Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings.

Sarah V Stevenage

Department of Psychology, University of Southampton, UK

CORRESPONDENCE:

Professor Sarah Stevenage, Department of Psychology, University of Southampton, Highfield, Southampton, Hampshire, UK, SO17 1BJ

Email: [svs1@soton.ac.uk](mailto:svs1@soton.ac.uk); Tel: +44 2380 592973; Fax: +44 2380 594597

Abstract

Thirty years on from their initial observation that familiar voice recognition is not the same as unfamiliar voice discrimination (van Lancker & Kreiman, 1987), the current paper reviews available evidence in support of a distinction between familiar and unfamiliar voice processing. Here, an extensive review of the literature is provided, drawing on evidence from four domains of interest: the neuropsychological study of healthy individuals, neuropsychological investigation of brain-damaged individuals, the exploration of voice recognition deficits in less commonly studied clinical conditions, and finally empirical data from healthy individuals. All evidence is assessed in terms of its contribution to the question of interest – is familiar voice processing distinct from unfamiliar voice processing. In this regard, the evidence provides compelling support for van Lancker and Kreiman’s early observation. Two considerations result: First, the limits of research based on one or other type of voice stimulus are more clearly appreciated. Second, given the demonstration of a distinction between unfamiliar and familiar voice processing, a new wave of research is encouraged which examines the transition involved as a voice is learned.

Keywords:

Familiar voice recognition; Unfamiliar voice discrimination; Multimodal person perception; Voice learning.

Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings.

1. Introduction

Over the last few years, voice research has gained substantially in momentum. With initial explorations styled very much on the more dominant field of face recognition, there now exists a considerable body of knowledge. This allows reflection on the neuropsychology of voice processing, the deficits that may result through brain damage or disorder, and the performance of healthy individuals under a range of empirical conditions. Despite this impressive rate of discovery in the field of voice processing, very little voice research has explored the intuitive realisation that the processing of a familiar voice is not equivalent to the processing of an unfamiliar voice. The purpose of the present review is to evaluate the evidence in support of such a distinction.

This distinction was first highlighted by van Lancker and Kreiman (1987) who described the cases of three brain-damaged patients capable of discriminating between unfamiliar voices whilst being unable to recognise familiar ones. More interesting were the cases of seven patients who showed the reverse pattern of being capable of recognising familiar voices without being able to discriminate between unfamiliar ones. Since this time, however, only a few papers have discussed this distinction. The first is provided by Neuner and Schweinberger (2000) who examined the performance of 36 brain-damaged patients. Their results revealed four cases who showed a voice specific deficit. Importantly, however, three of these four patients were capable of discriminating between unfamiliar voices but were unable to recognise the voices of famous celebrities, confirming van Lancker and Kreiman’s (1987) initial distinction. A second paper is provided by Schelinski, Borowiak and von Kriegstein (2016) who showed intact ability in high functioning individuals with autism when recognising familiar voices and yet a particular neuropsychological dysfunction when processing unfamiliar voices. Finally, a third paper is provided by Blank, Wieland and von Kriegstein (2014) who reviewed neuropsychological data from both brain-damaged patients and healthy individuals indicating a clear distinction between familiar voice recognition and unfamiliar voice discrimination.

This distinction between the processing of familiar and unfamiliar voices has been extensively discussed by Kreiman and Sidtis (2011), and echoes a similar distinction in the field of face research (Megreya & Burton, 2006). Across a series of experiments, Megreya and Burton showed that quite different perceptual processes were utilised for the two sets of stimuli, with performance only correlated when familiar faces were turned upside down. The authors concluded by setting a challenge for face researchers not to conflate familiar and unfamiliar face stimuli into a single overarching framework.

With this in mind, and with the weight of a considerable evidence base to now draw on, the current paper revisits the possibility of a distinction between familiar and unfamiliar voice processing raised by van Lancker and Kreiman (1987) 30 years ago. However, this review expands the reach of previous work by drawing together the evidence from four discrete areas of enquiry. These span the neuropsychological study of healthy individuals, neuropsychological investigation of brain-damaged individuals, the exploration of voice recognition deficits in less commonly studied clinical conditions, and finally, the empirical data from healthy individuals. The goal is to provide a comprehensive consideration of the evidence regarding voice recognition in general, and regarding a distinction between familiar and unfamiliar voice processing in particular.

