimed to explore the brain's processing of different types of voices according to their personal relevance or value, as well as implementing the social incentive delay task in order to investigate the neural bases for the behavioural effects observed in Chapter 4. To enable predictions to be made about the brain regions we may expect to observe in the current experiment, it is firstly important to outline and synthesise previous literatures relevant to the current chapter. Therefore, I will draw upon the findings of previous social and monetary incentive delay tasks to guide such predictions. Additionally, as the conditions also include differences in the degree of familiarity and personal importance of the speakers, previous investigations into voice/face familiarity will also be discussed. Taking together findings from the literature using incentive delay tasks and those probing familiarity and personal relevance will allow for informed predictions to be made.

5.1.1.1 Incentive delay tasks

Incentive delay tasks have used various types of rewarding stimuli, from monetary to social incentives. Reward and loss processing can generally be split into two temporal phases: an anticipation phase and an outcome or receipt phase. Using incentive delay tasks can enable an examination of the neural activity associated with either or both of these phases when encountering different types of rewards, and studies exist that have attempted to separate the

brain systems implicated in the anticipation and receipt of rewards (e.g. Rademacher, Krach, Kohls, Irmak, Gründer, & Spreckelmeyer, 2009). Numerous social incentive delay studies exist, spanning a wide variety of different socially rewarding stimuli from smiling faces to verbal or written praise (see Chapter 4). It is useful to discuss commonly observed brain regions in the anticipation and receipt of social rewards to identify likely structures to be observed in the current experiment. Moreover, it is also important to understand the processes these structures may be involved in, that centre around obtaining and processing rewards.

In the anticipation of rewards in the social incentive delay task, activation is very frequently reported in the striatum. The striatum is part of the basal ganglia and can be divided into ventral and dorsal portions. The ventral striatum, which includes the nucleus accumbens (NAcc) has been associated with motivation and subjective value, and is proposed to have a role in encoding the motivational salience of rewarding stimuli (Rademacher et al., 2014). Dopamine neurons project from the ventral tegmental area (VTA) in the midbrain to the NAcc. It is thought that these neurons are involved in detecting that a potential reward is nearby, and dopamine neurons in the VTA fire prior to actions that result in receiving rewards (Nestler, Hyman, Holtzman, & Malenka, 2014). Many incentive delay studies, both monetary and social, observe activity in the NAcc, with many reporting that this activity increases as the expected value or salience of the cued rewards increase (Rademacher et al., 2010, Rademacher et al., 2014; Spreckelmeyer et al., 2009; Bjork et al., 2004; Knutson, Adams, Fong, & Hommer, 2001, 2003; Wrase et al., 2007). This activity is accompanied by decreases in behavioural reaction times (Spreckelmeyer et al., 2009). Consequently, researchers have identified the NAcc as an important structure for the appetitive phase of reward processing (Rademacher et al., 2010, Spreckelmeyer et al., 2009).

Activation for the anticipation of social rewards has also been observed in the dorsal part of the striatum, that includes the putamen and caudate nucleus (Rademacher et al., 2010; Rademacher et al., 2014; Cremers et al., 2015). The dorsal striatum in reward processing has been associated with motivational processes that support decision-making and goal-directed action (Balleine, Delgado, & Hikosaka, 2007). Increases in activity have been observed in the dorsal striatum in response to the anticipation of primary and secondary rewards (Marche, Martel, & Apicella, 2017). Thus, while the ventral striatum is implicated in encoding the value of rewards, the dorsal portion is thought to be involved in action selection to achieve optimal outcomes (Oldham, Murawski, Fornito, Youssef, Yücel, & Lorenzetti, 2017). Other subcortical

regions such as the insula, thalamus and amygdala are also frequently reported in reward anticipation in incentive delay studies. Both the striatum and thalamus have connections with the insula (Ghaziri et al., 2018). The insula and thalamus have been implicated in the processing of salient stimuli, with the insula being a central node of the salience network (Menon & Uddin, 2010), as well as being involved in social cognition (Couto et al., 2013). It has been proposed that the striatum, insula, and thalamus may work in concert to process the salience of the cues in the social incentive delay task. Other commonly observed regions involve the left inferior frontal gyrus (IFG) and precentral gyrus in the SID task, and the supplementary motor area (SMA) in both monetary and social incentive delay tasks (Martins et al., 2021; Oldham et al., 2017). Both the SMA and precentral gyri are motor regions. In the anticipation phase of the SID task, the participant must process the salience or value of the rewards (as determined by their associated cues), as well as holding the cue in working memory. Sustained attention and motor preparation to make an action towards the target is also necessary to perform this task (Martins et al., 2021). Thus, the anticipation of rewards in the SID/MID tasks appears to be underpinned by activity in reward, motivational and salience networks that support these processes.

Receiving a reward (i.e. the receipt phase) likely involves basic sensory processing (e.g. visual/auditory), which can lead to encoding the subjective value of the received reward. Moreover, as incentive delay tasks are contingent upon participants' performance, and the reward is not always obtained, the discrepancy between predicted and received outcomes may be encoded, in addition to learning and strengthening associations between the cue, action, and outcome (Martins et al., 2021). Ventromedial prefrontal cortex (vmPFC), orbitofrontal cortex (OFC), anterior and posterior cingulate cortices (ACC, PCC), and amygdala are among the regions frequently observed in the receipt phase (e.g. Rademacher et al., 2010; Delmonte et al., 2012; Martins et al., 2021; Oldham et al., 2018). Some studies also find ventral striatum activity in the receipt of rewards, but this finding is not a consistent one (Dutra, Cunningham, Kober, & Gruber, 2015; Martins et al., 2021). Areas within the prefrontal cortex (e.g. vmPFC, OFC) as well as the PCC have been considered to have a role in reward valuation and magnitude of rewards (Cao et al., 2019), with activity in the vmPFC being found to positively correlate with subjective value (Delmonte et al., 2012; Piva, Velnoskey, Jia, Nair, Levy, & Chang, 2019). The OFC is activated by subjectively pleasant stimuli and has been found to have a role in learning associations between stimuli and rewarding outcomes or, in incentive tasks, cues and rewards (Rolls, Cheng, & Feng, 2020). One theory is that the OFC represents reward value,

learns associations between stimuli that are rewarding/not rewarding and transmits this information to regions such as the cingulate cortex that allow for actions that result in rewarding outcomes (Rolls, Cheng, & Feng, 2021). In the cingulate cortex, the ACC is often observed when the incentives are social. This region has been associated with social cognition, salience, and reward-based processing. The PCC on the other hand has been implicated in outcome monitoring, by tracking the environment and remembering past outcomes (Martins et al., 2021). In the SID task, some authors argue that social rewards engage the same brain regions generally associated with reward prediction and receipt, whereas others argue that in addition to these core regions, activity in regions associated with social cognitive functions such as the temporoparietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), precuneus, and superior temporal gyrus (STG) are also implicated (Martins et al., 2021).

Thus, it is clear that in the social and monetary incentive delay tasks in previous studies, engagement of the dopaminergic reward system is robustly observed across studies. Besides this, a distributed network of brain regions supporting the encoding of salience, motor processes, and outcome monitoring, as well as those underlying affective and social cognitive functions (in the SID task) is also observed. Understanding the brain systems that have been observed in previous studies for these two temporal phases is useful to enable predictions to be made, and allows for a more precise/informed interpretation of the findings.

In addition to reward and task-related activity, the current study also includes manipulations of familiarity and social relevance, which differs from previous social incentive delay studies. As mentioned in Chapter 4, social rewards are commonly studied from one of two perspectives: socially rewarding actions or interactions (e.g. receiving a thumbs up or social approval), or individuals as socially rewarding stimuli (e.g. attractive faces, loved ones' faces). Previous studies using effort-based tasks such as the SID tend to focus on socially rewarding actions or interactions (e.g. Spreckelmeyer et al., 2009), but even those that do focus on individuals as socially rewarding stimuli have only tended to use unfamiliar identities (e.g. attractive vs. neutral unfamiliar faces; Aharon et al., 2004), and thus familiarity is held constant across the different conditions or incentives. Other studies that do contrast personally relevant, familiar individuals with less relevant/familiar ones in a reward processing framework usually do not use a task, and instead participants passively engage with the stimuli (e.g. Acevedo et al., 2012). The current chapter explores the neural correlates of hearing voices that differ in terms of familiarity and social/personal relevance, using a task that measures a key component of

reward: that rewards have the ability to motivate behaviour. This study is therefore a combination of the passive studies that explore differences in the brain due to familiarity/personal relevance, and the incentive delay studies that probe the motivation for rewards and neural underpinnings of reward anticipation and receipt, using vocal stimuli. Consequently, we may expect to observe activation that is associated with differences in familiarity and value of individual voice identities. The inclusion of a familiar voice that is not personally relevant allows us to better understand effects that may be due to familiarity, and those due to personal relevance.

The processing of vocal identity in the brain is strongly associated with activity in the temporal cortices, particularly the anterior portion of the superior temporal gyrus/sulcus (aSTS/STG) in the right hemisphere (Aglieri, Cagna, Velly, Takerkart, & Belin, 2021). However, knowing a talker is also associated with biographical knowledge, episodic memories, and emotional responses, and prior research exists that demonstrates differences in the brain systems involved in the processing of familiar and unfamiliar people (e.g. Bethmann, Scheich, & Brechmann, 2012; Schall, Kiebel, Maess, & von Kriegstein, 2014; Latinus, Crabbe, & Belin, 2011). Due to the difficulty in obtaining voices of personally familiar people, fewer studies exist that explore the processing of personally familiar voices in the brain (McGettigan, 2015). However, within the few existing studies, several key regions are consistently reported. The processing of personally familiar voices and faces has been associated with activity in the PCC and precuneus (von Kriegstein & Giraud, 2004; Shah et al., 2001; Nakamura et al., 2001; Arnott, Heywood, Kentridge, & Goodale, 2008; Tsantani, Kriegeskorte, McGettigan, & Garrido, 2019). For example, a case study of an individual with prosopagnosia (face blindness) found that the intact recognition of familiar voices was associated with activity in the PCC and precuneus (Arnott et al., 2008). A meta-analysis of neuroimaging studies comparing specific familiar (i.e. personally familiar) to unfamiliar faces also finds involvement of the PCC (Horn et al., 2016). This region has previously been found to differentiate between familiar (other) and self-related processing. Thus, the PCC and precuneus have both been frequently observed when comparing personally familiar to unfamiliar voices. Other commonly observed brain regions include entorhinal cortex, frontal and temporal poles, which have been associated with autobiographical memory and social cognitive functions (McGettigan, 2015), and the fusiform gyrus associated with face processing (von Kriegstein & Giraud, 2004).

5.1.1.2 The current chapter

Thus, the study in the current chapter is interested in how voices that differ in familiarity and personal relevance may engage reward circuitry in the brain, as has been demonstrated for personally important and valued faces (e.g. Aharon et al., 2001, Acevedo et al., 2012), as well as how the motivational value of particular voices might involve and interact with voice and person perception networks. Therefore, the design used in this study reflects this and differs slightly from the design used in Chapter 4. In contrast to the conditions used in the previous chapter, a pure tone condition was not included. Instead, the three cued outcomes/incentives were as follows: the listener's musical idol voice, another famous familiar voice (not socially relevant to the listener), and an unfamiliar voice. In this way, each condition contains vocal stimuli that are similarly complex, whilst allowing for an exploration of neural differences based on familiarity (familiar vs. not), and the type of familiarity (socially relevant vs. not), with a voice. Lastly, on half of the trials, regardless of participant performance, the expected outcome (i.e. the speaker's voice) was not received, and the other half of the trials the cued speaker's voice was received. In this way, main effects of the voice could be explored, as well as interactions between voice and outcome (received reward vs. not).

Due to the behavioural findings in Chapter 4, and previous research demonstrating the positive biological effects of hearing loved voices, I anticipate that the voices of personally relevant others should be socially rewarding stimuli. I expect that this will manifest in the brain similarly to previous research examining other types of social rewards in the SID task. Thus, I predict that differences in activity will be observed in key regions associated with motivation and reward processing, particularly the NAcc, as well as the dorsal striatum, medial prefrontal regions, and cingulate cortices. Engagement of other regions reflective of salience, outcome monitoring and motor processing, such as the supplementary motor area, anterior insula, and thalamus may also be expected. In all of these regions, I expect the greatest activation in response to a familiar, socially relevant voice (the participant's musical idol). However, it is uncertain as to whether a familiar (non-relevant) voice would be expected to be more rewarding or motivating to hear than an unfamiliar voice. On one hand, it may be expected that what turns a voice into a rewarding stimulus is its personal relevance or value to the speaker, and thus one may expect to see no differences in the motivating qualities of a familiar (non-relevant) and unfamiliar voice. Support for this prediction comes from the previous chapter that found no evidence for familiarity effects on participant performance. On the other hand, a familiar voice

may be expected to be a more socially relevant signal than an unfamiliar voice (e.g. more efficient detection of social cues from familiar faces; see Visconti di Oleggio Castello, Guntupalli, Yang, & Gobbini, 2014), and thus we may expect to see more of a graded effect in the BOLD signal. Moreover, the processing of voices that differ in their familiarity and personal relevance are likely to differ irrespective of the task used. Therefore, I predict differences in activation in regions associated with knowing the talker, such as the PCC/precuneus (often observed in processing familiar people), social brain regions e.g. TPJ, and those implicated in episodic memory or person knowledge. Specifically, a graded increase in activation is expected as the degree of familiarity increases (musical idol > familiar celebrity > unfamiliar voice).

5.1.2 Methods

5.1.2.1 Participants

Twenty-six participants were recruited to take part in the current experiment (mean age = 22.88 years, SD = 4.97, age range = 18-39 years, 19 female). Twenty participants were super-fans of Taylor Swift, and 6 participants were Beyoncé super-fans. One Taylor Swift super-fan's data was excluded due to a technical issue at the MRI scanner. This left a total of 25 participants (mean age = 22.64 years, SD = 4.91, age range = 18-39 years, 19 female). Ethical approval was obtained via the UCL research ethics committee (Approval code: fMRI/2019/005), and informed consent was given by all participants.

5.1.2.2 Materials

Spontaneous speech excerpts were extracted from interviews uploaded to YouTube (https://www.youtube.com/) from three female speakers (Taylor Swift, Beyoncé, and Allie Long (US soccer player) to be used in the current experiment. Voice excerpts were neutral (i.e. not expressive), and speech content was non-identifying (i.e. did not mention cues to their identity, for example referencing or naming songs, teams, or individuals associated with the talker). There were 126 voice excerpts in total (3 voices x 42 stimuli per voice). All stimuli were saved as mono WAV files using PRAAT (Boersma & Weenink, 2010) and were normed

for RMS amplitude. Item durations ranged from 1.62 – 2.19 seconds (mean = 1.92, SD = 0.15 seconds). Visual stimuli created for the SID task included three black and white symbols to reflect different cues. These were circles with varying numbers of horizontal lines: 1, 2, and 3 lines were created. A white square symbol was also created to represent a target in the SID task.

5.1.2.3 Design & Procedure

In order to check that the design for the in-scanner task would not present issues with multicollinearity, a number of first-level design matrices were simulated including names, onsets, and durations of the 6 conditions in this task (see in-scanner task), comparing various inter-trial intervals (ITIs). Design orthogonality was then checked for each simulated design matrix in SPM12. Orthogonality depicts the degree to which regressors correlate with one another (Mumford, Poline, & Poldrack, 2015). It is important that regressors are orthogonal as if they are highly correlated, this can decrease power and cause instability in the fit of the model. It was found that the selected design showed suitably orthogonal regressors (regressors were not highly correlated with each other).

5.1.2.4 **Practice Session**

Prior to entering the scanning session, participants had the opportunity to practise a version of the task that they would be performing inside the scanner (social incentive delay task; Spreckelmeyer et al., 2009) and to learn cue-voice associations. The practice session was conducted on the online testing platform Gorilla.sc (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018), and was split into two parts: 1) an untimed practice, and 2) the task as it would be performed inside the scanner (timed). Task instructions and stimuli were always presented in white against a black background to mirror the task as it would be viewed inside of the MRI scanner.

In the untimed practice, participants were instructed to press the space key on their keyboard as fast as possible whenever they saw a white square appear on the screen. On each trial, participants viewed one of three symbols (same as the cues outlined below) for 250ms, followed by a fixation cross for 500ms. The target white square was then presented following a delay set to last between 500-1000ms. The white square remained on screen until the space

key was pressed, but participants were encouraged to respond as fast as possible. This practice included a total of 45 trials, and was used to collect participants' mean reaction times. On each trial, reaction time was calculated as the time from the onset of the screen displaying the target, until a key was pressed down (i.e. not on release of the key). These reaction times were stored on every trial using a custom script I created in Gorilla's code editor. From this, each individual participant's mean reaction time across the 45 trials was calculated using the stored trial-by-trial reaction times. This value for each participant was then passed on to the main task, used as the total duration of time per trial that the participant was given to respond to the target white square. It should be noted that at this stage, participants were not informed of the meanings of the three cues. Rather, the inclusion of these cues was to keep the task as consistent as possible in all phases.

In the main (timed) practice, participants were introduced to the social incentive delay task. They were instructed to again press the space key as fast as possible when the white square (target) appeared, but that now they would only have a limited amount of time to respond. Here, the timeout threshold was set to individual mean reaction times collected in the untimed practice phase. Participants were instructed to try to respond before the white square disappeared. The target remained on the screen either until the space key was pressed, or until the time limit had been reached. Associations between three cues and the three vocal identities were also learned here. The cues used were the same as in Chapter 4, whereby circles with differing numbers of horizontal lines were used to represent each possible outcome. These were as follows: three lines = musical idol, two lines = familiar neutral celebrity, one line = unfamiliar. Participants were presented with each cue and its associated vocal identity, and were asked to memorise each of these cue-voice pairings. They were informed that upon responding fast enough to the target, they would hear a short voice excerpt from the person associated with the preceding cue, but that if they responded too slowly, they would not hear anything and could assume that they were too slow. There were 72 trials in this practice phase presented in a fully randomised order. 24 voice clips from each of the three speakers were pre-selected by the experimenter and every participant heard the same voice excerpts. On each trial, a cue was presented for 250ms, followed by a fixation cross for 500ms, and a variable delay (between 500 and 1000ms). Next, the target was presented onscreen for the individually set duration. A single voice clip was played on each trial, provided the participant responded quickly enough to the target.

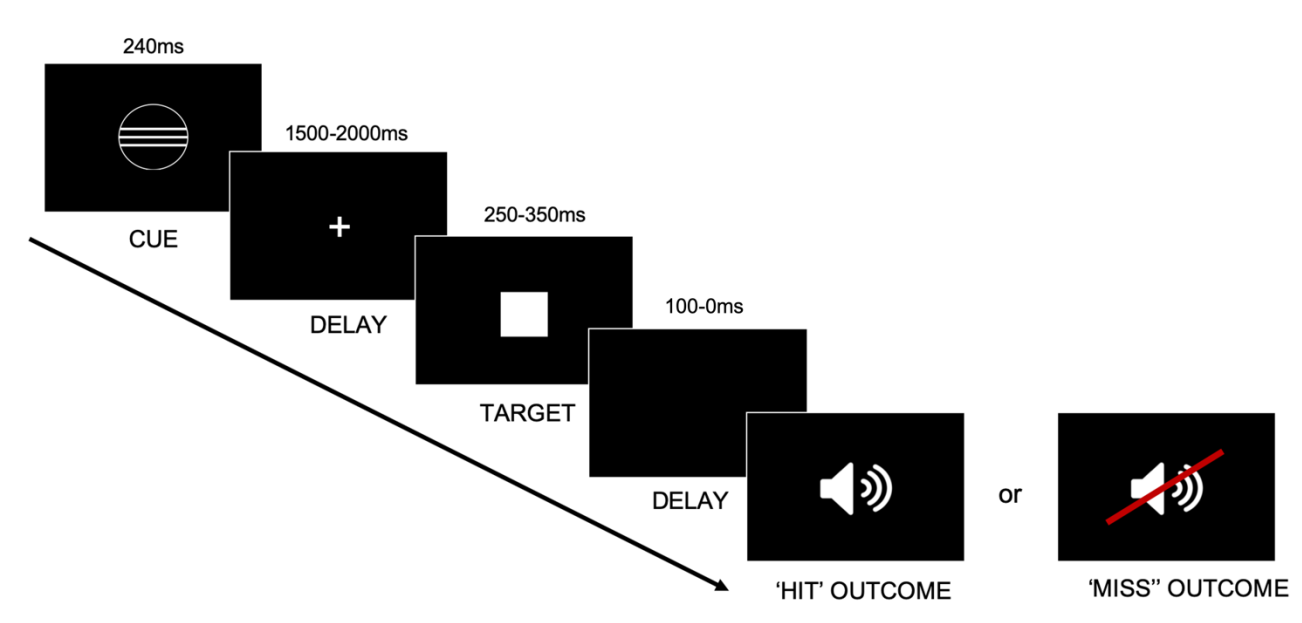

Figure 20. Experimental paradigm for the in-scanner social incentive delay task. Durations are displayed above each phase of the trial. On each trial, a cue provided information about the potential outcome participants could receive upon responding to a target (white square) within the set time window. In this version, the outcome ('HIT' or 'MISS') was not contingent on participants' performance as was their belief, but pre-set to 50%.