2. Neuropsychological data from healthy individuals

Within the first section of this review, the purpose is to examine evidence regarding the brain areas that may be specialised for voice processing. Furthermore, the question of relevance is whether there is a distinction at the neurological level between familiar and unfamiliar voice processing.

2.1 Localisation of voice processing in healthy individuals

In response to the first of these questions, a considerable literature now exists which indicates the importance of the temporal cortices for voice processing, with laterality effects favouring the right hemisphere. For example, Mathiak, Menning, Hertrich, Matiak, Zvyagintsev & Ackermann (2007) used fMRI to distinguish three distinct neural networks. This indicated that the brain areas involved in locating a speaker (left posterior STG), and in determining what they were saying (Broca’s areas), were quite distinct from those involved in identifying the speaker (temporal lobes). Similarly, Imaizumi, Mori, Kiritani, et al. (1997) provided PET scan data which also revealed the importance of bilateral temporal poles when attending to the speaker (rather than the speech content) in a voice clip.

Alongside these studies, Nakamura, Kawashima, Sugiura, Kato, Nakamura, Hatano, Nagumo, Kubota, Fukuda, Ito and Kojima (2001) examined the PET scans of nine healthy participants who were asked to listen to two brief speech clips and determine whether the speaker was familiar (friend or self) or unfamiliar. In comparison to a control task involving vowel discrimination, the PET scan data indicated involvement of the left frontal pole and the right temporal pole when processing the voice clips for identity. This separation of speech and voice processing areas has been confirmed in a more recent study using a powerful MEG approach (Schall, Kiebel, Maess & von Kriegstein, 2015). As above, Schall and colleagues identified activity in the right superior temporal sulcus (STS) of the temporal lobe at 200 ms post-onset during a voice recognition task. In contrast, the participants showed activity in the left STS during a speech-based task.

A number of studies also exist which have enabled greater precision in the identification of voice sensitive areas within the temporal lobes. For example, using functional magnetic resonance imaging (fMRI), Belin and colleagues defined a ‘temporal voice area’ (TVA) when recognising voices. This, they viewed as analogous in function to the fusiform face area when recognising faces (Belin, Zatorre, Lafaille, Ahad & Pike, 2000). This area extended along the superior bank of the right STS, and selectively responded when passively listening to vocal over non-vocal sounds. Increased activation was noted particularly in the posterior superior temporal gyrus (STG), and from posterior to anterior regions of the STS. Similarly, Belin, Zatorre and Ahad (2002) used fMRI to show greater activation of anterior portions of the right STS when passively listening to non-speech vocal sounds as opposed to scrambled counterparts.

The importance of the right STS in voice processing is confirmed in three further studies. First, Lattner, Meyer and Friederici, (2005) used fMRI data from 16 participants who completed a ‘naturalness decision’ in response to original, and pitch-modified, voices of unfamiliar speakers. The results confirmed greater activation of the right posterior STG when differentiating between the unfamiliar speakers. Similarly, Rämä and Courtney’s (2005) fMRI results showed heightened activation of the posterior STG and posterior to mid regions of the STS, again mainly in the right hemisphere, as 12 participants encoded and later recognised unfamiliar voices. Finally, the results of Warren, Scott, Price and Griffiths (2006) indicated greater activation of the posterior STS, mainly again in the right hemisphere, as 12 participants passively differentiated between unfamiliar voices. These results converge with the previous demonstration to indicate that the posterior portion of the STS/STG region is important, particularly within the right hemisphere, when processing speaker as opposed to speech information.

2.2 A distinction between sites for familiar and unfamiliar voice processing

Set against these studies are the findings of von Kriegstein, Eger, Kleinschmidt and Giraud (2003). They conducted an fMRI study with 14 participants who had been familiarised with 24 voices. During the fMRI study, participants were asked to perform two tasks to determine the pattern of activity associated with voice recognition and speech recognition respectively. At a behavioural level, the results indicated that participants were able to perform at above chance levels in the voice recognition task despite limited familiarity with the voices. Moreover, in common with the data from Schall et al. (2015), speech recognition was associated with a greater involvement of the left hemisphere (especially the left medial temporal gyrus), whilst speaker recognition was associated with a greater involvement of the right hemisphere. Importantly, however, when asked to recognise the familiarised speakers, the results indicated activation in *anterior* portions of the right STS. This contrasts with previously noted activation in posterior portions of the STS.