5.1.2.5 **fMRI Experiment**

Immediately prior to entering the scanner, participants were tested on their knowledge of the cue-voice pairings. They were shown each of the three symbols and asked to name the associated voice. Inside the scanner, each trial began with the presentation of a cue for 240ms signalling different outcomes (musical idol voice, familiar celebrity voice, unfamiliar voice). After a delay (jittered between 1500-2000ms), a target was presented for between 250-350ms. During the time the target was on screen and before it disappeared, participants had to make a button press with their right index finger in order to receive the cued outcome. There were two possible outcomes: hearing a voice excerpt from one of the three voice identities (HIT trial), or hearing no voice (i.e. silence; MISS trial; see Figure 20). The time between the target appearing and hearing a sound (or no sound) was always set to 350ms. The in-scanner task had a 3x2 factorial design, with the factors Voice Identity (musical idol, familiar neutral celebrity, unfamiliar), and Outcome (HIT, MISS). In order to have a similar number of trials in each cell

of the design, I pre-set the stimulus delivery to play no sounds on 50% of the trials for each voice condition, regardless of the participant's response time. This allowed for an investigation of the anticipation of the voice identities with and without the outcome (i.e. receiving vs. not receiving the voice). There was one exception to this, however. If a participant did not respond to the target, they would not hear a sound, regardless of if that trial was a HIT or MISS trial. This was to ensure that participants believed that their behaviour affected the outcome. Trial and stimulus order were fully randomised, and these randomisations were created offline, such that the order of presentation of the trials was different for each participant. There were 84 total trials per functional run (14 trials x 3 voice identities x 2 outcomes [HIT/MISS]). Voice excerpts were different in each functional run, meaning that participants maximally heard a total of 42 speech tokens from each of the three voices.

During data collection, it became apparent that participants varied in their ability to hit the target, as target duration in the scanner was not individually tailored to each participant. Therefore, the stimulus presentation script was updated part-way through data collection (from Participant 14 onwards) to improve participant performance, and to ensure that participants would get to hear enough exemplars from each of the voice identities. This change involved increasing the amount of time participants had to make a button press, by increasing the window of time in which the script would register a button press by 500ms, extending beyond when the target had disappeared from the screen. However, if a button was pressed whilst the target was on screen, this extra time was not executed. From the participant's perspective, this extra time was not seen, as the target was still displayed on screen for a duration of between 250-350ms.

5.1.2.6 **fMRI Image acquisition**

Scanning was performed on a 3T MR scanner (MAGNETOM Prisma, Siemens Healthineers, Erlangen, Germany), using a 32-channel head coil. Participants laid in a supine position, and were instructed to stay as still as possible during scanning. Head movement was additionally minimised by placing padding around the participant's head. Functional images were acquired over three runs, using an echo-planar imaging (EPI) sequence optimised for imaging the orbitofrontal cortex and amygdala (TR = 2.45 seconds, TA = 2.38 seconds, TE = 30ms, flip-angle = 90 degrees, 35 slices, in-plane resolution = 3mm x 3mm x 2.5 mm, with an inter-slice gap of 0.5mm, field-of-view = 192mm; ascending acquisition). Field of view was adjusted per

participant to encompass the entirety of the frontal and temporal lobes, meaning slice positioning excluded the very top of the parietal lobes. Each run lasted approximately 11 minutes, although the total number of volumes collected per participant and run varied as trial durations were affected by participants' reaction times. Five 'dummy' scans were presented immediately prior to the first trial of each run to allow for steady-state magnetisation to become established. These were discarded and not included in the analyses. Field maps were collected between runs 1 and 2 (short TE: 10ms, long TE: 12.46ms). A whole-brain T1-weighted anatomical image was acquired between runs 2 and 3 (MPRAGE; 160 sagittal slices, voxel size = 1 mm isotropic). During the functional runs, auditory stimuli were presented via MR-compatible earphones (Sensimetrics Corporation, Woburn, MA), with sounds being played via MATLAB (version R2018b, Mathworks Inc., Natick, MA), using the Psychophysics Toolbox extension (http://psychtoolbox.org). Visual information was presented onscreen via a projector connected to a stimulus presentation laptop outside of the scanning room. Participants viewed the screen via a mirror attached to the top of the head coil.

5.1.2.7 Data Analyses

5.1.2.8 Behavioural data

Data obtained from the practice task were analysed using linear mixed effects models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). To assess the impact of the three voice conditions on participants' reaction times, a LMM was run with participant reaction time in milliseconds as the outcome variable, the voice condition as a fixed effect (3 levels: musical idol, familiar neutral, unfamiliar), and participant as a random effect. Statistical significance was established via likelihood ratio tests comparing the full model that contained the fixed and random effects, to a reduced model where the relevant effect (i.e. voice condition) was dropped. Model estimates and confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero are significant.

Trials in which reaction times were faster than 150ms were discarded. Participants that failed to respond quickly enough to the target in the task on more than a third of trials (i.e. 24 trials) were also excluded from the behavioural analysis.

5.1.2.9 MRI pre-processing and analysis

MRI data were pre-processed using SPM12 (Wellcome Centre for Human Neuroimaging, London, UK), implemented in MATLAB (R2018b, Mathworks Inc., Sherborn, MA, USA). Each participant's functional images were realigned with the mean image to correct for slight head movements, and co-registered to align participants' functional images to their anatomical image. Data were then spatially normalised and converted into a shared anatomical space (Montreal Neurological Institute/MNI space). Voxel size changed to $1mm^3$ when these images were re-written. Finally, data were smoothed by convolving the functional images using a Gaussian kernel of 6mm Full Width at Half Maximum (FWHM).

An event-related statistical analysis was performed in a two-level mixed effects procedure. At the single subject level, a fixed effects General Linear Model (GLM) was generated for each participant. Event onsets for six conditions (3 voice conditions x 2 outcomes) were modelled as instantaneous events and were convolved with the canonical haemodynamic response function (HRF). In addition, the six rigid-body movement regressors (resulting from realignment) were included as regressors of no interest. Onsets for trials in which the participant did not make a button press were saved as a seventh condition but were not modelled so that each participant's design matrix had an equal number of regressors.

For the group level random effects model, a 3x2 within-subjects ANOVA was conducted with voice condition (3 levels: idol voice, familiar neutral celebrity, unfamiliar), and outcome (2 levels: hit/miss) as within-subject factors. A partitioned error approach was employed (Henson & Penny, 2005). Under this approach, a set of differential effects is computed at the first level. For our design, this corresponded to a first level contrast of [1 1 1 -1 -1 -1] for the main effect of outcome, followed by a one-sample t-test (with the contrast set to [1]) at the second level, as this factor has two levels. To test for a main effect with three levels (identity), differential effects are calculated at the first level, using contrasts [1 0 0 1 0 0], [0 1 0 0 1 0], and [0 0 1 0 0 1]. At the second level, these three contrasts are entered into a one-way within-subjects ANOVA using the contrast [1 -1 0; 0 1 -1], to test for the main effect of identity. For the interaction, the differences of differential effects are calculated for each subject (Henson & Penny, 2005). We used the contrasts [1 0 0 -1 0 0], [0 1 0 0 -1 0], [0 0 1 0 0 -1] at the first level. At the second level, these contrasts are entered into a one-way within-subjects ANOVA, using

the contrast [1 -1 0; 0 1 -1]. Whole-brain analyses were thresholded at $p < .05$, family-wise error corrected for multiple comparisons. Voxels surviving a less stringent uncorrected voxel-wise threshold of $p < .001$ are also reported. In the case of significant main effects or an interaction, mean parameter estimates from regions of interest (ROIs) for the six conditions vs baseline were extracted for further analyses. Significant whole clusters from the main effects and interaction were saved as ROIs. Next, parameter estimates were extracted from ROIs of interest, for the six conditions vs. baseline, as the mean estimates per cluster. These ROIs were created and parameter estimates extracted using the MarsBaR toolbox (Brett, Anton, Valabregue, & Poline, 2002).

Additionally, a region-of-interest (ROI) analysis was run to test for an effect of identity in the left and right nucleus accumbens (NAcc; at an FWE-corrected level of $p < .05$) using an anatomically defined mask of this region (WFU PickAtlas v3.0; Wake Forest University; Winston-Salem, NC, USA; https://www.nitrc.org/projects/wfu_pickatlas/). This region was chosen as it is a key structure of the brain's reward system that receives dopamine neuron projections from the VTA, has been reported to be involved in coding incentive salience or the value of rewards, and is observed in many studies using SID and MID tasks.

5.1.3 Results

5.1.3.1 Behavioural Data – Social Incentive Delay

Nine participants missed the target on over a third of trials in the behavioural SID practice. I analysed the data with and without these participants and it did not change the results (see Appendix C), therefore the results reported here include all participants. Reaction times faster than 150ms were also removed on a trial-by-trial basis (Whelan, 2008). 227 trials across 25 participants were excluded on this basis. Three trials also had reaction times that were extreme outliers (>1000ms) and were thus removed. Comparing a full model containing the fixed (voice condition) and random effects (participant) to a reduced model that did not contain the fixed effect showed that there was a significant effect of voice condition on participant reaction times ($\chi^2(2) = 15.1$, $p < .001$). Post-hoc pairwise comparisons using the *emmeans* package (FDR-corrected for multiple comparisons) showed that participants were significantly faster to respond to their musical idol (raw mean = 264.5ms) compared to the famous neutral (raw mean = 281.3ms; t ratio = -3.12, p = .003), and to the unfamiliar voice (raw mean =284.1; t ratio = -

3.53, p = .001). No significant differences in reaction time were observed between the famous neutral condition (raw mean = 281.3ms) and unfamiliar voice condition (raw mean = 284.1ms; t ratio = 0.45, p = .650; see Figure 21).

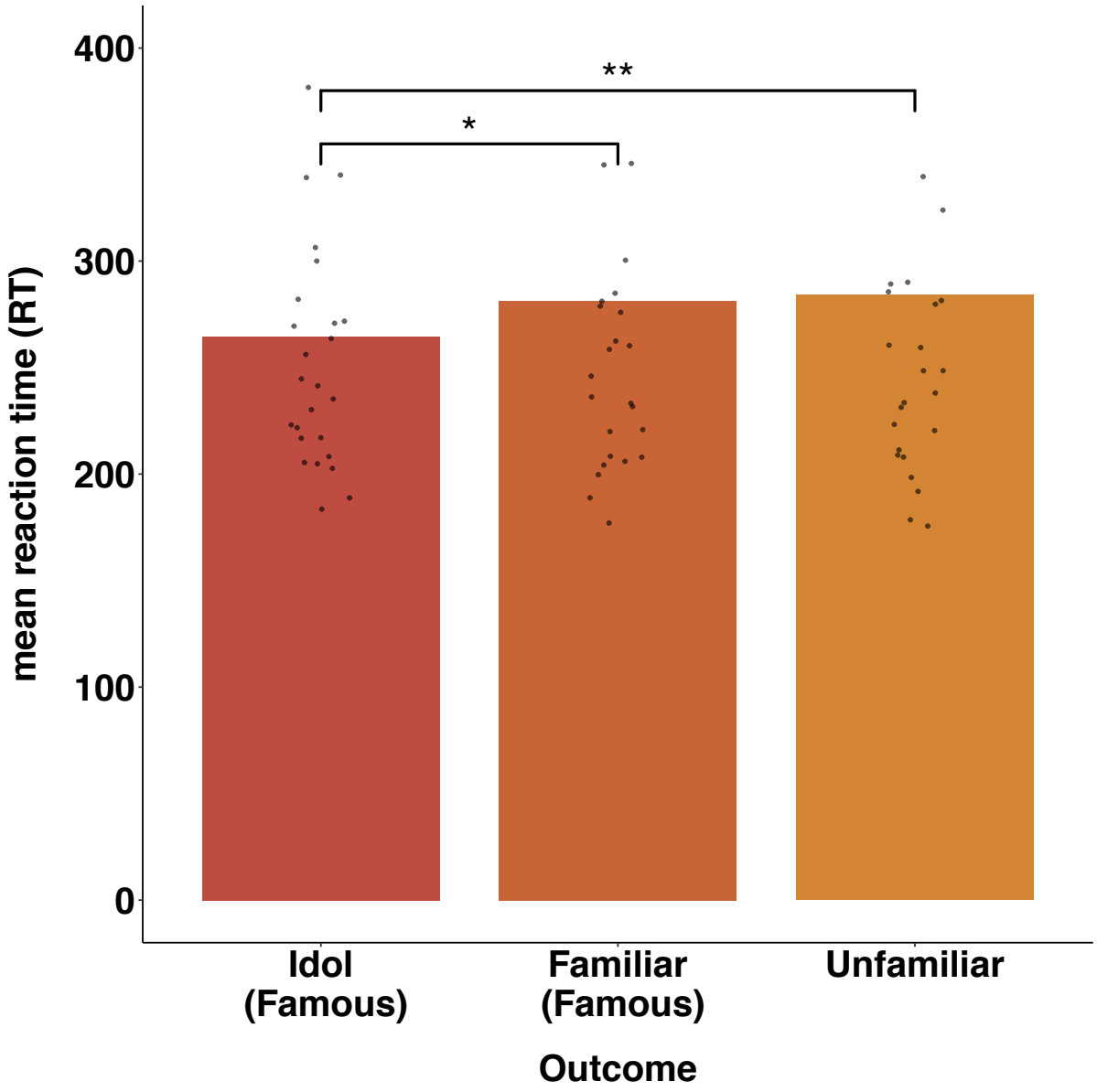

Figure 21. Bars display mean reaction times to the target in each outcome condition. Individual participants' mean reaction times are displayed as individual points. Asterisks denote significance of pairwise comparisons for reaction times between conditions. ** p = .001, * p < .01.

5.1.3.2 Functional MRI: Whole-brain analysis: 2x3 within-subjects ANOVA

5.1.3.2.1 Main effect of outcome

At the group level, a one-sample t-test was run to test for the main effect of outcome (HIT/MISS), and revealed activation that included large clusters in the superior and middle temporal gyri (STG/MTG; peaks in the planum temporale and temporal pole), inferior frontal

gyrus (IFG), and inferior occipital gyrus bilaterally, as well as clusters in the right precuneus, precentral gyrus, and posterior cingulate cortex (PCC; $p < .05$, FWE-corrected; Table 7, Figure 22). Contrasting HIT > MISS ($p < .05$, FWE-corrected) revealed activation in all of the above regions, with the exception of occipital and inferior temporal regions (see Table 7).

Table 7. Results of the main effect of outcome (hit/miss) across all identity conditions. The contrasts hit>miss and miss>hit are reported.

Region of peak activation	Hemisphere	MNI coordinates of peak			z-score	p value (FWE-corrected)	Cluster size (number of voxels)
		x	y	z			
Hit > Miss							
Planum Temporale	L	-45	-34	11	7.50	<.0001	895
Temporal Pole	R	48	14	-16	7.37	<.0001	859
Precuneus	R	3	-58	26	5.85	<.0001	38
Precentral Gyrus	R	54	-4	47	5.84	<.0001	22
IFG (pars opercularis)	L	-45	17	20	5.51	.001	18
PCC	R	3	-34	38	5.43	.001	3
Middle Cingulate Gyrus	R	9	5	32	5.28	.002	1
IFG (pars triangularis)	R	54	32	14	5.20	.004	6
IFG	R	57	23	26	5.19	.004	9
Amygdala/Parahippocampal gyrus	R	30	-1	-22	5.15	.005	2
IFG (pars orbitalis)	L	-39	29	-10	5.04	.009	3
IFG (pars opercularis)	R	57	23	17	4.95	.014	2
Angular Gyrus	R	48	-52	20	4.89	.019	2
IFG (pars triangularis)	R	54	32	5	4.86	.022	1

IFG (pars triangularis)	L	-54	29	17	4.71	.046	1
Medial Frontal Cortex	R	6	38	-19	4.71	.047	1
Miss > Hit							
Inferior Occipital Gyrus	R	45	-73	-4	6.02	<.0001	38
Inferior Occipital Gyrus	L	-39	-82	-7	5.45	.001	46
Inferior Occipital Gyrus	R	33	-88	-7	5.23	.003	17
Occipital Fusiform Gyrus	L	-18	-88	-13	5.01	.010	1
Inferior Temporal Gyrus	R	45	-61	-10	4.99	.011	2
Lingual Gyrus	R	21	-79	-7	4.96	.013	3
Occipital Fusiform Gyrus	R	24	-85	-10	4.83	.025	2
Superior Occipital Gyrus	R	30	-85	11	4.76	.036	2
Occipital Fusiform Gyrus	L	-27	-85	-13	4.76	.037	1
Occipital Fusiform Gyrus	R	30	-79	-16	4.74	.040	1
Occipital Pole	R	24	-94	20	4.73	.042	1

Co-ordinates are reported in Montreal Neurological institute (MNI) stereotactic space. All results are reported at a family-wise error (FWE) corrected threshold of $p < .05$. IFG = inferior frontal gyrus, PCC = posterior cingulate.

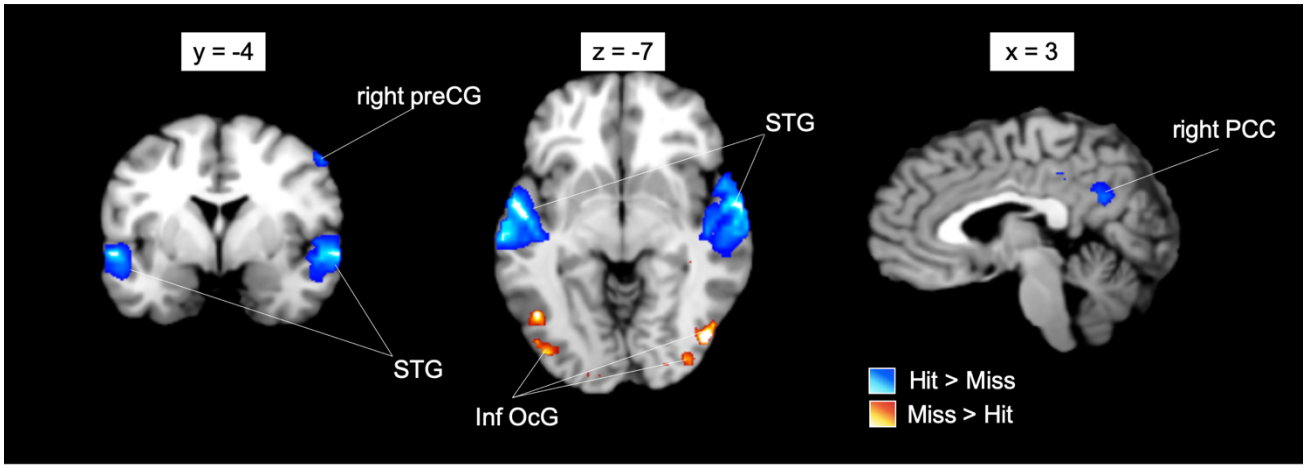

Figure 22. Significant clusters showing increased activation for trials in which voices were heard (hit) compared to silent trials (miss; blue). Clusters showing significant activity for the opposite contrast (miss > hit) are also displayed (red; FWE, $p < .05$). STG = superior temporal

gyrus, preCG = precentral gyrus, PCC = posterior cingulate cortex, Inf OcG = inferior occipital gyrus.

5.1.3.2.2 Main effect of identity

The main effect of identity was tested using a one-way within-subjects ANOVA and revealed a pattern of activation that included anterior insula (AI), IFG, and STG bilaterally (with a peak in the MTG in the right hemisphere), as well as the cerebellum and frontal operculum in the left hemisphere. At a less conservative threshold ($p < .001$, uncorrected, cluster-extent threshold of 10 voxels) further revealed activity in the right caudate, left anterior cingulate cortex (ACC), right cerebellum, and bilateral occipital poles (see Table 8, Figure 23).

Parameter estimates were also extracted from selected significant clusters of interest, and differences between identity conditions (collapsed across outcome conditions) were examined using one-way within-subjects ANOVAs. For significant results, pairwise comparisons (FDR-corrected for multiple comparisons) were run using the *emmeans* package. The results for these selected clusters are reported in Appendix D.

Table 8. Result of the main effect of identity (musical idol, famous, unfamiliar) across both outcome conditions.

Region of peak activation	**Hemisphere**	**MNI coordinates of peak**			**z-score**	**p-value (FWE)**	**p-value (uncorr)**	**Cluster size (number of voxels)**
		x	y	z				
Anterior Insula	R	36	17	-7	5.65	<.0001		39
Frontal Operculum	L	-42	26	-4	5.42	.001		22
IFG (pars opercularis)	R	54	17	-1	5.32	.002		11
Anterior Insula	L	-36	17	2	5.18	.004		2

Frontal Operculum	L	-42	8	2	5.10	.006		1
STG	L	-54	2	-10	5.03	.009		1
Cerebellum	L	-33	-52	-28	4.97	.013		3
Frontal operculum	L	-48	17	-7	4.95	.014		3
IFG (pars orbitalis)	R	42	29	-1	4.92	.016		2
MTG	R	57	-25	-4	4.85	.024		4
Cerebellum	L	-42	-55	-34	4.84	.025		1
Cerebellum*	R	30	-52	-31	4.39		<.0001	25
Caudate*	R	9	11	2	4.30		.003	14
ACC*	L	-6	35	11	4.02		<.0001	40
Inferior Occipital Gyrus*	L	-21	-97	-1	3.96		.002	15
Occipital Pole*	R	21	-100	2	3.69		.004	13

Main effects are reported at a family-wise error corrected threshold of $p < .05$. Co-ordinates are reported in Montreal Neurological institute (MNI) stereotactic space. Regions marked with an asterisk (*) denote peaks that survived a threshold of $p < .001$ (uncorrected) with a cluster-extent threshold of 10 voxels, but did not survive a family-wise error corrected threshold. IFG = inferior frontal gyrus, STG = superior temporal gyrus, MTG = middle temporal gyrus, ACC = anterior cingulate cortex.

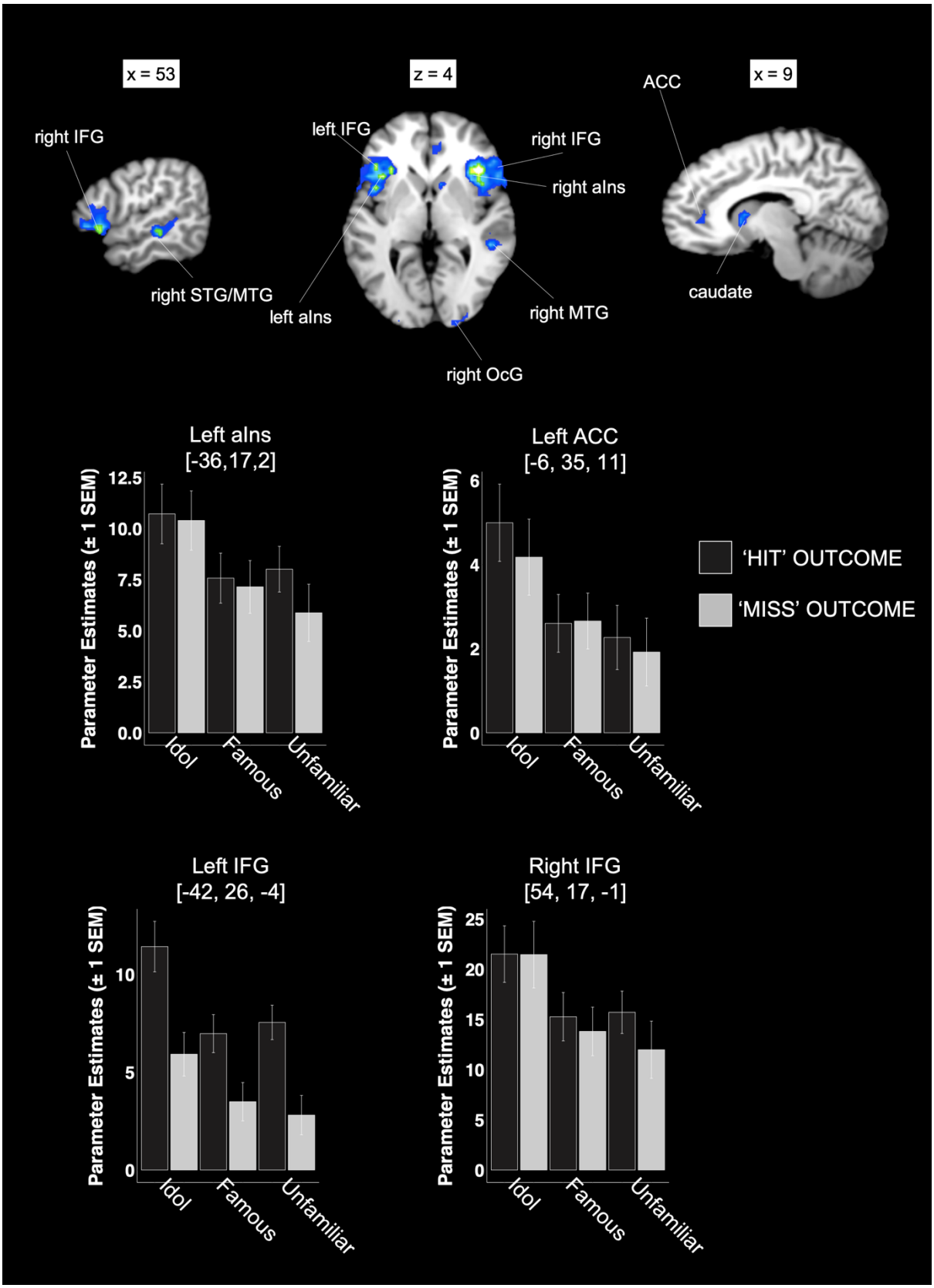

Figure 23. Significant clusters showing a main effect of identity. Activations are shown at both an uncorrected threshold of $p < .001$ (blue), and a FWE-corrected threshold of $p < .05$ (green). Plots show parameter estimates (± 1 S.E.M.) per identity condition, taken from selected

significant clusters (using the MarsBaR toolbox in SPM; Brett, Anton, Valabregue, & Poline, 2002). Note that the main effect is calculated across outcome condition, but for visualisation purposes, parameter estimates are also shown by outcome. Coordinates are given in Montreal Neurological Institute stereotactic space. IFG = inferior frontal gyrus, STG = superior temporal gyrus, MTG = middle temporal gyrus, aIns = anterior insula, OcG = occipital gyrus, ACC = anterior cingulate cortex.

5.1.3.2.3 Interaction between outcome and identity

To test for the interaction between outcome and identity, a one-way within-subjects ANOVA was conducted. This analysis did not yield any significant clusters of activation at the correction level of $p < .05$ (FWE). Using an uncorrected threshold ($p < .001$, cluster-extent threshold of 5 voxels) revealed significant clusters of activation in regions including the right STG, left precuneus, bilateral supramarginal gyrus (SMG) and middle frontal gyrus, and right frontal operculum (Table 9, Figure 24).

Parameter estimates were also extracted from selected significant clusters of interest, and the interaction between identity and outcome condition were examined using 2x3 within-subjects ANOVAs. For significant results, pairwise comparisons (FDR-corrected for multiple comparisons) were run using the *emmeans* package. The results for these selected clusters are reported in Appendix D.

Table 9. Result of the interaction between identity (musical idol, famous, unfamiliar) and outcome condition (HIT, MISS).

Region of peak activation	Hemisphere	MNI coordinates of peak			z-score	p-value (uncorr)	Cluster size
		x	y	z			
STS/STG	R	51	-34	5	4.30	.009	10
SMG	L	-60	-25	29	4.26	<.0001	22
SMG	R	66	-28	32	4.12	.001	17
Precuneus	L	-6	-61	20	4.10	.012	9
dlPFC	R	39	23	41	3.99	.004	13

Frontal Operculum	R	39	20	5	3.66	.017	8
IFG/precentral gyrus	L	-36	5	26	3.51	.024	7

Main effects are reported at an uncorrected threshold of $p < .001$ (cluster-extent threshold of 5 voxels). Co-ordinates are reported in Montreal Neurological institute (MNI) stereotactic space. STS/STG = superior temporal sulcus/gyrus, SMG = supramarginal gyrus, dlPFC = dorsolateral prefrontal cortex.

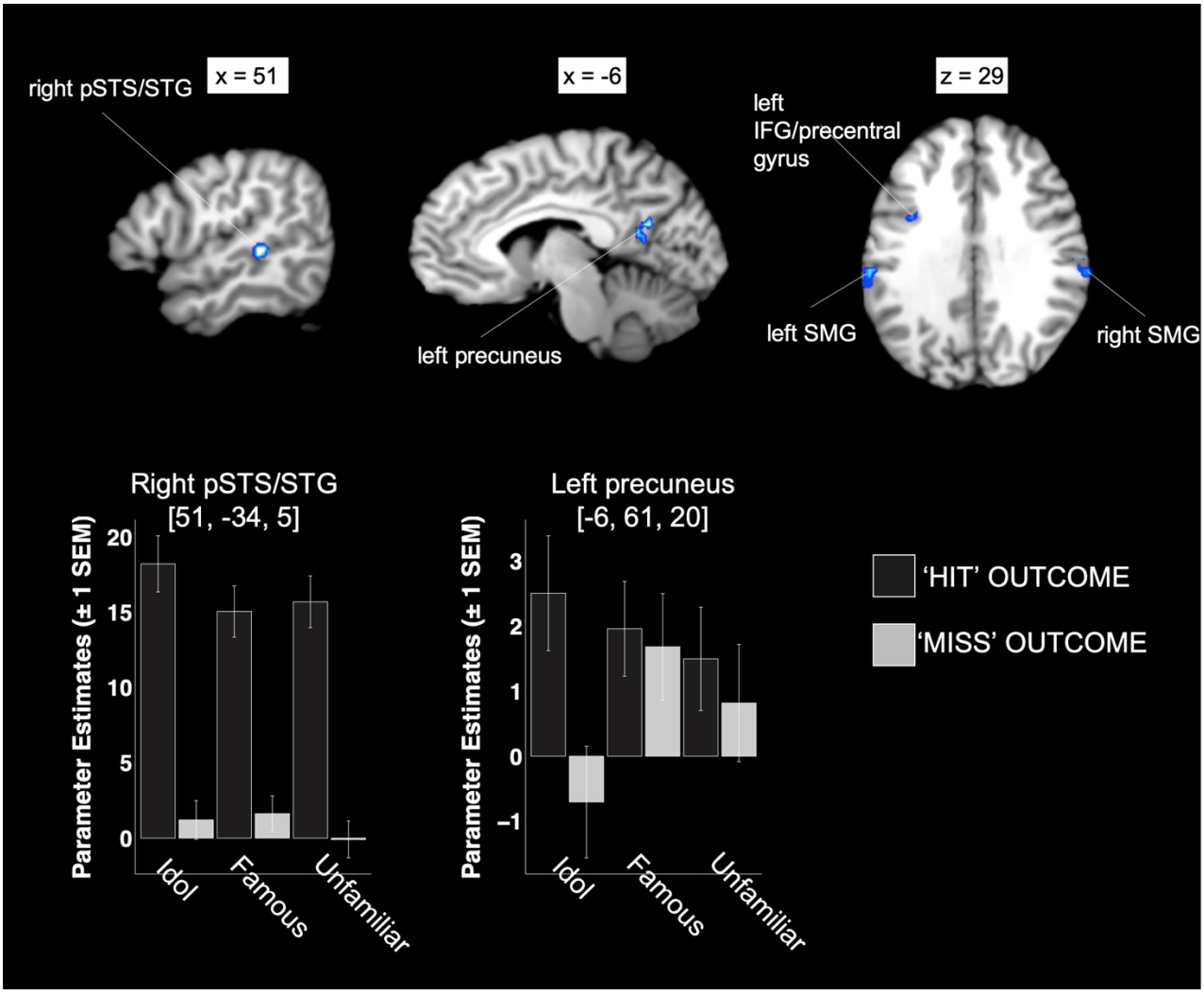

Figure 24. Clusters showing a significant interaction between identity (musical idol, famous, unfamiliar) and outcome (HIT/MISS). Activations are shown at an uncorrected threshold of $p < .001$. Plots show parameter estimates (± 1 S.E.M.) per identity condition and outcome condition, taken from selected significant clusters (using the MarsBaR toolbox in SPM; Brett

et al., 2002). Coordinates are given in Montreal Neurological Institute stereotactic space. STS = superior temporal sulcus, IFG = inferior frontal gyrus, SMG = supramarginal gyrus.

5.1.3.3 **Region of Interest analysis**

Testing for a main effect of identity in the anatomically defined bilateral NAcc revealed that there was no significant effect of voice on neural activity in this ROI.

5.1.4 Discussion

In the present study, I examined whether and how voices of differing familiarity and personal value engaged the brain differently. Speech excerpts from three voices (musical idol, familiar neutral celebrity, unfamiliar voice) were used as rewarding outcomes in the social incentive delay task. The aim of this experiment was to explore whether known, personally relevant/valued voices could be socially rewarding stimuli, specifically questioning if this would manifest as increased neural activity in brain regions associated with reward and motivation processing, as has been observed for other types of social rewards (e.g. attractive faces, positive facial expressions etc.). Due to the nature of comparing voices that differed in their personal relevance/value, this study also necessarily included differences in familiarity. Therefore, activation in regions implicated in familiarity and voice processing more broadly was also anticipated.

Contrary to expectations, the current experiment did not observe a main effect of vocal identity in the nucleus accumbens (NAcc). A region-of-interest analysis in the NAcc bilaterally also did not reveal any significant differences in this region. The ventral striatum, which includes the NAcc, is a key component of the brain's reward system, and has particularly been observed to be involved in the appetitive phase of reward processing, independent of reward type (Rademacher et al., 2010; Kirsch et al., 2003; Dreher et al., 2007; Spreckelmeyer et al., 2009). Studies using incentive delay tasks frequently report activation in the NAcc, and activity in this region often increases with the increasing expected reward value of the stimuli (Rademacher et al., 2010; Spreckelmeyer et al., 2009; Rademacher et al., 2014). Therefore, the fact that activity in this region was not found to be associated with the different value levels of voices in the current experiment is unexpected. Previous SID/MID studies often include four possible

outcomes: three rewarding outcomes that increase in expected reward value or magnitude, and one neutral outcome. This differs from the current experiment which included three outcomes. Additionally, it was not clear from the outset whether the other familiar neutral celebrity (and unfamiliar) voice would be rewarding to a certain degree, solely due to voices being socially relevant signals. That is, we may have expected to observe a graded effect with regard to the reward value of the vocal stimuli in the three conditions, as is often the case in previous studies. However, regions in the current experiment in which differential engagement was observed based on identity, such as the anterior cingulate cortex (ACC), caudate, and anterior insula (aIns), as well as the behavioural findings, suggest against graded effects. That is, only the musical idol voice appeared to be more motivationally salient compared with the other voices in this experiment. This, coupled with generally smaller voice-related effects seen in this thesis (in comparison to effects observed in the face perception literature) may have meant activity in the NAcc was not detected in this study. Nonetheless, in a recent meta-analytic study of neuroimaging findings in the anticipation and receipt of rewards in the SID task, activation in the NAcc was not observed in either temporal phase (Martins et al., 2021). In contrast, a meta-analysis of monetary incentive delay studies did observe the NAcc as a region commonly observed in neuroimaging studies using the MID task, for both reward anticipation and outcome (Oldham et al., 2018). This may raise some questions as to how regularly the NAcc is observed particularly for social rewards, and thus further research may be needed to delineate the precise circumstances under which this region is engaged.

Whilst the current study did not observe activity in the NAcc, the main effect of vocal identity did reveal activation that differed based on the voice condition, regardless of whether the voice was heard or not. It should be reiterated that this main effect captured the effect of vocal identity across the entire trial, including both the anticipation and outcome phases. Typically in studies using the SID task, this main effect is interpreted as the main effect of "reward level." Activation was observed in the aIns, temporal lobe regions (superior temporal sulcus/gyrus, right middle temporal gyrus; MTG), inferior frontal gyrus (IFG) bilaterally, as well as the left cerebellum. Clusters that did not survive family-wise error correction but survived at a less conservative threshold were located in the right caudate, ACC, and bilateral occipital poles. Importantly, activity in these regions was strongest for the musical idol voice condition compared to the familiar neutral celebrity and the unfamiliar voice conditions. There were also no significant differences in the parameter estimates between the familiar neutral celebrity and unfamiliar voice conditions (see Appendix D).

The aIns, together with the dorsal ACC, are key structures of the salience network, which is involved in detecting and selecting the most relevant stimuli in the environment in order to help guide behaviour (Menon & Uddin, 2010). The aIns and ACC parts of the salience network have been proposed to occupy distinct roles within this network (Uddin, 2015). The aIns has been implicated in the detection of salient or behaviourally relevant stimuli in the external environment, as well as being found to be sensitive to internal signals such as autonomic processes (e.g. heart rate, respiration). This region receives inputs from the ventral striatum, VTA, and amygdala, which provides connections between reward, saliency, and emotional processing regions (Uddin, 2015). The current study found that activation in this region was significantly greater for the musical idol condition compared to the remaining two conditions, suggesting that cues to this speaker were more salient. The aIns has been particularly implicated in the anticipation phase of reward processing. Visual inspection of the parameter estimates in this region showed that the aIns was not sensitive to the outcome (i.e. the presence or absence of the voices), suggesting that it may be modulated by the differential anticipation of the three voice identities. Moreover, the right IFG showed a similar pattern of activation to the aIns, suggesting that this region might have a related function. Indeed, the right IFG has been found to be implicated in attentional processes, such as directing attention to salient or task-relevant stimuli (Hampshire, Chamberlain, Monti, Duncan, & Owen, 2010). The right IFG and aIns have been proposed to function as an 'alerting' system allowing for communication between a ventral and dorsal attentional network (Cazzoli, Kaufmann, Paladini, Müri, Nef, & Nyffeler, 2021). During the anticipation phase of the SID task, participants must process the salience or value of the rewards, as determined by their associated cue, and maintain attention throughout the trial. Thus, the aIns and right IFG activity seen here may reflect these processes. In order to inform the amount of effort worth exerting to receive a particular outcome, the salience of each of the cues must be weighted, and attention directed towards the most relevant cues (Martins et al., 2021). The pattern of responses in the extracted parameter estimates mirrors the behavioural findings from the practice task. That is, reaction times to the target in the musical idol condition were significantly faster than both other conditions, and it is also the musical idol condition that generated the strongest activation in the aIns and right IFG. Moreover, there were no significant differences in reaction times to the target between the familiar neutral celebrity voice and unfamiliar voice conditions, which corresponds with the similar levels of activation to these two identities in the aIns and right IFG. Together, the

neuroimaging and behavioural findings indicate preferential processing of the musical idol voice, which represents a more salient and valued stimulus than the other identities.

Also for the main effect of vocal identity, regions were observed that did not survive family-wise error correction, but were observed at a less conservative threshold. These included the ACC, right caudate, and middle temporal gyrus (MTG), as well as bilateral cerebellum and inferior occipital gyri. The dorsal ACC (dACC) is also a region central to the brain's salience network. However, whereas the aIns receives multimodal sensory input and is involved in salience detection, the dACC (and associated dorsomedial prefrontal cortex; dmPFC) transmits motor output, and has connections to the spinal cord, implicating this region in action control (Uddin, 2015). It should be noted, though, that the portion of the ACC observed in the current study was located more rostrally than that commonly identified as belonging to the salience network. Broadly, the ACC has been implicated in a myriad of cognitive functions, including reward and motivation, decision-making, error-detection, social cognition, and motor control (Brockett & Roesch, 2021; Rigney, Koski, & Beer, 2018; Apps, Rushworth, & Chang, 2016; Lavin et al., 2013). This region can be divided into the subgenual ACC (sACC), rostral ACC (rACC), and the dorsal ACC (dACC), although different labelling and methods of dividing this region exist (e.g. the sACC and rACC are sometimes grouped together to form the ventral ACC; Rigney, Koski, & Beer, 2018; Stevens, Hurley, & Taber, 2011; Tang et al., 2019). The sACC is located under the genu of the corpus callosum and has been associated with motivation, emotion, and determining value. The dACC sits adjacent to the PCC and has been associated with salience, as well as motor planning and control (Asemi, Ramaseshan, Burgess, Diwadkar, & Bressler, 2015; Stevens, Hurley & Taber, 2011; Menon & Uddin, 2010). The rACC, the location of the significant ACC activation in the current study, lies between and has connections with the sACC and dACC. Thus, this portion of the ACC sits between motivational and action control networks and may play an important role in the transition from value and choice, to action (Tang et al., 2019).

The caudate nucleus, is part of the dorsal striatum and has been associated with reward processing, goal-directed action, and emotional processing (Graff-Radford, Williams, Hones, & Benarroch, 2017; Driscoll, Bollu, & Tadi, 2021; Grahn, Parkinson, & Owen, 2008). The dorsal striatum has also been associated with selecting or initiating motor responses to achieve optimal outcomes (Oldham et al., 2017). Thus, both the ACC and caudate have been shown to have roles in reward, motivation, and motor actions. In a study by Kurniawan, Guitart-Masip,

Dayan, and Dolan (2013), the amount of effort needed to achieve monetary rewards was manipulated. It was found that during the anticipation phase, the ACC, dorsal striatum, and SMA showed increased activity for higher effort trials. These results again highlight that these regions may be important for integrating the value of rewarding outcomes with the motor actions or amount of effort needed to achieve them (Oldham et al., 2017). Further, a previous study using the SID task found activation in the NAcc, ACC and caudate that specifically displayed increased activity with increasing reward value, suggesting that these regions are implicated in processing aspects of the cue relevant to resultant differences in task performance, such as reward valuation, saliency, or motor planning (Rademacher et al., 2010). In the current study, activity in the ACC and caudate also mirrored the behavioural findings, whereby there were significantly larger BOLD responses in the musical idol condition (across both HIT and MISS trials) compared to the familiar neutral celebrity voice, and unfamiliar voice, as well as no significant difference in activation between the familiar neutral celebrity voice and the unfamiliar voice. Taken together, this implies that perhaps these regions respond to the reward value of the possible outcomes which subsequently inform the amount of effort exerted/motor responses needed to obtain these outcomes. However, as activity in these regions did not survive family-wise error correction, these results should be interpreted with caution.