These data are important in the context of the current review as they serve to provide the first indication of a neurological separation of areas responsible for familiar and unfamiliar voice processing. More specifically, they suggest the involvement of posterior regions of the right STS when processing unfamiliar voices, and anterior regions of the right STS when processing familiar voices. These data echo the suggestion of Warren et al. (2006) of a posterior-to-anterior axis along which voice processing becomes refined to support the individuation of familiar speakers rather than the mere discrimination of unfamiliar speakers.

Further support for the separation of familiar and unfamiliar voice processing comes from a number of studies. For example, von Kriegstein and Giraud (2004) used an fMRI recognition study with familiar and unfamiliar voices. They revealed increased activity in anterior regions of the right STS in response to both familiar and unfamiliar voices. However, posterior regions of the right STS showed a stronger activation to unfamiliar voices. Similarly, von Kriegstein, Kleinschmidt, Sterzer & Giraud (2005) used fMRI to examine activation when processing personally familiar versus unfamiliar voices, and found differential activation in bilateral anterior regions of the STS/STG together with the temporal poles. When using fMR imagery to determine the Blood Oxygenation Level Dependant (BOLD) response, Bethmann and Brechmann (2014) also showed a greater response in the anterior middle temporal gyrus (MTG) when participants were presented with familiar celebrity voices as opposed to non-vocal sounds. Together, these data confirm von Kriegstein et al.’s (2003) indication of a posterior/anterior separation when processing unfamiliar and familiar voices respectively.

These data are further supported by three studies which take a different approach. Belin and Zatorre (2003) used an fMRI adaptation study to investigate brain activity to voices as they became familiar through passive exposure. In line with the results previously described, Belin and Zatorre observed repetition suppression effects in terms of a smaller activation in right anterior regions of the STS in response to an adapted (learned) voice. More specifically, activation decreased in the area of the right anterior STS as the voice of a particular speaker was repeated. Similarly, Latinus, Crabbe and Belin (2011) measured the BOLD response through an fMRI adaptation design, and showed activation in the superior temporal pole when processing acoustic aspects of unfamiliar voices, but showed activity in areas towards the anterior STS and along the convexity of the inferior frontal cortices when processing identity-related aspects of the adapted voices. Finally, Zäske, Volberg, Kovács & Schweinberger (2014) used ERP measurement to record the P3 waveform that emerged when participants were adapted to a voice (Schweinberger, Walther, Zäske & Kovács, 2011). They showed a larger P300 response to learned voices in parietal areas, which the authors attributed to episodic memory influences. However, they also found a suppression of beta-band oscillations (at 16-17Hz) between 290 and 370 ms post-onset in central and right temporal sites when novel clips of the learned voices were presented. Given that ERP data better indicate temporal rather than spatial patterns of activation, the identification of familiarity-driven differences in wave forms within right temporal areas provides cautious convergence with the fMRI data described above.

Together, this block of studies provides a rich and consistent picture (summarised in Table 1). Whilst the right temporal lobes, and the areas associated with the right STS/STG in particular, show increased activation when processing voices *per se*, there does appear to be support for the dissociation of familiar and unfamiliar voice processing. Specifically, posterior areas of the STS/STG appear to be more involved in unfamiliar voice discrimination, whilst anterior areas of the STS/STG are more involved in familiar voice recognition.

(Please insert Table 1 about here)

3. Consideration of voice-processing deficits

Given the identification of regions of the brain sensitive to voice processing, the next natural focus of interest relates to the existence of voice-specific processing deficits. The purpose of this section is to examine the evidence from brain-damaged individuals with a primary complaint involving voice recognition in order to determine whether a distinction between familiar and unfamiliar voice processing is indicated.