Nevertheless, the brain regions observed for the main effect of voice identity are suggestive of a network of regions involved in directing attention to salient stimuli, reward processing, and motor preparation, all of which are processes integral to the social incentive delay task. Crucially, these regions do not discriminate between whether the voices are heard or not (i.e. they are not observed in the main effect of outcome, or in the interaction), and show increased activation for the participants' musical idol condition, implying that these voices are identified by their corresponding cues (during reward anticipation) as being more salient and rewarding, and thus more attention is paid to try to obtain the chance to hear them. Interestingly, one region that was observed in the main effect of identity that showed a slightly different profile in the parameter estimates was the left IFG. Unlike the right IFG, this region showed stronger activation on HIT trials compared to MISS trials, for all voice conditions. Additionally, activity was significantly larger in the musical idol condition, similarly to the observed activation in the AI, ACC, caudate, and right IFG (see Appendix D). This region therefore responded differently depending on whether or not a voice excerpt was heard, and thus was influenced by the outcome. Aside from its implication in linguistic processes described above in association with responses to heard speech, this region has been implicated in the social brain network

(Blakemore, 2008; Tso, Rutherford, Fang, Angstadt, & Taylor, 2018). The social brain is a network of brain regions involved in understanding other people's intentions, mental states, and is thought to support social interaction (Tso et al., 2018). Regions thought to be involved in social cognition include the medial prefrontal cortex, IFG, pSTS, ACC, AI, and amygdala (Blakemore, 2008), most of which were engaged by one or both experimental manipulations in the current study. Studies comparing neural differences between neurotypicals and those with ASD have found that activation in regions of the social brain network such as the IFG, insula, STG, and fusiform face area can distinguish these two groups (Patriquin, DeRamus, Libero, Laird, & Kana, 2016). Previous studies using the SID task have also observed activation in the IFG (Kollmann, Scholz, Linke, Kirsch, & Wessa, 2017; Barman et al., 2015; Dutra, Cunningham, Kober, & Gruber, 2015). One study compared neural activity during a monetary and a social incentive delay task (MID and SID, respectively) in participants with bipolar disorder. During the receipt of rewards, activity was observed in the bilateral IFG and orbitofrontal cortex (OFC) but only in the SID task, not the MID task (Dutra et al., 2015). Another study also compared activation in the MID and SID tasks and again supported this finding of bilateral IFG activation during receipt of social but not monetary rewards (Rademacher et al., 2014). In these studies, it was concluded that this region is involved in social information processing and these regions may be associated with processing the social outcomes (Dutra et al., 2015; Grecucci, Giorgetta, Bonini, & Sanfey, 2013). Thus, engagement of the right IFG in the current study may be related to the receipt of rewards, particularly when these rewards are social in nature, and the stronger activation of this region for the musical idol voice indicates that this voice may be more socially relevant.

An interaction between identity (musical idol, familiar neutral celebrity, unfamiliar) and the outcome (HIT/MISS trial) was observed in the right posterior superior temporal sulcus/gyrus (pSTS/STG) and left precuneus, amongst other frontal and parietal regions, although no clusters survived family-wise error correction here, and thus the results should be interpreted tentatively. Observing activation in the pSTS/STG fits with a general role for this region in voice processing, particularly in the right hemisphere. Maguinness and colleagues describe a core voice system in the brain that supports identity recognition (Maguinness, Roswandowitz, & von Kriegstein, 2018). This system includes Heschl's gyrus in the primary auditory cortex and planum temporale, as well as temporal voice areas along the STS/STG, predominantly in the right hemisphere. Based on the current knowledge of vocal identity processing in the brain, it is thought that there is a posterior-anterior gradient in the right superior temporal lobe, where

the posterior portion appears to be more sensitive to the acoustic features of voice identity, whereas the anterior portion has been implicated in matching these features to specific identities (i.e. voice recognition; Luthra, 2021). I found activation in pSTS/STG, perhaps suggesting low-level acoustic processing of the vocal identities. This task was not a voice recognition task, and this could be the reason that more anterior regions were not observed. Thus, this result may reflect differences in the processing of the acoustic properties or the acoustic representations of voices of differing familiarity and relevance. The pSTS/STG has also been proposed to be a modality-general region, responsive to both face- and voice-identity information (Campanella & Belin, 2007). For instance, one study found that the right pSTS was able to discriminate pairs of face identities from response patterns to the corresponding voice identities (Tsantani et al., 2019). In the current study, the musical idol voice was highly relevant and valued, but listeners do not solely value the voice of their musical idol, rather they value the person as a whole. Therefore, activation in pSTS may reflect a multi-modal response to the musical idol compared to the other voices. However, the pattern of responses for the other two voices (familiar neutral celebrity and unfamiliar) in this region are not significantly different from one another. Yet the voice of the unfamiliar speaker could not be represented multimodally as their face was not known to the participant and one would therefore expect less activation in the pSTS for this speaker in comparison to the familiar neutral celebrity voice. This casts some doubt on whether this region was indeed reflective of a multimodal response to the individual identities. Overall, activation in the pSTS may be tentatively interpreted as indicative of differences between voices that vary in terms of familiarity and personal relevance, and these could be low-level acoustic differences, reflect multimodal integration, or perhaps a combination of both.

In the left precuneus, an interaction was also observed. This region, and the adjacent PCC have commonly been associated with the processing of familiarity in voice and face (von Kriegstein & Giraud, 2004; Shah et al., 2001; Nakamura et al., 2001; Arnott et al., 2008; Tsantani et al., 2019). It has been proposed that the precuneus is responsible for determining the familiarity of others, retrieving person knowledge and others' mental states (Lee, Leung, Lee, Raine, & Chan, 2013). A study by Nakamura and colleagues (2001) is one of a few existing neuroimaging studies to use personally familiar voices (friends, colleagues) to explore brain responses to familiar and unfamiliar voices. The authors found activation in the left precuneus, as well as the right entorhinal cortex, left frontal and right temporal poles that showed greater activation for familiar compared to unfamiliar voices. Von Kriegstein and Giraud (2004)

further reported precuneus/PCC activity in response to familiar voices but not to unfamiliar voices. This region has been commonly observed in face perception research too (Cavanna & Trimble, 2006). For instance, a study by Gobbini, Leibenluft, Santiago, and Haxby (2004) found stronger precuneus activity when participants viewed personally familiar faces, compared to viewing famous faces. Thus, this region is found to be robustly activated when determining the familiarity of others, and this appears to be a modality-general region (Shah et al., 2001). The pattern of responses observed in the precuneus are suggestive of an involvement in familiarity processing, whereby the strongest activation was displayed for the musical idol voice, followed by the neutral famous voice, and lastly the unfamiliar voice (for the HIT outcome). However, it must be highlighted that these differences between voice identities for the HIT outcome were not statistically significant, rather the differences represented a trend (see Appendix D). The precuneus was indeed part of an interaction, and the parameter estimates revealed that in the MISS outcome, there was a deactivation in the precuneus for the musical idol condition. This was not observed for any of the other conditions. The reason for this is not entirely clear, however, perhaps this deactivation could reflect a suppression effect: in reward processing, activity in dopamine-rich regions increases if a "better than expected" outcome occurs, whereas these same regions show decreases in activity for a negative prediction error (i.e. no or less than expected reward occurs; Schultz, 2015). In this task, when the listener expected to hear the musical idol voice but did not, perhaps this was reflected in a decrease in precuneus activity. This is a very tentative interpretation, but nonetheless, the pattern of activity in the precuneus highlights that there is a difference in the brain's response to the musical idol compared to the other voices, which varies as a function of whether or not the listener gets to hear them.

An interaction was also observed in other frontal and parietal regions, including bilateral supramarginal gyrus (SMG), right dorsolateral prefrontal cortex (dlPFC), right frontal operculum, and left middle frontal gyrus. Again, none of these regions survived family-wise error correction. In these regions, a disordinal interaction was often observed, whereby activation for MISS outcomes is largest for the musical idol voice and smallest for the unfamiliar voice, whereas for HIT outcomes, the greatest activation is observed in the unfamiliar voice condition, and least in the musical idol condition. The observed regions make up the frontoparietal network, which primarily consists of the dlPFC and posterior parietal cortex, and is involved in problem-solving, attention and working memory (Chenot, Lepron, Boissezon, & Scannella, 2021). This network also has proposed functions in executive

functioning, such as goal-oriented cognition, task switching, and inhibition (Uddin, Yeo, & Spreng, 2019). Thus, differences in activation in the three voice identity conditions in the HIT and MISS outcomes may reflect differences in task-related cognitive functions, such as in attention or goal-directed cognition.

In sum, the study in this chapter used the SID task with vocal stimuli as the rewarding outcomes. It aimed to explore whether particular vocal identities could be socially rewarding stimuli, by using a task that measures a defining feature of rewards - their ability to motivate behaviour – investigating the neural underpinnings of this. Taken together, differences between the voice conditions were observed in regions implicated in vocal identity perception, particularly in areas involved in the acoustic-based processing of voices, as well as in regions commonly engaged in familiarity. These results build upon and support previous neuroimaging findings that explore activation in response to voices that vary in terms of familiarity. In addition to these core regions, I also found evidence for regions associated with the processing of salience, value, as well as social cognition and motor planning. Importantly, activation in these regions was consistently greater for a personally relevant voice (musical idol) whilst showing little differentiation between the other two voice identities. This pattern of activity reliably mirrored the behavioural findings that participants would exert more effort (illustrated by faster reaction times) to hear their musical idol, and showed no significant difference in effort for the other two voices. These results indicate that the voices of specific personally relevant, valued others can influence behaviour by motivating listeners to exert more effort to hear them, implying that these voices are rewarding to listen to. Secondly, this is reflected in the recruitment of these extra brain systems identified in previous studies using the SID task to be involved in detecting salience, encoding the value of rewarding outcomes, and social cognition. Previously, the voice has been largely "typecast" as a vehicle for speech, and studies that have utilised personally familiar voices generally frame investigations in terms of basic recognition. Therefore, the results of the current chapter are important as they demonstrate that particular voice identities can engage additional systems observed for other types of social rewards and illuminate a frequently neglected aspect of particular familiar voices: that they can be personally meaningful, valued signals, capable of motivating behaviour.

6 Discussion

Familiar voices are listened to and engaged with on a daily basis. However, the study of voice recognition has largely neglected those voices at the upper end of the familiarity continuum, with the focus instead being on readily available voices of famous people, or voices that can be trained to be familiar in the lab. In addition to being able to recognise who is speaking, familiar voices also often belong to individuals that we care about, and thus may hold strong personal and emotional relevance to a listener. Yet this is an underexplored element of familiar voice perception. This thesis began to address questions surrounding the recognition of different types of familiar voices, framing familiarity as a continuum whereby the extent and type of experience one has with familiar voices can vary widely and have effects on recognition ability (Experiments 1-5). It also asked questions about the personal relevance of familiar voices and the effects of this on behaviour; i.e. exploring the socially rewarding nature of personally relevant voices and their ability to motivate behaviour (Experiments 6 and 7).

6.1 Summary of the Findings

The experiments in the first half of this thesis compared recognition of voices that differed in their degree of familiarity, using the same experimental tasks. When voices are highly personally familiar, such as the familiarity we have with close friends, relatives, and romantic partners, listeners are able to recognise them to a high degree of accuracy, even under challenging listening conditions. For instance, listeners were 91.3% accurate on average for recognising their romantic partner from conversational filler sounds, that had an average duration of 0.59s and contained minimal linguistic information (Experiment 1). Similarly, Experiment 2 also revealed a high level of accuracy overall for recognising personally familiar voices that had been acoustically modulated. However, familiarity advantages for speech intelligibility were not observed in Experiment 3, but this may have been due to the choice of task rather than being reflective of an absence of familiarity advantages. Personally familiar voices are thought to be underpinned by robust, fine-tuned representations, and the findings from this thesis are consistent with this proposal. Unlike personally familiar voices, lab-trained voices proved to be no match for the highly accurate recognition that accompanies personal familiarity. Recognition of these voices was vulnerable to challenges to perception introduced experimentally, and representations were not stable enough to contend with such changes to

the vocal signal. I also found that manipulating the degree of familiarity within lab-trained voices, by adjusting the amount of training listeners received, did yield significant improvements to recognition, suggesting a relatively rapid updating of stored reference patterns, but there was little evidence to suggest a refinement of stored representations as observed for voices at the highest levels of familiarity. For instance, a simple change in speaking style from training to test in Experiment 4 was enough to disrupt recognition and mask any benefits associated with increasing familiarity training with these lab-trained voices. Therefore, the results from Experiments 1-5 illustrated that voice recognition abilities in the same task can be extremely different depending on the degree of familiarity with the voices to be identified. This demonstrates a need for models of voice processing to consider the ways that familiarity is defined and the effect that this has on vocal identity perception. These findings also inevitably raise further questions about the mechanism and speed with which initial voice representations are established, as well as the processes underpinning the refinement of such representations. That is, how do stored representations transform from those associated with the familiarity observed in the lab to robust stable representations associated with high personal familiarity? Although research has already begun to explore this, future work is needed to continue to investigate the factors important for the formation of robust stored representations accompanying highly personally familiar voices, such as the type and amount of experience necessary.

A second main aim of the current thesis was to explore the social and emotional relevance of certain voices and whether these voices can function as social rewards. Behaviourally, the voice of a personally relevant musical idol was found to be a socially rewarding stimulus, as listeners exerted more effort to obtain the chance to hear this voice in Experiment 6, and this was replicated in the behavioural portion of Experiment 7. Experiments 6 and 7 were the first investigations to establish specific familiar voice identities as socially meaningful signals, capable of inducing approach behaviour and a motivation to engage. This was supported by neural activity in brain regions implicated in reward, motivation, and social cognition, that mirrored the patterns observed behaviourally. Importantly, taking these experiments together, the findings suggest that only certain voices are rewarding, and this appears to be determined by their personal relevance or importance to the listener. No significant behavioural differences in motivational value – as indexed by reaction times – were observed between an unfamiliar voice and a non-vocal stimulus (pure tone) in Experiment 6, and the equivalent response times to familiar (non-idol) and unfamiliar voices in Experiment 7 further showed that basic

familiarity could not explain the effects. That is, listeners worked equally hard to hear a familiar, non-relevant voice and an unfamiliar voice. This finding is particularly compelling as it highlights that another facet of voice processing exists in addition to familiarity, namely the personal relevance or significance of particular identities. It is likely that some voices are innately personally relevant, such that an infant bonds with their caregiver instantaneously. Other voices may become personally relevant with repeated interactions and potentially a high-level attribution of an individual as personally relevant i.e. a learned personal relevance as a relationship develops through social interaction. It is the work of future research to explore the full range of circumstances under which voices are personally meaningful or capable of motivating behaviour, and what the underlying function of this is. Prior work by Abrams and colleagues (2013, 2016, 2019) found that children with Autism Spectrum Disorders (ASD) presented with differences in resting-state functional neural connectivity between voice-selective cortex and regions associated with reward, motivation, and emotional processing (Abrams et al., 2013). The extent of this under-connectivity was linked to the severity of communication deficits in this group, and thus connectivity between vocal and reward pathways was proposed to be potentially important for social development in children. This raises the question as to the possibility that the socially rewarding nature of personally relevant voices may serve some communicative function, such as being important for motivating and maintaining social relationships that are important to us. Sugiura (2014) argued that for social survival, an individual must understand how to respond appropriately towards another person. The argument is that we possess a behavioural readiness to respond, and the type of response depends on the person's relationship to the perceiver. In Sugiura's review of neuroimaging studies into the recognition of personally familiar people, the recognition of friends and colleagues was associated with social cognition, as well as memory retrieval and self-referential processing, whereas the recognition of a loved one was associated with motivation, reward, and affective processing. Therefore, perhaps the brain systems engaged when a voice is anticipated and/or recognised (e.g. reward/motivation) can aid in informing the listener of the appropriate way to interact with the speaker. Therefore, the results from Experiments 6 and 7, together with the earlier experiments finding highly robust recognition of personally familiar voices, reinforces the notion that voice identities can be more than merely familiar stimuli, and thus sets the precedent for future research, in order to continue to study when voices may be rewarding to hear, and the potential functions of this for voice processing and communication.

6.2 Familiarity: Definitions, Representations, and Future Directions

The experiments in this thesis have used a variety of familiar voices, including those of personally familiar, famous, and lab-trained identities. Experiments 1 and 2 in particular directly compared identity perception for highly personally familiar and lab-trained voices within a single group of listeners. This is an uncommon approach within the vocal identity literature. Key differences in identification accuracy for voices of differing degrees of familiarity were found, illustrating that familiarity is not a binary concept (i.e. familiar vs. unfamiliar), but that it is better conceptualised as a continuum. This may seem evident, yet familiarity is often studied as a binary variable, such that the type or degree of familiarity is often unacknowledged as meaningful. Defining familiarity on a spectrum is not an entirely new idea, however (e.g. see Bindemann & Johnston, 2017, for a discussion for faces). In the face perception literature, Clutterbuck and Johnston (2002) used a same-different judgement task to compare discrimination ability for unfamiliar, moderately familiar, and highly familiar faces. They found that the most familiar faces were discriminated more accurately and faster than moderately familiar faces, which were discriminated more accurately and faster than unfamiliar faces. The authors concluded that familiarity does not have an "all or nothing" effect on performance, but rather that it is graded. For voices, Yarmey, Yarmey, Yarmey, and Parliament (2001) used voices of four levels of familiarity: high, moderate, and low familiar, as well as unfamiliar speakers. These authors found that the more familiar the voice, the better and sooner they were recognised from voiced or whispered speech. Moreover, the effect of whispered speech on recognition was least detrimental for voices that listeners were most familiar with. These were people such as immediate family members or a best friend. These findings are similar to those observed in this thesis, whereby voices at the highest levels of familiarity (romantic partners) were better recognised in the face of challenges to perception, in comparison to other, less familiar voices. The way that familiarity is defined affects the ways that voice perception is studied, and thus assuming that we are uniformly familiar with all voices that we know is an oversimplification. Therefore, the results in this thesis illustrate the importance of studying voices as a continuous variable in order to understand how voices are learned, recognised, and represented.

The findings in the first half of this thesis also contribute to the discussion around how voices are represented. It is well-reported that familiarity provides benefits to person recognition, and

this has been explained in terms of differences in the underlying representations between familiar and unfamiliar (or less familiar) voices (Lavan, Burton et al., 2019). Greater familiarity is reflected in more successful generalisation ability across different types of stimuli and speaking styles (Lavan, Burton, et al., 2019). In contrast, the processing of identity from unfamiliar voices (or unfamiliar vocalisations produced by familiar voices) is prone to error. The research detailed in this thesis supports this distinction, finding that personally familiar listeners recognised their romantic partner's voice with a high degree of accuracy and were able to compensate when encountering changes to the way the voices were presented (e.g. when cues were removed (Experiment 1) or became less reliable (Experiment 2)). On the other hand, recognition of lab-trained voices across experiments 1-5 were disrupted to a greater extent due to differences in the type and extent of experiences the listeners had with these voices. The results in this thesis link with previous findings into voice identity recognition that show vulnerabilities where familiarity with the speakers and/or the stimuli used disrupts accurate recognition or discrimination ability (e.g. Lavan, Scott, & McGettigan, 2016), as well as those observing that familiarity allows for a greater ability to generalise across variable stimuli (voice sorting studies; e.g. Lavan, Burston, & Garrido, 2019a). Differences in generalisation ability may be because high familiarity is thought to be a result of lengthy and varied exposure to a speaker. As a consequence, this prolonged and varied exposure is proposed to be useful in the formation of detailed and robust representations that contain critical information about the ways in which a voice can sound under different circumstances (Kramer, Young, & Burton, 2018; Lavan, Burton, et al., 2019).

Voice training studies can also provide insights into how voice representations are formed through learning. The training studies in this thesis (Experiments 4 and 5) explored the effect of the amount of training participants had on the recognition of acoustically modulated voice excerpts. These studies showed that increasing the duration of training led to overall improvements in recognition ability in Experiment 5. However, in Experiment 4, these longer exposure benefits were offset by a change in speaking style, such that performance for unmodulated voice excerpts dropped to 61% correct (from an average of 76.6% correct during training) in both training groups. Thus, just as variability in the experience one has with naturally acquired voices affects recognition and the underlying representations, the stimuli used in voice training studies can also affect the outcomes observed. Lavan, Knight, and colleagues (2019b) conducted a series of experiments exploring the effects of the type of training – high vs. low variability training sets – on participants' ability to make old/new voice

identity judgements on different types of test stimuli. Low variability training items contained vocalisations produced under one speaking style in a single recording session. High variability training items were taken from a number of recording sessions, speaking styles and environments. In one experiment, the training and test sets overlapped for the low variability condition, and less so for the high variability condition. In this study, the authors found a low variability advantage. Similarly, when Holmes, To, and Johnsrude (2021) explored how the amount of training with voices affected subsequent recognition abilities using a consistent speaking style across training and test, it was observed that the voices were recognised well overall and there were no significant differences in recognition accuracy after 10, 20, or 60 minutes of training. Therefore, when there is overlap in this way, there seems to be a general base level ability to recognise voices that appears to develop quite rapidly. However, in a second experiment, Lavan Knight and colleagues (2019b) again trained listeners on low or high variability training sets, but this time, the test items did not overlap in speaking style with the training sets. In this experiment, a *high*-variability advantage was observed. The authors argued that this may be due to being exposed to a range of each speaker's potential variability, making it easier to generalise to novel vocalisations. The voice learning experiments in Chapter 3 of the current thesis showed that increasing the amount of training exposure to voices can have some benefits for recognition, but that representations are still relatively unstable. That is, a change in speaking style from training to test was enough to conceal any potential effects of the amount of exposure. The differences observed in this thesis between recognition of lab-trained voices (Chapters 2 and 3) and personally familiar voices (Chapter 2) illustrate that the familiarity we have with different voices is bound up in the degree and type of experience we have with them, which in turn affects how well we can recognise them. In particular, what appears to set low familiar/trained-to-familiar voices apart from those that are 'truly familiar' is the ability to generalise from existing stored reference patterns to novel instantiations of a speaker's voice.

The fMRI study in the current thesis used voices that differed in terms of familiarity and personal relevance, and began to explore the brain systems implicated in these processes. However, to gain more of an understanding of some of the familiarity benefits as observed in Experiments 1 and 2 for instance, neuroimaging research in future could capitalise upon multivariate analysis techniques to compare patterns of neural activity for voices that differ in the degree of familiarity. For example, representational similarity analysis (RSA) has been commonly employed to compare the similarity of neural patterns associated with different

conditions to a hypothetical model about how these conditions relate to each other. The idea is that stimuli that are represented similarly should show similar neural patterns of activity in regions that support the process of interest (Kriegeskorte, Mur, & Bandettini, 2008).

A few studies have used RSA to identify regions involved in voice and speech perception. For instance, a study by Tsantani, Kriegeskorte, McGettigan, & Garrido (2019) used fMRI to examine regions that were sensitive to faces, voices, and both faces and voices. RSA was used to construct and compare representational dissimilarity matrices (RDMs), which provide information about the (dis)similarity of neural activation patterns between pairs of conditions/stimuli, for responses to faces and voices, with the hypothesis that regions involved in multimodal representations of identity should show higher similarity (lower dissimilarity) of identity-related responses across the two modalities. The authors also explored whether any regions could discriminate response patterns for faces based on the discriminants computed for pairs of voices, and vice versa. A region in the right pSTS was found to be able to do this, and thus this region was argued to be implicated in representing person identity, incorporating both face and voice information. Thus, RSA was useful here in exploring brain regions implicated in multimodal processing as well as modality-specific regions. In a similar way, this method could be useful for furthering our understanding of familiarity benefits in voice recognition. One recent study that examined familiarity benefits for speech intelligibility is worth discussing here. Holmes and Johnsrude (2021) used RSA to explore brain regions implicated in representing the intelligibility of speech, particularly those underpinning familiar talker benefits commonly observed for understanding speech in noise. The authors hypothesised that if there is a familiarity benefit for comprehending masked speech, neural activity when listening to this speech might be more similar to hearing this speech alone when the talker is familiar, compared with when the talker is unfamiliar. Therefore, the authors expected smaller dissimilarity values (higher similarity) when comparing familiar alone (i.e. unmasked) and familiar masked conditions, relative to dissimilarity values when comparing unfamiliar alone and masked conditions. Multiple regions within the temporal lobes showed this pattern, and thus RSA was useful for determining the neural underpinnings of intelligibility benefits observed in behavioural research.

The experiments in Chapter 2 in this thesis illustrated that personally familiar voices have more robust stored representations and that this can allow for a stable percept of identity to be maintained despite the huge variability that can exist within a single speaker. Similarly to

Holmes and Johnsrude's (2021) study above, one might also expect higher similarity in neural response patterns for familiar listeners in representing identity across highly variable exemplars of an individual's speech. Specifically, neural activation patterns in response to highly personally familiar and less familiar voices could be compared, using high-variability stimulus sets. RSA could be used to compare the similarity of neural response patterns across these highly variable vocalisations, with the expectation that there would be higher similarity between pairs of vocal stimuli produced by a personally familiar speaker, whereas one might expect lower within-talker similarity in the neural responses to lab-trained or unfamiliar voices. For less familiar speakers, high similarity may only be expected for within-speaker comparisons that also share the same vocalisation type (cf Lavan, Scott, & McGettigan, 2016). Searchlight RSA within the right temporal lobes as well as an exploratory whole-brain analysis could be used to identify regions that might underpin the perceptual benefits for recognition of highly familiar voices that I have observed behaviourally.

6.3 Vocal Learning and Social Factors – Possible Interactions?

The second half of this thesis explored the socio-affective qualities associated with particular vocal identities. To do this, Experiments 6 and 7 looked at the motivation to seek out the voice of a valued other (the listener's musical idol), using an effort-based decision-making task. In both experiments, participants exerted more effort for the opportunity to hear their musical idol's voice, suggesting that particular vocal identities can be socially rewarding to listen to. If certain voices are responded to differently, for example displaying differences in their social/motivational properties, does this also have effects on other aspects of processing, such as how voices are encoded, learned, or represented? Familiar voice representations are never purely perceptual - instead familiar voices are encoded as meaningful social stimuli, associated with conceptual information, memories, and emotions (Sidtis & Kreiman, 2012). Yet the recognition of familiar voices is mostly tested in laboratory settings, and voice learning is often removed from social interaction, person knowledge, or social motivation to learn new voices. Therefore, potential interactions between social or affective aspects of voices and voice learning or recognition have not been explored. There is, however, some evidence in the face perception literature suggestive of social or motivational factors on the quality of person representations. Wilson, See, Bernstein, Hugenberg, and Chartier (2014) found that when participants expected future interactions with people in their "outgroup" (whose faces were presented to them in an experiment), recognition of those faces was improved. Similarly,

Hugenberg, Wilson, See, and Young (2013) suggested that social factors, such as perceiver motivation and social importance, as well as prior experience, affect recognition ability and contribute to the own-race bias. The social importance placed upon particular faces may motivate individuals to attend to the specific diagnostic features of these faces in order to be able to recognise them. On the other hand, they argue that it is not useful to individuate faces not personally relevant to us and thus these faces are not worthy of attention (Hugenberg et al., 2013). Therefore, social expectations or intentions may affect the subjective importance placed on being able to recognise particular faces, which in turn impacts the fidelity with which the brain encodes them when first encountered. Taking a different approach, Schwartz and Yovel (2019) also directly explored the effects of social factors and motivation on face recognition. The researchers compared the ability for learning new voices when these were learned as concepts versus percepts. Conceptual encoding involved rating social traits in the faces (e.g. "how intelligent is the face?") whereas perceptual encoding involved focusing on specific facial features (e.g. "how round are the eyes?"). The authors found that focusing on conceptual knowledge led to increased recognition ability, and attributed these findings to the notion that familiar faces are not purely perceptual, but they are associated with conceptual knowledge and social information.

My thesis has explored recognition and motivation for beloved voices, where commitment to these voices was established either via a long-term relationship (romantic couples) or devoted fan activity (musical idols). Whether a social motivation to engage with specific individuals influences earlier development of voice representations remains to be known. Instead of learning to recognise others from disembodied voices or faces, it may be important to consider social factors, remembering that semantic and social information is associated with knowing another person, and that this may not solely be "extra" information but could in fact be an important part of the learning process. Future investigations could explore this by comparing novel to-be-learned voices that have been associated with conceptual, semantic, or social importance to other voices learned solely by voice. Voices are not acquired in isolation, rather they are learnt within a social context, and thus vocal identity perception should be studied as such.

6.4 Integrating Personally Familiar Voice Representations into Current Theoretical Models of Vocal Identity

As outlined previously, several models exist that aim to explain the processing of familiar and unfamiliar voices. A model proposed by Maguinness, Roswandowitz, and von Kriegstein (2018) incorporates elements of both the prevailing prototype model (Lavner, Rosenhouse, & Gath, 2001) and the "auditory face model" (Belin, Fecteau, & Bedard, 2004). This integrative model is particularly relevant to the current thesis as it includes a mechanism for how voices may become familiar over time. Briefly, the model argues that familiar and unfamiliar voice processing separates at the point that an incoming signal is compared to stored reference patterns. As reference patterns do not exist for unfamiliar voices, these need to be established. Reference patterns for recently familiarised or unfamiliar voices are thought to be established via an iterative loop, and signals are subjected to this loop when the voice is not recognised, or if a listener knows from the outset that a voice will be unfamiliar (Maguinness, Roswandowitz, and von Kriegstein, 2018). Repeated iterations through this perceptual processing loop allow for robust representations to be built up over time, and the number of iterations needed is thought to depend on factors such as the distinctiveness of the new voice. Whilst this model provides an explanation as to how an initial weak representation can become a robust stored representation through repeated exposure, there is a need for further specification of this process to account for the wide variation in the degree of familiarity a listener can have with a voice. For instance, the model argues that iterations through the perceptual processing loop will continue until a robust reference pattern is established, and this will then join other stored reference patterns. However, this does not explain the type or frequency of exposure that is necessary for a robust reference pattern to be established, nor does it explicitly define what a robust reference pattern is. Framing the model in terms of different processes for familiar and unfamiliar voices can be useful to understand why dissociations have been found between discrimination and identification after brain injury for example, however, these dissociations are built on a confound. That is, in the neuropsychology literature, unfamiliar voice processing is always tested with discrimination tasks, whereas familiar voice processing is only tested via naming/recognition. This erroneously alludes to familiarity as a binary concept. In Maguinness, Roswandowitz and von Kriegstein's (2018) model, the iterative loop is thought to be accessed only when a voice is not recognised, and a reference pattern does not exist. However, even familiar voices may not be recognised under all circumstances, particularly if the listener has only experienced the voice in certain contexts, and the experiments in this thesis support this idea. Lab-trained voices in this thesis were familiar in that they could be recognised accurately under similar conditions to which they were trained. However, a change in speaking style or other manipulation was detrimental to recognition, highlighting that familiarity is not always

equivalent to accurate recognition under all circumstances. These findings fit into the notion of exemplar-based or episodic processing of voices, that argues that specific instances of a speaker's voice are stored in long-term memory, and recognition is achieved by matching an incoming signal to the nearest matching exemplar (Lavan, Burton, et al., 2019). Thus, to be able to recognise voices robustly and flexibly, particularly in generalising to novel vocalisations, there may need to be a shift from exemplar-based processing to prototype or norm-based coding, and this may be achieved via prolonged exposure and social motivation (Fontaine et al., 2017). Indeed, a high level of personal familiarity – acquired through varied, naturalistic exposure – was associated with highly robust recognition. Whilst not explicitly stated, the model can easily incorporate both the formation of initial stored reference patterns, and the refinement of existing reference patterns to account for differences in the robustness of representations within the range of voices we are familiar with. For example, a familiar voice may be recognised if heard in the contexts a listener is used to hearing it in (e.g. neutral, modal voice), but the underlying representation for this voice may be relatively underspecified. However, the same voice may not be recognised when heard in a new context, and thus although a reference pattern already exists for this speaker, it needs to be updated to incorporate newly experienced vocalisations/exemplars.

A slight clarification to the model could help to better explain the findings in this thesis, and beyond, in relation to the differences in how different types of familiar voices are recognised across different contexts. In the original model, the iterative perceptual processing loop is not required for familiar voices as these voices should be immediately recognised via stored representations. However, instead of the perceptual voice processing loop only being accessed for unfamiliar or newly-familiar voices in need of forming a reference pattern, this loop may also be accessed to update existing reference patterns for familiar voices (see Figure 25 for an updated model). This may interact with face processing systems such that a familiar person may be recognised by their face, whilst hearing an aspect of their voice not encountered previously, and this could allow for the existing voice reference patterns to be updated accordingly - arguably this could be an unconscious and automatic process. Therefore, under this updated model, there may be various possible processes that are engaged upon hearing a voice: If an incoming signal is compared to existing reference patterns and the distance is smaller than the perceptual threshold for recognition, the voice will be recognised as familiar and reference patterns will not be refined. If a speaker is not recognised as familiar by voice, but is recognised via other cues such as the face or content of speech (e.g. "it's me, NAME"),

then the perceptual processing loop may be accessed to update existing reference patterns. Lastly, if a voice is not recognised, and existing reference patterns do not exist, then the process of reference pattern establishment may be engaged. This subtle clarification to the model allows us to understand familiarity as a continuum and to further specify how voice learning is a gradual and continuous process.

There is likely continual refinement of existing reference patterns with continued and varied exposure, in order to encompass within-speaker variability. This helps us to understand the findings in Chapter 2, whereby highly personally familiar voices could be recognised robustly despite the perceptual challenges introduced experimentally. Thus, the robust representations that exist for these voices is proposed to be due to prolonged and varied exposure to familiar speakers, via refinement and updating of reference patterns, potentially in interaction with face and speech processes via social interaction. As voices become increasingly familiar, representations may expand from a singular point to a region in representational space (i.e. the formation of a within-person voice space), allowing for generalisation to novel instances of these voices (Stevenage, Symons, Fletcher, & Coen, 2020). Within our existing voice space, some voices may be represented as singular points, whereas others may be represented as larger areas in space (of varying size and specificity depending on the degree of familiarity) that contain the possible variations in a speaker's voice, with that speaker's individual prototype at the centre, allowing for these representations to be able to tolerate variations in the signal. However, questions still remain as to how within- and between-person voice spaces might interact to produce the robust recognition as observed for highly personally familiar voices observed in the current thesis, and whether a single speaker is represented by a single voice space or multiple voice spaces to reflect different types of vocalisations that may sound very different from each other. Theoretically, it could even be that both norm-based and exemplar-based coding could co-exist for accurately recognising a single speaker. For instance, there may exist a within-speaker voice space, as well as specific exemplars that lie outside of this, e.g. vocalisations that are extreme deviations from the speaker's "normal" voice (i.e. outliers), and that are very rarely experienced.

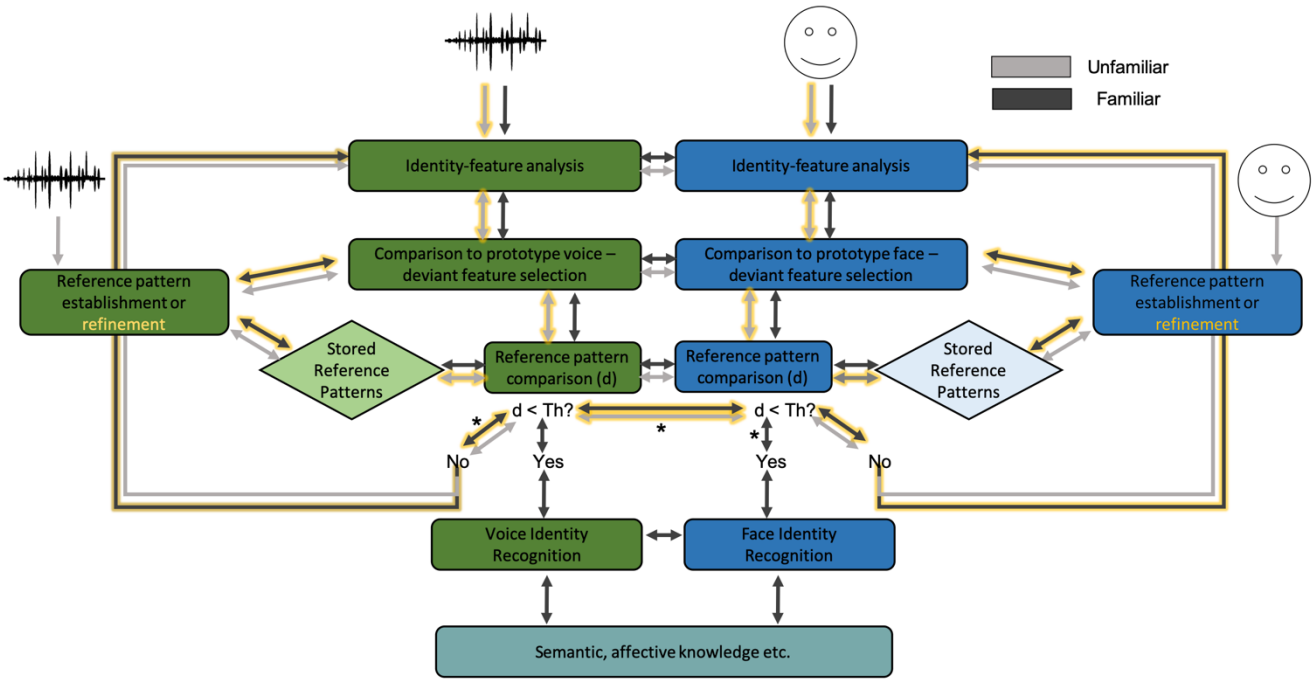

Figure 25. Updated audio-visual integrative model. Changes to the model are highlighted in yellow. Dark grey arrows depict processing for familiar identities, and light grey for unfamiliar identities. For familiar voices, processing either leads to recognition or to the refinement of stored representations. For truly unfamiliar voices, a failure to recognise results in the establishment of reference patterns, as well as in cases where the listener knows from the outset that a voice they are encountering is unfamiliar (e.g. meeting someone new; depicted at the far left and right of the figure). Proposed modifications to the audio-visual integrative model are largely characterised by a greater overlap in familiar and unfamiliar processing. Asterisks (*) in the model highlight a possible route/mechanism by which familiar but unrecognised voices may be incorporated into existing reference patterns, via a successful recognition of the face, and vice versa. Adapted from Maguinness, Roswandowitz, & von Kriegstein (2018).

As well as finding that personally familiar voices could be recognised flexibly, the findings in Experiment 2 also showed that personally familiar listeners began to assign modulated voice tokens belonging to their partner as belonging to another speaker, whilst very rarely perceiving the lab-trained and unfamiliar voices as their romantic partner. In the integrative and prototype models, the distance between the extracted deviations from the prototype are compared to existing stored reference patterns. If the distance is smaller than a perceptual threshold (d' <

Th), the voice is recognised as familiar (Maguinness, Roswandowitz, & von Kriegstein, 2018). Perhaps this stage of the model could account for the findings in Experiment 2 and further explain the nature of the underlying reference patterns. It may be that the perceptual threshold or decision-boundary changes as a function of increasing familiarity. That is, as a listener becomes increasingly familiar with a speaker, there is sharper tuning to the features that represent that individual, and therefore the perceptual distance between the incoming signal and the stored representation needs to be smaller for the listener to identify the incoming signal as a personally familiar other. Figure 26 gives a schematic as to what this may look like conceptually. When a speaker is low- or moderately-familiar, the threshold (Th) is higher and thus deviations with larger distances from existing reference patterns may be recognised as belonging to that individual. Whereas for highly familiar speakers, there needs to be stronger evidence as the listener has a greater sense of what this voice sounds like, and thus the threshold may be lowered. This is supported by previous face perception research using face morphing, whereby there was an observed shift in the categorical boundary in deciding whether a face morphed between the personally familiar other and an unfamiliar identity belonged to the familiar person (Chauhan & Gobbini, 2020). This shift was in the direction of the personally familiar face, meaning that a morph needed to contain a higher percentage of the personally familiar face for it to be labelled as such. The authors argued that increasing familiarity leads to an amplification of perceptual distances between the representation of the familiar face and small changes introduced via e.g. morphing (Chauhan & Gobbini, 2020). Taken together, insights into the recognition of personally familiar voices under perceptually challenging conditions highlights that representations may both be expanded to form within-person voice spaces (as evidenced by Lavan, Knight, & McGettigan, 2019a) that make it possible to recognise a speaker across a range of vocalisations and contexts, whilst at the same time possessing sharper tuning to allow for high accuracy in recognising when an incoming signal does or does not belong to a personally familiar other.

However, it should be noted that there is a potential confound in the findings in Experiment 2 in this thesis that affects this interpretation. That is, recognition of the personally familiar voice was higher than recognition of the lab-trained voice across all conditions in the modulation task. It could be argued that particularly high recognition for the unmodulated voice tokens meant that there was further potential for performance to drop before reaching chance performance compared to shallower "tuning" to lab-trained voices that were closer to, but still above, chance. Therefore, it is possible that this observed change in decision-boundary may

only be an appropriate explanation in particular experimental conditions. Despite this, the types of errors made showed that personally familiar listeners showed an increasing rejection of personally familiar voice tokens that no longer fit their stored representations. Thus, existing models need to be able to explain these processes, and more research is needed to test aspects of these theoretical models with highly personally familiar voices, to further specify and refine our understanding of familiar voice recognition. For instance, if novel voices could be trained sufficiently such that baseline recognition performance was similar to that for personally familiar voices, this would allow for a better comparison for potential differences in "tuning." The training provided to participants in Chapter 3, even in the longer training condition, was not sufficient to do this, and thus training may have to be much more extensive. If there is still sharper tuning for personally familiar voices, this would provide support for this notion of an expanded but precisely defined voice space for the voices we are most familiar with.

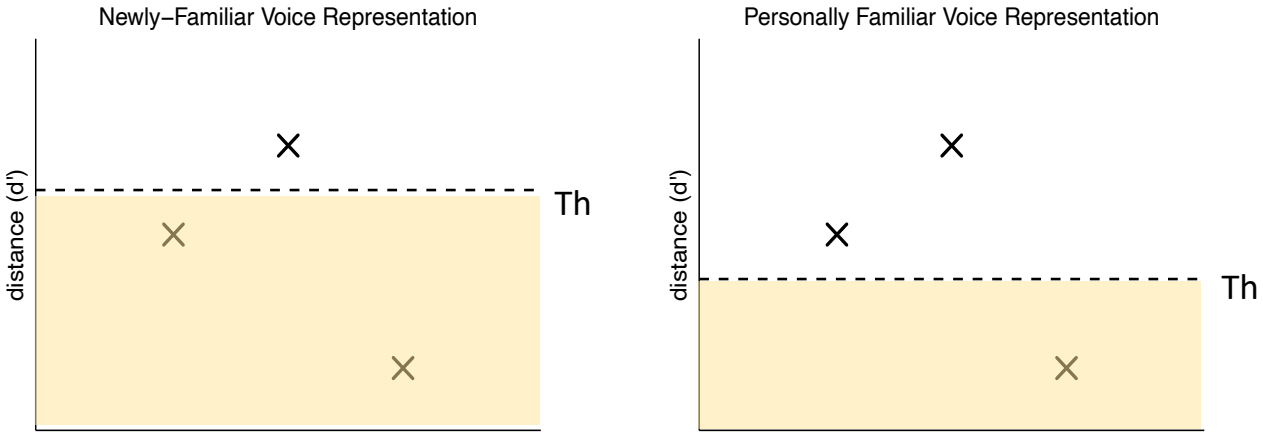

Figure 26. Conceptualisation of familiarity decisions based on the degree of familiarity with a speaker. The distance (d') between the deviations from the prototype and existing reference patterns (depicted by X's above) needs to be smaller than a perceptual threshold (Th) or decision boundary to be recognised as familiar. For newly-familiar voices, this threshold may be higher, and thus incoming signals further away from existing reference patterns may be correctly or incorrectly recognised as familiar (yellow highlighted area), whereas for personally familiar voices, this threshold is lowered, and smaller distances to the stored representation are needed to perceive the voice as a personally familiar other.

6.5 Conclusion

Overall, this thesis highlights that our current understanding of familiar voice processing is incomplete. The definition of a familiar voice can be vastly different depending on the type and range of experiences one has with a speaker, and all familiar voices are not equal in their social and emotional relevance. A high level of familiarity with a speaker's voice acquired naturally is associated with a refined and robust representation, and recognition of these voices is extremely accurate. As familiarity develops, a reliance on low-level acoustic cues may be reduced such that recognition improves. Yet representations robust enough to withstand natural and artificial variability may require prolonged, repeated, and varied exposure to be formed, although this remains to be fully determined. Some familiar voices are deemed personally important to a listener, and this emotional relevance is associated with consequences for the brain and behaviour, characterised by differences in reward and motivational processes. The findings of this thesis therefore provide an updated view of the nature and extent of familiar voice processing, necessitating an approach for prospective work that embraces the full range of possible familiar voices, as well as examining voice processing within a framework that recognises the social and emotional significance of particular vocal identities.

7 References:

Abrams, D. A., Chen, T., Odriozola, P., Cheng, K. M., Baker, A. E., Padmanabhan, A., ... & Menon, V. (2016). Neural circuits underlying mother's voice perception predict social communication abilities in children. *Proceedings of the National Academy of Sciences*, *113*(22), 6295-6300. https://doi.org/10.1073/pnas.1602948113

Abrams, D. A., Lynch, C. J., Cheng, K. M., Phillips, J., Supekar, K., Ryali, S., ... & Menon, V. (2013). Underconnectivity between voice-selective cortex and reward circuitry in children with autism. *Proceedings of the National Academy of Sciences*, *110*(29), 12060-12065. https://doi.org/10.1073/pnas.1302982110

Abrams, D. A., Padmanabhan, A., Chen, T., Odriozola, P., Baker, A. E., Kochalka, J., ... & Menon, V. (2019). Impaired voice processing in reward and salience circuits predicts social communication in children with autism. *Elife*, *8*, e39906. https://doi.org/10.7554/eLife.39906

Acevedo, B. P., Aron, A., Fisher, H. E., & Brown, L. L. (2012). Neural correlates of long-term intense romantic love. *Social Cognitive and Affective Neuroscience*, *7*(2), 145-159. https://doi.org/10.1093/scan/nsq092

Aglieri, V., Cagna, B., Velly, L., Takerkart, S., & Belin, P. (2021). FMRI-based identity classification accuracy in left temporal and frontal regions predicts speaker recognition performance. *Scientific Reports*, *11*(1), 1-13. https://doi.org/10.1038/s41598-020-79922-7

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron*, *32*(3), 537-551. https://doi.org/10.1016/S0896-6273(01)00491-3

Anderson, B. A. (2016). Social reward shapes attentional biases. *Cognitive Neuroscience*, *7*(1-4), 30-36. https://doi.org/10.1080/17588928.2015.1047823

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

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. https://doi.org/10.1080/17470218.2014.1003949

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The anterior cingulate gyrus and social cognition: tracking the motivation of others. *Neuron*, *90*(4), 692-707. https://doi.org/10.1016/j.neuron.2016.04.018

Arias-Carrión, Ó., & Pöppel, E. (2007). Dopamine, learning, and reward-seeking behavior. *Acta Neurobiologiae Experimentalis*, *67*(4), 481–488.

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-González, M., & Pöppel, E. (2010). Dopaminergic reward system: a short integrative review. *International Archives of Medicine*, *3*(1), 1-6. https://doi.org/10.1186/1755-7682-3-24

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. https://doi.org/10.1348/174866407X246131

Aron, A., Fisher, H., Mashek, D. J., Strong, G., Li, H., & Brown, L. L. (2005). Reward, motivation, and emotion systems associated with early-stage intense romantic love. *Journal of Neurophysiology*, *94*(1), 327-337. https://doi.org/10.1152/jn.00838.2004

Asemi, A., Ramaseshan, K., Burgess, A., Diwadkar, V. A., & Bressler, S. L. (2015). Dorsal anterior cingulate cortex modulates supplementary motor area in coordinated

unimanual motor behavior. *Frontiers in Human Neuroscience*, *9*, 309. https://doi.org/10.3389/fnhum.2015.00309

Asutay, E., & Västfjäll, D. (2016). Auditory attentional selection is biased by reward cues. *Scientific Reports*, *6*(1), 1-6. https://doi.org/10.1038/srep36989

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761-770. https://doi.org/10.3758/s13428-011-0075-y

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, *27*(31), 8161-8165. https://doi.org/10.1523/JNEUROSCI.1554-07.2007

Barman, A., Richter, S., Soch, J., Deibele, A., Richter, A., Assmann, A., ... & Schott, B. H. (2015). Gender-specific modulation of neural mechanisms underlying social reward processing by Autism Quotient. *Social Cognitive and Affective Neuroscience*, *10*(11), 1537-1547.https://doi.org/10.1093/scan/nsv044

Barsics, C., & Brédart, S. (2012). Recalling semantic information about newly learned faces and voices. *Memory*, *20*(5), 527-534. https://doi.org/10.1080/09658211.2012.683012

Barsics, C., 2014. Person Recognition Is Easier from Faces than from Voices. *Psychologica Belgica*, 54(3), 244–254. http://doi.org/10.5334/pb.ap

Bartels, A., & Zeki, S. (2000). The neural basis of romantic love. *Neuroreport*, *11*(17), 3829-3834.

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language & the Law*, *22*(2), 229-248. https://doi.org/10.1558/IJSLL.V22I2.23101

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412-414. https://doi.org/10.1098/rsbl.2006.0509

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711-725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129-135. https://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*(16), 2105-2109. https://doi.org/10.1097/00001756-200311140-00019

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17-26. https://doi.org/10.1016/S0926-6410(01)00084-2

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

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35*(7), 1124-1143.

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, *86*(3), 646-664. https://doi.org/10.1016/j.neuron.2015.02.018

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences*, *26*(9), 507-513. https://doi.org/10.1016/S0166-2236(03)00233-9

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist*, *71*(8), 670-679. https://doi.org/10.1037/amp0000059

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), 1-15. https://doi.org/10.1371/journal.pone.0047626

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 61-73. https://doi.org/10.1002/wcs.1266

Bindemann, M., & Johnston, R. A. (2017). Understanding how unfamiliar faces become familiar: Introduction to a special issue on face learning. *Quarterly Journal of Experimental Psychology*, *70*(5), 859-862. https://doi.org/10.1080/17470218.2016.1267235

Bjork, J. M., Knutson, B., Fong, G. W., Caggiano, D. M., Bennett, S. M., & Hommer, D. W. (2004). Incentive-elicited brain activation in adolescents: similarities and differences from young adults. *Journal of Neuroscience*, *24*(8), 1793-1802. https://doi.org/10.1523/JNEUROSCI.4862-03.2004

Blakemore, S. J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, *9*(4), 267-277. https://doi.org/10.1038/nrn2353

Bliss-Moreau, E., Owren, M. J., & Barrett, L. F. (2010). I like the sound of your voice: Affective learning about vocal signals. *Journal of Experimental Social Psychology*, *46*(3), 557-563. https://doi.org/10.1016/j.jesp.2009.12.017

Boersma, P., & Weenink, D. (2010). Praat: Doing phonetics by computer. Phonetic Sciences, University of Amsterdam.

Brédart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology*, *21*(7), 1013-1021. https://doi.org/10.1080/09541440802591821

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002, June). Region of interest analysis using an SPM toolbox. In *8th International Conference on Functional Mapping of the Human Brain* (Vol. 16, No. 2, p. 497).

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. https://doi.org/10.1121/1.1910246

Brock, J. (2013). Connectivity and cognition in autism spectrum disorders: Where are the links? *Proceedings of the National Academy of Sciences*, *110*(42), E3973-E3973. https://doi.org/10.1073/pnas.1311907110

Brockett, A. T., & Roesch, M. R. (2021). The ever-changing OFC landscape: What neural signals in OFC can tell us about inhibitory control. *Behavioral Neuroscience*, *135*(2), 129-137. https://doi.org/10.1037/bne0000412

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

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

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223. https://doi.org/10.1111/cogs.12231

Camalier, C. R., & Kaas, J. H. (2011). Sound. In J.A. Gottfried (ed.), *Neurobiology of Sensation and Reward* (pp.183-199), CRC Press.

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

Cao, Z., Bennett, M., Orr, C., Icke, I., Banaschewski, T., Barker, G. J., ... & IMAGEN Consortium. (2019). Mapping adolescent reward anticipation, receipt, and prediction error during the monetary incentive delay task. *Human Brain Mapping*, *40*(1), 262-283. https://doi.org/10.1002/hbm.24370

Carter, R. M., MacInnes, J. J., Huettel, S. A., & Adcock, R. A. (2009). Activation in the VTA and nucleus accumbens increases in anticipation of both gains and losses. *Frontiers in Behavioral Neuroscience*, *3*(21), 1-15. https://doi.org/10.3389/neuro.08.021.2009

Cazzoli, D., Kaufmann, B. C., Paladini, R. E., Müri, R. M., Nef, T., & Nyffeler, T. (2021). Anterior insula and inferior frontal gyrus: where ventral and dorsal visual attention systems meet. *Brain Communications*, *3*(1), 1-6. https://doi.org/10.1093/braincomms/fcaa220

Chase, H. W., Michael, A., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2010). Paradoxical enhancement of choice reaction time performance in patients with major depression. *Journal of Psychopharmacology*, *24*(4), 471-479. https://doi.org/10.1177/0269881109104883

Chauhan, V., Kotlewska, I., Tang, S., & Gobbini, M. I. (2020). How familiarity warps representation in the face space. *Journal of Vision*, *20*(7), 1-15. https://doi.org/10.1167/jov.20.7.18

Chenot, Q., Lepron, E., De Boissezon, X., & Scannella, S. (2021). Functional Connectivity Within the Fronto-Parietal Network Predicts Complex Task Performance: A fNIRS Study. *Frontiers in Neuroergonomics*, 22(718176), 1-14. https://doi.org/10.3389/fnrgo.2021.718176

Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of Cognitive Neuroscience*, *20*(6), 941-951. https://doi.org/10.1162/jocn.2008.20062

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994. https://doi.org/10.1068/p3335

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. doi:10.1177/001316446002000104.

Couto, B., Manes, F., Montañés, P., Matallana, D., Reyes, P., Velázquez, M., ... & Ibáñez, A. (2013). Structural neuroimaging of social cognition in progressive non-fluent aphasia and behavioral variant of frontotemporal dementia. *Frontiers in Human Neuroscience*, *7*(467), 1-11. https://doi.org/10.3389/fnhum.2013.00467

Cremers, H. R., Veer, I. M., Spinhoven, P., Rombouts, S. A., & Roelofs, K. (2015). Neural sensitivity to social reward and punishment anticipation in social anxiety disorder. *Frontiers in Behavioral Neuroscience*, *8*(439), 1-9. https://doi.org/10.3389/fnbeh.2014.00439

Daniel, R., & Pollmann, S. (2014). A universal role of the ventral striatum in reward-based learning: evidence from human studies. *Neurobiology of Learning and Memory*, *114*, 90-100. https://doi.org/10.1016/j.nlm.2014.05.002

Davey, C. G., Allen, N. B., Harrison, B. J., Dwyer, D. B., & Yücel, M. (2010). Being liked activates primary reward and midline self-related brain regions. *Human Brain Mapping*, *31*(4), 660-668. https://doi.org/10.1002/hbm.20895

Delmonte, S., Balsters, J. H., McGrath, J., Fitzgerald, J., Brennan, S., Fagan, A. J., & Gallagher, L. (2012). Social and monetary reward processing in autism spectrum disorders. *Molecular Autism*, *3*(1), 1-13. https://doi.org/10.1186/2040-2392-3-7

Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied, 26*(2), 236–247. https://doi.org/10.1037/xap0000247

Dreher, J. C., Schmidt, P. J., Kohn, P., Furman, D., Rubinow, D., & Berman, K. F. (2007). Menstrual cycle phase modulates reward-related neural function in women. *Proceedings of the National Academy of Sciences*, *104*(7), 2465-2470. https://doi.org/10.1073/pnas.0605569104

Driscoll ME, Bollu PC, Tadi P. Neuroanatomy, Nucleus Caudate. [Updated 2021 Jul 31]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557407/

Dutra, S. J., Cunningham, W. A., Kober, H., & Gruber, J. (2015). Elevated striatal reactivity across monetary and social rewards in bipolar I disorder. *Journal of Abnormal Psychology*, *124*(4), 890–904. https://doi.org/10.1037/abn0000092

Fareri, D. S., & Delgado, M. R. (2014). Social rewards and social networks in the human brain. *The Neuroscientist*, *20*(4), 387-402. https://doi.org/10.1177/1073858414521869

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Feinberg, D. R., & Cook, O. (2020, September 15). VoiceLab: Automated Reproducible Acoustic Analysis. https://doi.org/10.31234/osf.io/v5uxf

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, *102*(2), 1213-1222. https://doi.org/10.1121/1.421048

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology*, *8*(1180), 1-9. https://doi.org/10.3389/fpsyg.2017.01180

Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. (2009, September). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. In *Interspeech 2009*. https://doi.org/10.21437/Interspeech.2009-54

Ghaziri, J., Tucholka, A., Girard, G., Boucher, O., Houde, J. C., Descoteaux, M., ... & Nguyen, D. K. (2018). Subcortical structural connectivity of insular subregions. *Scientific Reports*, *8*(1), 1-12. https://doi.org/10.1038/s41598-018-26995-0

Gossen, A., Groppe, S. E., Winkler, L., Kohls, G., Herrington, J., Schultz, R. T., ... & Spreckelmeyer, K. N. (2014). Neural evidence for an association between social proficiency and sensitivity to social reward. *Social Cognitive and Affective Neuroscience*, *9*(5), 661-670. https://doi.org/10.1093/scan/nst033

Graff-Radford, J., Williams, L., Jones, D. T., & Benarroch, E. E. (2017). Caudate nucleus as a component of networks controlling behavior. *Neurology*, *89*(21), 2192-2197. https://doi.org/10.1212/WNL.0000000000004680

Grahn, J. A., Parkinson, J. A., & Owen, A. M. (2008). The cognitive functions of the caudate nucleus. *Progress in Neurobiology*, *86*(3), 141-155. https://doi.org/10.1016/j.pneurobio.2008.09.004

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: the role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience*, *7*(523), 1-12. https://doi.org/10.3389/fnhum.2013.00523

Guntupalli, J. S., & Gobbini, M. I. (2017). Reading faces: From features to recognition. *Trends in Cognitive Sciences, 21*(12), 915–916. https://doi.org/10.1016/j.tics.2017.09.007

Halahakoon, D. C., Kieslich, K., O'Driscoll, C., Nair, A., Lewis, G., & Roiser, J. P. (2020). Reward-processing behavior in depressed participants relative to healthy volunteers: A systematic review and meta-analysis. *JAMA Psychiatry*, *77*(12), 1286-1295. https://doi.org/10.1001/jamapsychiatry.2020.2139

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: Inhibition and attentional control. *Neuroimage*, *50*(3), 1313-1319. https://doi.org/10.1016/j.neuroimage.2009.12.109

Hanley, J. R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory*, *17*(8), 830-839. https://doi.org/10.1080/09658210903264175

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 179-195. https://doi.org/10.1080/713755751

Hatfield, E., & Sprecher, S. (1986). Measuring passionate love in intimate relationships. *Journal of Adolescence*, *9*(4), 383-410. https://doi.org/10.1016/S0140-1971(86)80043-4

Hepper, P. G., Scott, D., & Shahidullah, S. (1993). Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, *11*(3), 147-153. https://doi.org/10.1080/02646839308403210

Herzmann, G., Schweinberger, S. R., Sommer, W., & Jentzsch, I. (2004). What's special about personally familiar faces? A multimodal approach. *Psychophysiology*, *41*(5), 688-701. https://doi.org/10.1111/j.1469-8986.2004.00196.x

Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(8), 1465–1476. https://doi.org/10.1037/xlm0000823

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, *29*(10), 1575-1583. https://doi.org/10.1177/0956797618779083

Holmes, E., To, G., & Johnsrude, I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to

voice training. *Psychological Science*, *32*(6), 903-915. https://doi.org/10.1177/0956797621991137

Horn, M., Jardri, R., D'Hondt, F., Vaiva, G., Thomas, P., & Pins, D. (2016). The multiple neural networks of familiarity: A meta-analysis of functional imaging studies. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(1), 176-190. https://doi.org/10.3758/s13415-015-0392-1

Huckvale, M., & Kristiansen, A. L. (2012, June). Effectiveness of electronic voice disguise between friends. In *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*. Audio Engineering Society.

Hugenberg, K., Wilson, J. P., See, P. E., & Young, S. G. (2013). Towards a synthetic model of own group biases in face memory. *Visual Cognition*, *21*(9-10), 1392-1417. https://doi.org/10.1080/13506285.2013.821429

Hung, L. W., Neuner, S., Polepalli, J. S., Beier, K. T., Wright, M., Walsh, J. J., ... & Malenka, R. C. (2017). Gating of social reward by oxytocin in the ventral tegmental area. *Science*, *357*(6358), 1406-1411. https://doi.org/10.1126/science.aan4994

Husain, M., & Roiser, J. P. (2018). Neuroscience of apathy and anhedonia: a transdiagnostic approach. *Nature Reviews Neuroscience*, *19*(8), 470-484. https://doi.org/10.1038/s41583-018-0029-9

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284-294. https://doi.org/10.1016/j.neuron.2008.03.020

Jack, R. E., & Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology*, *25*(14), R621-R634. https://doi.org/10.1016/j.cub.2015.05.052

Jaensch, M., van den Hurk, W., Dzhelyova, M., Hahn, A. C., Perrett, D. I., Richards, A., & Smith, M. L. (2014). Don't look back in anger: The rewarding value of a female face is discounted by an angry expression. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2101–2105. https://doi.org/10.1037/a0038078

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

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995-2004. https://doi.org/10.1177/0956797613482467

Kawahara, H., & Irino, T. (2004), "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation," in Speech separation by humans and machines, edited by P.L. Divenyi (Kluwer Academic, Massachusetts), pp. 167–180. https://doi.org/10.1007/0-387-22794-6_11

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *18*(3), 327-336. https://doi.org/10.1002/acp.974

Kirsch, P., Schienle, A., Stark, R., Sammer, G., Blecker, C., Walter, B., ... & Vaitl, D. (2003). Anticipation of reward in a nonaversive differential conditioning paradigm and the brain reward system: An event-related fMRI study. *Neuroimage*, *20*(2), 1086-1095. https://doi.org/10.1016/S1053-8119(03)00381-1

Knight, S., Lavan, N., Torre, I., & McGettigan, C. (2021). The influence of perceived vocal traits on trusting behaviours in an economic game. *Quarterly Journal of Experimental Psychology*, *74*(10), 1747-1754. https://doi.org/10.1177/17470218211010144

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16), RC159-RC159. https://doi.org/10.1523/JNEUROSCI.21-16-j0002.2001

Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with

rapid event-related fMRI. *Neuroimage*, *18*(2), 263-272. https://doi.org/10.1016/S1053-8119(02)00057-5

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage*, *12*(1), 20-27. https://doi.org/10.1006/nimg.2000.0593

Kobayashi, M., Watanabe, K., and Nakamura, K. (2020) "Attractive faces are rewarding irrespective of face category: Motivation in viewing attractive faces in Japanese viewers," *12th International Conference on Knowledge and Smart Technology (KST)*, 2020, pp. 186-190, https://doi.org/10.1109/KST48564.2020.9059336.

Kok, R., Taubert, J., Van der Burg, E., Rhodes, G., & Alais, D. (2017). Face familiarity promotes stable identity recognition: exploring face perception using serial dependence. *Royal Society Open Science*, *4*(3), 1-13. https://doi.org/10.1098/rsos.160685

Kollmann, B., Scholz, V., Linke, J., Kirsch, P., & Wessa, M. (2017). Reward anticipation revisited-evidence from an fMRI study in euthymic bipolar I patients and healthy first-degree relatives. *Journal of Affective Disorders*, *219*, 178-186. https://doi.org/10.1016/j.jad.2017.04.044

Koranyi, N., Brückner, E., Jäckel, A., Grigutsch, L. A., & Rothermund, K. (2020). Dissociation between wanting and liking for coffee in heavy drinkers. *Journal of Psychopharmacology*, *34*(12), 1350-1356.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* West Sussex: Wiley-Blackwell.

Kreitewolf, J., Gaudrain, E., & von Kriegstein, K. (2014). A neural mechanism for recognizing speech spoken by different speakers. *Neuroimage*, *91*, 375-385. https://doi.org/10.1016/j.neuroimage.2014.01.005

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology*, *8*(1584), 1-8. https://doi.org/10.3389/fpsyg.2017.01584

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1-28. https://doi.org/10.3389/neuro.06.004.2008

Krix, A. C., Sauerland, M., & Schreuder, M. J. (2017). Masking the identities of celebrities and personally familiar individuals: Effects on visual and auditory recognition performance. *Perception*, *46*(10), 1133-1150. https://doi.org/10.1177/0301006617710621

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar representations of emotions across faces and voices. *Emotion*, *17*(6), 912. http://dx.doi.org/10.1037/emo0000282

Kurniawan, I. T., Guitart-Masip, M., Dayan, P., & Dolan, R. J. (2013). Effort and valuation in the brain: the effects of anticipation and execution. *Journal of Neuroscience*, *33*(14), 6160-6169. https://doi.org/10.1523/JNEUROSCI.4777-12.2013

Lammert, A. C., & Narayanan, S. S. (2015). On short-time estimation of vocal tract length from formant frequencies. *PloS one*, *10*(7), e0132193. https://doi.org/10.1371/journal.pone.0132193

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*(175), 1-12. https://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143-R145. https://doi.org/10.1016/j.cub.2010.12.033

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

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

Lavan, N., & McGettigan, C. (2019). Toward a unified account of person perception from familiar and unfamiliar voices. *PsyArXiv,* https://doi.org/10.31234/osf.io/shxa6

Lavan, N., Burston, L. F., & Garrido, L. (2019a). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576-593. https://doi.org/10.1111/bjop.12348

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019b). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology*, *72*(9), 2240-2248. https://doi.org/10.1177/1747021819836890

Lavan, N., Burton, A., Scott, S.K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin, & Review, 26*(1), 90–102, https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Knight, S., & McGettigan, C. (2019a). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 1-9. https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019b). The effects of high variability training on voice identity learning. *Cognition*, *193*, 1-9. https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Merriman, S. E., Ladwa, P., Burston, L. F., Knight, S., & McGettigan, C. (2020). 'Please sort these voice recordings into 2 identities': Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology*, *111*(3), 556-569. https://doi.org/10.1111/bjop.12416

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(12), 1604–1614. https://doi.org/10.1037/xge0000223

Lavan, N., Short, B., Wilding, A., & McGettigan, C. (2018). Impoverished encoding of speaker identity in spontaneous laughter. *Evolution and Human Behavior, 39*(1), 139–145. https://doi.org/10.1016/j.evolhumbehav.2017.11.002

Lavin, C., Melis, C., Mikulan, E. P., Gelormini, C., Huepe, D., & Ibañez, A. (2013). The anterior cingulate cortex: an integrative hub for human socially-driven interactions. *Frontiers in neuroscience*, *7*(64), 1-4. https://doi.org/10.3389/fnins.2013.00064

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30*(1), 9–26. https://doi.org/10.1016/S0167-6393(99)00028-X

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

Lee, S. H., Shim, H. J., Yoon, S. W., & Lee, K. W. (2009). Effects of Various Background Noises on Speech Intelligibility of Normal Hearing Subjects. *Korean Journal of Otorhinolaryngology-Head and Neck Surgery*, *52*(4), 307-311. https://doi.org/10.3342/kjorl-hns.2009.52.4.307

Lee, T., Leung, M. K., Lee, T. M., Raine, A., & Chan, C. C. (2013). I want to lie about not knowing you, but my precuneus refuses to cooperate. *Scientific Reports*, *3*(1), 1-5. https://doi.org/10.1038/srep01636

Leibenluft, E., Gobbini, M. I., Harrison, T., & Haxby, J. V. (2004). Mothers' neural activation in response to pictures of their children and other children. *Biological Psychiatry*, *56*(4), 225-232. https://doi.org/10.1016/j.biopsych.2004.05.017

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.3.01. https://CRAN.R-project.org/package=emmeans

Leotti, L. A., & Delgado, M. R. (2011). The inherent reward of choice. *Psychological Science*, *22*(10), 1310-1318. https://doi.org/10.1177/0956797611417005

Levi, S. V. (2017). Another bilingual advantage? Perception of talker-voice information. *Bilingualism: Language and Cognition, 21*(3), 523-536. https://doi.org/10.1017/S1366728917000153

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053-4062. https://doi.org/10.1121/1.3651816

Liccione, D., Moruzzi, S., Rossi, F., Manganaro, A., Porta, M., Nugrahaningsih, N., ... & Allegri, N. (2014). Familiarity is not notoriety: phenomenological accounts of face recognition. *Frontiers in Human Neuroscience*, *8*(672), 1-10. https://doi.org/10.3389/fnhum.2014.00672

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1219-1236. https://doi.org/10.1016/j.neubiorev.2010.12.012

López, S., Riera, P., Assaneo, M. F., Eguía, M., Sigman, M., & Trevisan, M. A. (2013). Vocal caricatures reveal signatures of speaker identity. *Scientific Reports*, *3*(1), 1-7. https://doi.org/10.1038/srep03407

Luthra, S. (2021). The role of the right hemisphere in processing phonetic variability between talkers. *Neurobiology of Language, 2*(1), 138–151. https://doi.org/10.1162/nol_a_00028

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS One*, *4*(3), 1-14. https://doi.org/10.1371/journal.pone.0004638

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179-193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Marche, K., Martel, A. C., & Apicella, P. (2017). Differences between dorsal and ventral striatum in the sensitivity of tonically active neurons to rewarding events. *Frontiers in Systems Neuroscience*, *11*(52), 1-12. https://doi.org/10.3389/fnsys.2017.00052

Martins, D., Rademacher, L., Gabay, A. S., Taylor, R., Richey, J. A., Smith, D. V., ... & Paloyelis, Y. (2021). Mapping social reward and punishment processing in the human

brain: A voxel-based meta-analysis of neuroimaging findings using the social incentive delay task. *Neuroscience & Biobehavioral Reviews*, *122*, 1-17. https://doi.org/10.1016/j.neubiorev.2020.12.034

Mattarozzi, K., Colonnello, V., Russo, P. M., & Todorov, A. (2019). Person information facilitates memory for face identity. *Psychological Research*, *83*(8), 1817-1824. https://doi.org/10.1007/s00426-018-1037-0

Matyjek, M., Meliss, S., Dziobek, I., & Murayama, K. (2020). A multidimensional view on social and non-social rewards. *Frontiers in Psychiatry*, 11(818), 1-8. https://doi.org/10.3389/fpsyt.2020.00818

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS One*, *9*(3), 1-9. https://doi.org/10.1371/journal.pone.0090779

McGettigan, C. (2015). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience, 9*(129), 1-4. https://doi.org/10.3389/fnhum.2015.00129

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure and Function*, *214*(5), 655-667.

Mitchell, R. L., & Ross, E. D. (2013). Attitudinal prosody: What we know and directions for future study. *Neuroscience & Biobehavioral Reviews*, *37*(3), 471-479. https://doi.org/10.1016/j.neubiorev.2013.01.027

Mumford, J. A., Poline, J. B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PloS One*, *10*(4), 1 – 11. https://doi.org/10.1371/journal.pone.0126255

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, *68*(5), 1294-1300. https://doi.org/10.1121/1.385122

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., ... & Kojima, S. (2001). Neural substrates for recognition of familiar voices: a PET

study. *Neuropsychologia*, *39*(10), 1047-1054. https://doi.org/10.1016/S0028-3932(01)00037-9

Nestler, E. J., Hyman, S. E., Holtzman, M. D., Malenka, R.C. (2014). *Molecular neuropharmacology: A foundation for clinical neuroscience.* McGraw Hill.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85-103. https://doi.org/10.1016/j.wocn.2005.10.004

Novak, B. K., Novak, K. D., Lynam, D. R., & Foti, D. (2016). Individual differences in the time course of reward processing: stage-specific links with depression and impulsivity. *Biological Psychology*, *119*, 79-90. https://doi.org/10.1016/j.biopsycho.2016.07.008

Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology: Applied*, *25*(2), 280–290. https://doi.org/10.1037/xap0000213

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355-376. https://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42-46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yücel, M., & Lorenzetti, V. (2018). The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human Brain Mapping*, *39*(8), 3398-3418. https://doi.org/10.1002/hbm.24184

Olff, M., Frijling, J. L., Kubzansky, L. D., Bradley, B., Ellenbogen, M. A., Cardoso, C., ... & Van Zuiden, M. (2013). The role of oxytocin in social bonding, stress regulation and mental health: an update on the moderating effects of context and interindividual differences. *Psychoneuroendocrinology*, *38*(9), 1883-1894. https://doi.org/10.1016/j.psyneuen.2013.06.019

Ortigue, S., Bianchi-Demicheli, F., Hamilton, A. D. C., & Grafton, S. T. (2007). The neural basis of love as a subliminal prime: an event-related functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, *19*(7), 1218-1230. https://doi.org/10.1162/jocn.2007.19.7.1218

Patriquin, M. A., DeRamus, T., Libero, L. E., Laird, A., & Kana, R. K. (2016). Neuroanatomical and neurofunctional markers of social cognition in autism spectrum disorder. *Human Brain Mapping*, *37*(11), 3957-3978. https://doi.org/10.1002/hbm.23288

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., ... & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, *119*(1), 164-174. https://doi.org/10.1016/j.neuroimage.2015.06.050

Perrachione, T. (2018). Recognizing Speakers Across Languages. In S. Frühholz, & P. Belin (eds). *The Oxford Handbook of Voice Perception*, 514-538. doi: https://doi.org/10.1093/oxfordhb/9780198743187.013.23

Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice*, *31*(2), 256-e13 – 256.e17. http://dx.doi.org/10.1016/j.jvoice.2016.06.004

Piva, M., Velnoskey, K., Jia, R., Nair, A., Levy, I., & Chang, S. W. (2019). The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *Elife*, *8*, e44939. https://doi.org/10.7554/eLife.44939.001

Plante-Hébert, J., Boucher, V. J., & Jemel, B. (2021). The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *Plos One*, *16*(4), 1-20. https://doi.org/10.1371/journal.pone.0250214

Puglisi-Allegra, S., & Ventura, R. (2012). Prefrontal/accumbal catecholamine system processes high motivational salience. *Frontiers in Behavioral Neuroscience*, *6*(31), 1-13. https://doi.org/10.3389/fnbeh.2012.00031

Purhonen, M., Kilpeläinen-Lees, R., Valkonen-Korhonen, M., Karhu, J., & Lehtonen, J. (2004). Cerebral processing of mother's voice compared to unfamiliar voice in 4-

month-old infants. *International Journal of Psychophysiology*, *52*(3), 257-266. https://doi.org/10.1016/j.ijpsycho.2003.11.003

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rademacher, L., Krach, S., Kohls, G., Irmak, A., Gründer, G., & Spreckelmeyer, K. N. (2010). Dissociation of neural networks for anticipation and consumption of monetary and social rewards. *Neuroimage*, *49*(4), 3276-3285. https://doi.org/10.1016/j.neuroimage.2009.10.089

Rademacher, L., Salama, A., Gründer, G., & Spreckelmeyer, K. N. (2014). Differential patterns of nucleus accumbens activation during anticipation of monetary and social reward in young and older adults. *Social Cognitive and Affective Neuroscience*, *9*(6), 825-831. https://doi.org/10.1093/scan/nst047

Ramon, M., & Gobbini, M. I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. *Visual Cognition*, *26*(3), 179-195. https://doi.org/10.1080/13506285.2017.1405134

Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour*, *65*(3), 519-530. https://doi.org/10.1006/anbe.2003.2078

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *i-Perception*, *8*(5),1-21. https://doi.org/10.1177/2041669517731115

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America*, *66*(4), 1023-1028. https://doi.org/10.1121/1.383321

Rice, K., & Redcay, E. (2016). Interaction matters: A perceived social partner alters the neural processing of human speech. *NeuroImage*, *129*, 480-488. https://doi.org/10.1016/j.neuroimage.2015.11.041

Rigney, A. E., Koski, J. E., & Beer, J. S. (2018). The functional role of ventral anterior cingulate cortex in social evaluation: disentangling valence from subjectively

rewarding opportunities. *Social Cognitive and Affective Neuroscience*, *13*(1), 14-21. https://doi.org/10.1093/scan/nsx132

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology*, *7*(6), 475-481. https://doi.org/10.1002/acp.2350070603

Rolls, E. T., Cheng, W., & Feng, J. (2020). The orbitofrontal cortex: reward, emotion and depression. *Brain Communications*, *2*(2), 1-25. https://doi.org/10.1093/braincomms/fcaa196

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147-1153. https://doi.org/10.1093/cercor/bhl024

Salimpoor, V. N., van den Bosch, I., Kovacevic, N., McIntosh, A. R., Dagher, A., & Zatorre, R. J. (2013). Interactions between the nucleus accumbens and auditory cortices predict music reward value. *Science*, *340*(6129), 216-219. https://doi.org/10.1126/science.1231059

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology*, *65*(1), 111–116. https://doi.org/10.1037/0021-9010.65.1.111

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice identity recognition: functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280-291. https://doi.org/10.1162/jocn_a_00707

Scheele, D., Wille, A., Kendrick, K. M., Stoffel-Wagner, B., Becker, B., Güntürkün, O., ... & Hurlemann, R. (2013). Oxytocin enhances brain reward system responses in men viewing the face of their female partner. *Proceedings of the National Academy of Sciences*, *110*(50), 20308-20313. https://doi.org/10.1073/pnas.1314190110

Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological Reviews*, *95*(3), 853-951.https://doi.org/10.1152/physrev.00023.2014

Schwartz, L., & Yovel, G. (2019). Learning faces as concepts rather than percepts improves face recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1733–1747. https://doi.org/10.1037/xlm0000673

Schweinberger, S. R. (2001). Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia*, *39*(9), 921-936. https://doi.org/10.1016/S0028-3932(01)00023-9

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453-463. https://doi.org/10.1044/jslhr.4002.453

Seltzer, L. J., Prososki, A. R., Ziegler, T. E., & Pollak, S. D. (2012). Instant messages vs. speech: hormones and why we still need to hear each other. *Evolution and Human Behavior*, *33*(1), 42-45. https://doi.org/10.1016/j.evolhumbehav.2011.05.004

Seltzer, L. J., Ziegler, T. E., & Pollak, S. D. (2010). Social vocalizations can release oxytocin in humans. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1694), 2661-2666. https://doi.org/10.1098/rspb.2010.0567

Sescousse, G., Caldú, X., Segura, B., & Dreher, J. C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *37*(4), 681-696. https://doi.org/10.1016/j.neubiorev.2013.02.002

Shah, N. J., Marshall, J. C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H. J., et al. (2001). The neural correlates of person familiarity—a functional magnetic resonance imaging study with clinical implications. *Brain* 124, 804–815. https://doi.org/10.1093/brain/124.4.804

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, *46*(2), 146-159. https://doi.org/10.1007/s12124-011-9177-4

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177-3186. https://doi.org/10.1121/1.2047107

Smith, D. V., & Delgado, M. R. (2015). Reward Processing. In A. W. Toga (Ed.), Brain Mapping: An Encyclopedic Reference (1 ed., Vol. 3, pp. 361-366). Waltham, MA: Academic Press. https://doi.org/10.1016/B978-0-12-397025-1.00255-4

Smith, H. M., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272-287. https://doi.org/10.1002/acp.3478

Sørensen, M. H. (2012). Voice line-ups: Speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech Language and the Law*, *19*(2), 145–158. https://doi.org/10.1558/ijsll.v19i2.145

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(08), 689-700. https://doi.org/10.3766/jaaa.24.8.6

Spreckelmeyer, K. N., Krach, S., Kohls, G., Rademacher, L., Irmak, A., Konrad, K., ... & Gründer, G. (2009). Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women. *Social Cognitive and Affective Neuroscience*, *4*(2), 158-165. https://doi.org/10.1093/scan/nsn051

Stark, R., Bauer, E., Merz, C. J., Zimmermann, M., Reuter, M., Plichta, M. M., ... & Herrmann, M. J. (2011). ADHD related behaviors are associated with brain activation in the reward system. *Neuropsychologia*, *49*(3), 426-434. https://doi.org/10.1016/j.neuropsychologia.2010.12.012

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology*, *89*(1), 39-57. https://doi.org/10.1111/j.2044-8295.1998.tb02672.x

Stevenage, S. V., Symons, A. E., Fletcher, A., & Coen, C. (2020). Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting

task. *Quarterly Journal of Experimental Psychology*, *73*(4), 519-536. https://doi.org/10.1177/1747021819888064

Stevens, F. L., Hurley, R. A., & Taber, K. H. (2011). Anterior cingulate cortex: unique role in cognition and emotion. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *23*(2), 121-125. https://doi.org/10.1176/jnp.23.2.jnp121

Sugiura, M. (2014). Neuroimaging studies on recognition of personally familiar people. *Frontiers in Bioscience*, *19*(1), 672-686. https://doi.org/10.2741/4235

Tamir, D. I., & Hughes, B. L. (2018). Social rewards: from basic social building blocks to complex social behavior. *Perspectives on Psychological Science*, *13*(6), 700-717. https://doi.org/10.1177/1745691618776263

Tang, W., Jbabdi, S., Zhu, Z., Cottaar, M., Grisot, G., Lehman, J. F., ... & Haber, S. N. (2019). A connectional hub in the rostral anterior cingulate cortex links areas of emotion and cognitive control. *Elife*, *8*, 1-25. https://doi.org/10.7554/eLife.43761

Taylor, M. J., Arsalidou, M., Bayless, S. J., Morris, D., Evans, J. W., & Barbeau, E. J. (2009). Neural correlates of personally familiar faces: parents, partner and own faces. *Human Brain Mapping*, *30*(7), 2008-2020. https://doi.org/10.1002/hbm.20646

Tong, F., & Nakayama, K. (1999). Robust representations for faces: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1016–1035. https://doi.org/10.1037/0096-1523.25.4.1016

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*, 106215. https://doi.org/10.1016/j.chb.2019.106215

Torre, I., Goslin, J., White, L. (2015, August 10–14). Investing in accents: How does experience mediate trust attributions to different voices? [Conference session]. Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015), Glasgow, UK.

Torre, I., White, L., Goslin, J. (2016, May). Behavioural mediation of prosodic cues to implicit judgements of trustworthiness [Conference session]. Proceedings of the Eighth International Conference on Speech Prosody 2016, Boston, MA, United States.

Tsantani, M., Kriegeskorte, N., McGettigan, C., & Garrido, L. (2019). Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. *NeuroImage*, *201*, 116004. https://doi.org/10.1016/j.neuroimage.2019.07.017

Tso, I. F., Rutherford, S., Fang, Y., Angstadt, M., & Taylor, S. F. (2018). The "social brain" is highly sensitive to the mere presence of social information: An automated meta-analysis and an independent study. *PloS One*, *13*(5), 1-13. https://doi.org/10.1371/journal.pone.0196503

Tye-Murray, N., Spehar, B., Sommers, M., & Barcroft, J. (2016). Auditory training with frequent communication partners. *Journal of Speech, Language, and Hearing Research*, *59*(4), 871-875. https://doi.org/10.1044/2016_JSLHR-H-15-0171

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*(1), 55-61. https://doi.org/10.1038/nrn3857

Uddin, L. Q., Yeo, B. T., & Spreng, R. N. (2019). Towards a universal taxonomy of macro-scale functional human brain networks. *Brain Topography*, *32*(6), 926-942. https://doi.org/10.1007/s10548-019-00744-6

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. https://doi.org/10.1080/14640749108400966

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex*, *24*(2), 195-209. https://doi.org/10.1016/S0010-9452(88)80029-7

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologica, 25*(5), 829-834, https://doi.org/10.1016/0028-3932(87)90120-5

Visconti di Oleggio Castello, M., Guntupalli, J. S., Yang, H., & Gobbini, M. I. (2014). Facilitated detection of social cues conveyed by familiar faces. *Frontiers in Human Neuroscience*, *8*(678), 1-11. https://doi.org/10.3389/fnhum.2014.00678

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. *Cognitive Brain Research*, *17*(1), 48-55. https://doi.org/10.1016/S0926-6410(03)00079-X

Von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), 1089 - 1820. https://doi.org/10.1371/journal.pbio.0040326

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

Von Kriegstein, K., Smith, D. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience*, *30*(2), 629-638. https://doi.org/10.1523/JNEUROSCI.2742-09.2010

Von Kriegstein, K., Warren, J. D., Ives, D. T., Patterson, R. D., & Griffiths, T. D. (2006). Processing the acoustic effect of size in speech sounds. *Neuroimage*, *32*(1), 368-375. https://doi.org/10.1016/j.neuroimage.2006.02.045

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3-28. https://doi.org/10.1007/BF00987006

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. In *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco* (pp. 1381-1385).

Walton, B. R. P., & Hills, P. J. (2012). Face distortion aftereffects in personally familiar, famous, and unfamiliar faces. *Frontiers in Psychology*, *3*(258), 1-8. https://doi.org/10.3389/fpsyg.2012.00258

Warnell, K. R., Sadikova, E., & Redcay, E. (2018). Let's chat: developmental neural bases of social motivation during real-time peer interaction. *Developmental Science*, *21*(3), 1-14. https://doi.org/10.1111/desc.12581

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

Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58(3), 475-482. https://doi.org/10.1007/BF03395630

Wilson, J. P., See, P. E., Bernstein, M. J., Hugenberg, K., & Chartier, C. (2014). Differences in anticipated interaction drive own group biases in face memory. *PloS One*, *9*(3), 1-6. https://doi.org/10.1371/journal.pone.0090668

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America, 123*(6), 4524-4538. https://doi.org/10.1121/1.2913046

Wittfoth-Schardt, D., Gründing, J., Wittfoth, M., Lanfermann, H., Heinrichs, M., Domes, G., ... & Waller, C. (2012). Oxytocin modulates neural reactivity to children's faces as a function of social salience. *Neuropsychopharmacology*, *37*(8), 1799-1807. https://doi.org/10.1038/npp.2012.47

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072. https://doi.org/10.3758/s13414-017-1361-2

Wrase, J., Schlagenhauf, F., Kienast, T., Wüstenberg, T., Bermpohl, F., Kahnt, T., ... & Heinz, A. (2007). Dysfunction of reward processing correlates with alcohol craving in detoxified alcoholics. *Neuroimage*, *35*(2), 787-794. https://doi.org/10.1016/j.neuroimage.2006.11.043

Xu, X., Aron, A., Brown, L., Cao, G., Feng, T., & Weng, X. (2011). Reward and motivation systems: A brain mapping study of early-stage intense romantic love in Chinese participants. *Human Brain Mapping*, *32*(2), 249-257. https://doi.org/10.1002/hbm.21017

Xu, X., Brown, L., Aron, A., Cao, G., Feng, T., Acevedo, B., & Weng, X. (2012). Regional brain activity during early-stage intense romantic love predicted relationship outcomes after 40 months: An fMRI assessment. *Neuroscience Letters*, *526*(1), 33-38. https://doi.org/10.1016/j.neulet.2012.08.004

Yager, L. M., Garcia, A. F., Wunsch, A. M., & Ferguson, S. M. (2015). The ins and outs of the striatum: role in drug addiction. *Neuroscience*, *301*, 529-541. https://doi.org/10.1016/j.neuroscience.2015.06.033

Yan, X., Young, A. W., & Andrews, T. J. (2017). The automaticity of face perception is influenced by familiarity. *Attention, Perception, & Psychophysics*, *79*(7), 2202-2211. https://doi.org/10.3758/s13414-017-1362-1

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(3), 283-299. https://doi.org/10.1002/acp.702

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

Zarate, J., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 1-9. https://doi.org/10.1038/srep11475

Zeki, S., & Romaya, J. P. (2010). The brain reaction to viewing faces of opposite-and same-sex romantic partners. *PloS One*, *5*(12), 1-8. https://doi.org/10.1371/journal.pone.0015802

Zhang, L., Li, Y., Zhou, H., Zhang, Y., Shu, H. (2021). Language-familiarity effect on voice recognition by blind listeners. *The Journal of the Acoustical Society of America, 1*(5), 1-6. https://doi.org/10.1121/10.0004848

Zhou, X., & Mondloch, C. J. (2016). Recognizing "Bella Swan" and "Hermione Granger": No own-race advantage in recognizing photos of famous faces. *Perception*, *45*(12), 1426-1429. https://doi.org/10.1177/0301006616662046

8 Appendix A

Above chance performance for Chapter 2

Comparisons with chance performance for Experiments 1-3

An exploratory analysis examined whether couples and control participants' performance was above chance for all experiments and conditions. Due to the use of unbiased hit rates for Experiments 1 and 2, chance performance was no longer 1/3. Instead, chance was calculated by estimating the joint probability that a stimulus and response of the corresponding category (i.e. a 'hit') would occur by chance for each participant and condition (Wagner, 1993). Therefore, chance rates were estimated per condition and per participant. Paired samples t-tests were run to compare participants' H_u scores to chance to determine whether, as a group, Couples and Controls displayed above chance performance for the two vocal identity tasks (Experiments 1 & 2). Shapiro-Wilk's tests, as well as boxplots, were used to verify that the data met the assumption of normality. Where this assumption was violated, Wilcoxon signed-rank tests were run instead.

For the speech intelligibility Experiment (Experiment 3), one sample t-tests were run per condition and group (Couples & Controls) to determine whether performance was above chance (1/32; 4 colours x 8 number options). Results per Experiment, per group, and per condition are detailed below.

Experiment 1: Voice identity recognition from non-verbal vocalisations

Couples

Comparing participants' performance with chance for each of the three familiarity conditions (personally familiar, lab-trained Anna/Adam, unfamiliar) showed that participants displayed above-chance accuracy for all conditions (ps < .0001; adjusted alpha = .0167) in the fillers task.

Controls

Control participants' performance was also significantly above chance in the fillers task for the two lab-trained voices (ps < .0001), and the unfamiliar voice (p < .001; adjusted alpha = .0167).

Experiment 2: Voice identity recognition in the context of acoustic modulation

Couples

For the modulation task, paired samples t-tests (or Wilcoxon signed rank tests) were run at each modulation step for each familiarity condition, comparing H_u scores to those which would be expected by chance. Three paired t-tests were run per modulation step; thus, the adjusted alpha was set to be p < .0167. The results showed that participants were significantly above chance for all familiarity conditions and modulation steps (all ps <. 0167; see Table 10).

Table 10. Shows significance at each modulation step for each familiarity condition.

	MODULATION STEP				
FAMILIARITY	-2	-1	0	+1	+2
Personally familiar	***	***	***	***	***
Lab-trained Anna/Adam	***	***	***	***	*
Unfamiliar	**	***	***	**	**

*** p < .0001, ** p < .001, * p < .01.

Controls

For the control group, performance was, for the most part, significantly above chance. However, there were a few conditions where participants in this group showed performance that was on average at chance levels (see Table 11) - this was observed predominantly for the unfamiliar voice condition.

Table 11. Displays significance of paired t-tests comparing performance to chance at each modulation step and familiarity condition.

	MODULATION STEP				
FAMILIARITY	-2	-1	0	+1	+2
Lab-trained Beth/Ben	*	***	***	***	***
Lab-trained Anna/Adam	**	*	***	*	n.s
Unfamiliar	*	n.s	n.s	n.s	n.s

*** p < .0001, ** p < .001, * p < .01, n.s: not significant.

Experiment 3: Speech perception from personally-familiar voices

Couples

A Shapiro-Wilk's test was run and data violated the assumption of normality ($p < .0001$), therefore a one-sample Wilcoxon test was run. Results of this test showed that participants performed significantly above chance level (3.13%) in the speech intelligibility task for both personally familiar voice trials ($V = 351$, $p < .0001$), and unfamiliar voice trials ($V = 351$, $p < .0001$). Note that due to ties, p values are not exact.

Controls

A Shapiro-Wilk's test was run and showed that the data violated the assumption of normality ($p < .0001$). Therefore, a one-sample Wilcoxon test was run. Results of this test showed that control participants also performed significantly above chance (3.13%) in the speech perception task for both lab-trained voice

trials (V = 350, p < .0001), and unfamiliar voice trials (V = 350, p < .0001). Again, p values are not exact due to ties.

9 Appendix B

Experiment 6 Quiz Questions

Beyoncé quiz questions:

1) What is Beyonce's middle name?
    a. Solange
    b. Desiree
    c. Celestine
    d. Giselle

2) Beyoncé rose to fame in the late 1990s as the lead singer of what R&B girl-group?
    a. En Vogue
    b. Destiny's Child
    c. The Spice Girls
    d. TLC

3) Beyoncé made her big screen debut in what movie?
    a. The Pink Panther
    b. Zoolander
    c. Dreamgirls
    d. Austin Powers in Goldmember

4) What was the name of Beyoncé's solo debut album?
    a. B'day
    b. Dangerously in Love
    c. I am … Sasha Fierce
    d. Survivor

5) In what song does Beyoncé sing: "I swore I'd never fall again, but this don't even feel like falling"?
    a. "Broken-hearted girl"
    b. "Halo"
    c. "Crazy in Love"
    d. "All night"

6) What is the name of Beyoncé's all-female tour band?
    a. Suga mama
    b. Parliament funkadelic
    c. Heartbreakers
    d. Sweet things

7) Who did Beyoncé marry in 2008?
    a. Jay-Z
    b. Sean Combs
    c. Ice Cube
    d. Kanye West

8) Destiny's Child released their major label debut song on the soundtrack of what film?
    a. I, Robot
    b. Men in Black
    c. Enemy of the State

d. Bad Boys

9) Who did Beyoncé portray in the movie *Cadillac Records*?
    a. Billie Holiday
    b. Aretha Franklin
    c. Etta James
    d. Ella Fitzgerald
10) What perfume did Beyoncé develop with Tommy Hilfiger?
    a. Enchanted
    b. Dreamgirl
    c. True star
    d. Truth or Dare

Taylor Swift quiz questions:

1) What was the title of Taylor's first album?
    a. Taylor Swift
    b. Our Song
    c. Fearless
    d. Speak Now

2) What is Taylor Swift's middle name?
    a. Elizabeth
    b. Renee
    c. Alison
    d. Sue
3) What famous rapper interrupted Taylor Swift's speech at the 2009 VMAs?
    a. Jay-Z
    b. Kanye West
    c. Snoop Dogg
    d. Eminem

4) Who did Taylor write "We Are Never Ever Getting Back Together" about?
    a. Tom Hiddleston
    b. John Mayer
    c. Joe Jonas
    d. Jake Gyllenhaal

5) In what song does Taylor sing: "Cause the players gonna play, play, play, play, play"?
    a. "You belong with me"
    b. "Shake it off"
    c. "I knew you were trouble"
    d. "we are never ever getting back together"

6) Where did Taylor spend her early years?
    a. A missile silo
    b. A Christmas tree farm
    c. The biosphere
    d. An African safari

7) Taylor Swift met actor Taylor Lautner on the set of what movie?
    a. Twilight
    b. The Giver
    c. Valentine's Day
    d. The Adventures of Sharkboy and Lavagirl

8) Which of Taylor's songs earned her a Guinness World Record for fastest-selling digital single?
    a. "You belong with me"
    b. "We are never ever getting back together"
    c. "Shake it off"
    d. "Mine"

9) Taylor was the spokesperson for which NHL team?
    a. Flames
    b. Predators
    c. Kings
    d. Flyers
10) What was the lead single on Taylor Swift's debut album?
    a. "Our song"
    b. "Picture to burn"
    c. "Tim McGraw"
    d. "Teardrops on my Guitar"

Justin Bieber quiz questions:

1) Where did Justin Bieber's talent manager discover him?
    a. American Idol
    b. The Voice
    c. Star Search
    d. YouTube

2) What was the name of Justin's debut album?
    a. My House
    b. My World
    c. My Life
    d. My Girl

3) On which popular TV show did Justin guest star in 2010?
    a. The Big Bang Theory
    b. CSI
    c. Pretty Little Liars
    d. Glee

4) What was Justin's high school GPA?
    a. 1.97
    b. 4.0

c. 2.35
d. 3.84

5) What is Justin's favourite food?
   a. Peanut Butter & Jelly
   b. Spaghetti
   c. Swedish fish
   d. Pizza

6) What song earned Justin his first Grammy Award?
   a. "Baby"
   b. "Purpose"
   c. "Where are Ü now"
   d. "Love Yourself"

7) Although Usher ultimately won out, what other singer wanted to mentor Justin?
   a. Michael Jackson
   b. Eminem
   c. Jay-Z
   d. Justin Timberlake

8) How many songs from Justin's debut album made the Billboard Hot 100?
   a. 3
   b. 5
   c. 7
   d. 1

9) What was the name of Justin's second studio album?
   a. Believe
   b. Purpose
   c. Urban Behavior
   d. Under the mistletoe

10) Justin serves as a celebrity spokesperson for what charity?
   a. Parliament of Promise
   b. Packs of Promise
   c. Pencils of Promise
   d. Projects of Promise

Harry Styles quiz questions:

1) What is Harry Styles' middle name?
   a. William
   b. Edward
   c. Thomas
   d. Arthur

2) When is Harry's birthday?
    a. March 23, 1994
    b. February 1, 1994
    c. February 28, 1993
    d. March 12, 1993

3) What is the name of Harry's older sister?
    a. Gemma
    b. Poppy
    c. Claire
    d. Imogen

4) Harry made his film debut in which film?
    a. 1917
    b. Call Me by Your Name
    c. Dunkirk
    d. The Shape of Water

5) Which of these fruits is NOT featured in a Harry Styles song?
    a. Cherry
    b. Apple
    c. Kiwi
    d. Watermelon
6) What was the name of Harry's debut solo tour?
    a. Harry Styles – Live on Tour
    b. Self-titled
    c. Treat People with Kindness
    d. Sign of the Tour

7) Which fictional island does the "Adore You" music video take place on?
    a. Narnia
    b. Eroda
    c. Avalon
    d. Nedlog

8) In which song does Harry sing ""And I'm well aware I write too many songs about you"?
    a. "Fine Line"
    b. "Cherry"
    c. "She"
    d. "Falling"

9) Which of the following songs was featured on Harry's FIRST album?
    a. "Little White Lies"
    b. "Ever Since New York"
    c. "She"
    d. "Seeing Blind"

10) What is the name of the bakery where Harry used to work as a teenager?

a. W. Mandeville
b. B. Warburton
c. T. Maudsley
d. J. Huntley

10 Appendix C

Experiment 7 – Analysis of behavioural data with 9 participants excluded

Nine participants failed to respond quickly enough to the target on over a third of trials in the behavioural SID practice in Experiment 7. Removing these participants from the analysis, a linear mixed effects model was run, identical to the analysis described in Section 5.1.2.8. Comparing a full model containing the fixed (voice identity condition) and random effects (participant) to a reduced model that did not contain the fixed effect showed that there was a significant effect of voice condition on participant reaction times ($\chi^2(2) = 13.2$, p =.001). Post-hoc pairwise comparisons using the *emmeans* package (FDR-corrected for multiple comparisons) showed that participants were significantly faster to respond to their musical idol (raw mean = 276.9ms) compared to the famous neutral (raw mean = 294.0ms; E = -15.2, p < .01), and to the unfamiliar voice (raw mean =297.8.1; E = -17.5, p < .01). No significant differences in reaction time were observed between the famous neutral condition (raw mean = 294.0ms) and unfamiliar voice condition (raw mean = 297.8ms; E = 2.34, p = .660).

11 Appendix D

Experiment 7 parameter estimates pairwise comparisons

Table 12. Displays output from one-way within-subjects ANOVAs for parameter estimates in selected significant clusters for the main effect of identity and the results of 2x3 within-subjects ANOVAs for the interaction between identity and outcome.

	df	Mean sum of squares	F-statistic	η2	p value
Main effect of identity					
L aIns	1.36, 32.70	5.14	27.95	.54	<.0001
R aIns	1.29, 30.94	2.84	43.33	.64	<.0001
L IFG	1.38, 33.01	3.24	45.00	.65	<.0001
R IFG	1.30, 31.16	19.80	34.66	.59	<.0001
L ACC	1.54, 37.00	2.63	21.21	.47	<.0001
Identity x Outcome interaction					
R STG	1.25, 29.95	98.66	73.71	.75	<.0001
L Precuneus	3.82, 91.74	5.07	8.17	.25	<.0001

Table 13. Displays post-hoc pairwise comparisons comparing parameter estimates in selected significant clusters (FDR-corrected for multiple comparisons), for the main effect of identity and the interaction between identity and outcome.

	Estimate	SE	df	t ratio	p (FDR-corrected)
Main eff: Identity					
L aIns					
Idol – Famous	3.20	0.53	48	6.05	<.0001
Idol – Unfamiliar	3.61	0.53	48	6.83	<.0001
Famous – Unfamiliar	0.42	0.53	48	0.79	.436
R aIns					
Idol – Famous	2.91	0.38	48	7.61	<.0001
Idol – Unfamiliar	3.23	0.38	48	8.45	<.0001
Famous – Unfamiliar	0.32	0.38	48	0.84	.406
L IFG					
Idol – Famous	3.44	0.42	48	8.15	<.0001

Idol – Unfamiliar	3.50	0.42	48	8.28	<.0001
Famous - Unfamiliar	0.06	0.42	48	0.13	.896
R IFG					
Idol – Famous	6.94	1.01	48	6.85	<.0001
Idol – Unfamiliar	7.63	1.01	48	7.53	<.0001
Famous – Unfamiliar	0.69	1.01	48	0.68	.498
L ACC					
Idol – Famous	1.96	0.40	48	4.86	<.0001
Idol – Unfamiliar	2.49	0.40	48	6.19	<.0001
Famous – Unfamiliar	0.54	0.40	48	1.33	.189
Interaction: Identity x Outcome					
R STG					
Idol hit – Idol miss	16.99	1.4	120	12.11	<.0001
Idol hit – Famous hit	3.16	1.4	120	2.25	.039
Idol hit – Famous miss	16.59	1.4	120	11.82	<.0001
Idol hit – Unfamiliar hit	2.52	1.4	120	1.79	.103
Idol hit – Unfamiliar miss	18.30	1.4	120	13.04	<.0001
Idol miss – Famous hit	-13.84	1.4	120	-9.86	<.0001
Idol miss – Famous miss	-0.41	1.4	120	-0.29	.769
Idol miss – Unfamiliar hit	-14.48	1.4	120	-10.32	<.0001
Idol miss – Unfamiliar miss	1.30	1.4	120	0.93	.411
Famous hit – Famous miss	13.42	1.4	120	9.56	<.0001
Famous hit – Unfamiliar hit	-0.65	1.4	120	-0.46	.693
Famous hit – Unfamiliar miss	15.14	1.4	120	10.79	<.0001
Famous miss – Unfamiliar hit	-14.07	1.4	120	-10.02	<.0001
Famous miss – Unfamiliar miss	1.71	1.4	120	1.22	.281

Unfamiliar hit – Unfamiliar miss	15.78	1.4	120	11.25	<.0001
L Precuneus					
Idol hit – Idol miss	3.21	0.56	120	5.77	<.0001
Idol hit – Famous hit	0.55	0.56	120	0.98	.410
Idol hit – Famous miss	0.83	0.56	120	1.48	.212
Idol hit – Unfamiliar hit	1.01	0.56	120	1.81	.136
Idol hit – Unfamiliar miss	1.69	0.56	120	3.04	.009
Idol miss – Famous hit	-2.67	0.56	120	-4.79	<.0001
Idol miss – Famous miss	-2.39	0.56	120	-4.29	.0002
Idol miss – Unfamiliar hit	-2.20	0.56	120	-3.96	.0005
Idol miss – Unfamiliar miss	-1.52	0.56	120	-2.74	.018
Famous hit – Famous miss	0.28	0.56	120	0.50	.663
Famous hit – Unfamiliar hit	0.46	0.56	120	0.83	.471
Famous hit – Unfamiliar miss	1.14	0.56	120	2.05	.091
Famous miss – Unfamiliar hit	0.18	0.56	120	0.33	.741
Famous miss – Unfamiliar miss	0.87	0.56	120	1.55	.205
Unfamiliar hit – Unfamiliar miss	0.68	0.56	120	1.22	.305