3.1: Definition and demonstration of phonagnosia:

The question of whether a voice processing deficit exists owes much to the demonstration of a similar deficit in the face processing domain. Indeed, when considering face processing, a right occipito-temporal area – the fusiform face area – has been identified, and damage to this area has been associated with face processing difficulties (see Benton & van Allen, 1968). The corresponding prediction, therefore, is that damage to the TVA may be associated with voice processing difficulties. Surprisingly, however, whilst the identification of a voice processing area has been the focus of a significant research initiative, the exploration of voice processing deficits has attracted less attention. This is potentially because such deficits are less commonly appreciated or reported by the patients themselves. Moreover, there has been less effort towards the generation of standardised diagnostic tests of voice processing capability. This said, examples of specific voice processing deficits do exist.

The first such cases were documented by van Lancker and Canter (1982) in their investigation of 30 adult brain-damaged patients (21 left-hemisphere patients, 9 right-hemisphere patients) with lesions acquired through cerebro-vascular accident or injury. Patients were given a famous faces recognition test and a famous voices recognition test in order to test the assumption that both types of deficit may be exhibited more often by right-hemisphere patients than by left-hemisphere patients. The results supported predictions in that voice and face deficits were indeed more likely for patients with right- rather than left-hemisphere damage. However, the results also revealed that these deficits could dissociate from one another. Of the nine right-hemisphere patients, four showed a deficit in voice processing, four showed a deficit in face processing and, of these, three patients were common to both groups. Consequently, there was one patient who exhibited a pure deficit in face processing, and there was one patient who exhibited a pure deficit in voice processing. In line with the term ‘prosopagnosia’ to describe a pure face deficit (Bodamer, 1947), van Lancker and Canter coined the term ‘phonagnosia’ to describe a pure voice deficit.

The second paper of relevance to this section is provided by Hailstone, Crutch, Vestergaard, Patterson & Warren (2010) who described a single case study of a patient (QR) who displayed what may be called ‘progressive phonagnosia’ due to lobar degeneration. The value of this case study comes through the reporting of brain lesion data, as well as behavioural data. On testing, QR showed markedly impaired voice recognition (with an inability to recognise the voices of her children on the telephone). This, however, sat in stark contrast to her spared recognition of faces resulting in her identification as a pure phonagnosic. Of more interest, however, brain lesions were located in bilateral fronto-temporal areas, with particular atrophy in the right anterior temporal lobe, extending backwards and encompassing the STS. These lesion locations overlapped with the voice-processing locations identified through the study of healthy individuals in Section 2, providing a ready explanation for QR’s deficit with familiar voices.

Together, these explorations confirm the existence of a selective voice-processing deficit, following brain damage largely to right hemisphere areas. This was localised at least in the case of QR to the right superior temporal sulcus. Given this, the next question is whether any cases exist to support a clinical dissociation between the processing involved in familiar voice recognition and that involved in unfamiliar voice discrimination.

3.2 A dissociation between familiar and unfamiliar voice processing in phonagnosia

Given the identification in Section 2 of discrete brain regions associated with familiar, and with unfamiliar, voice processing, lesions to these discrete areas may support the prediction of a clinical dissociation of deficits. The work of van Lancker and colleagues has been instrumental in the investigation of this dissociation, and a number of studies are relevant.

The first is provided by van Lancker and Kreiman (1987) and indeed represents the impetus for the current review. Together, the authors reviewed the voice processing capabilities of 45 patients (21 right-hemisphere patients, 18 left-hemisphere patients, 6 bilateral patients) by testing famous voice recognition to 25 celebrity voices and unfamiliar voice discrimination to 10 unfamiliar voices (a sub-set of patients only completed the recognition test). Performance was compared to that of 48 healthy controls and revealed several findings of note. First, performance in the healthy controls showed that both famous voice recognition and unfamiliar voice discrimination were above chance but showed only a moderate correlation, suggesting the possibility of a different basis for familiar and unfamiliar voice processing. Second, within the patient sample, famous voice recognition deficits were associated with right hemisphere lesions, whereas unfamiliar voice discrimination deficits were associated with lesions in either hemisphere supporting the neurological data described previously. Third, deficits in familiar voice recognition could occur independently of deficits in unfamiliar voice discrimination, with three patients showing normal discrimination without the ability to recognise familiar voices, and seven patients showing normal recognition without the ability to discriminate between unfamiliar voices.

The latter group of seven patients are of particular interest given a naïve assumption that low-level perceptual discrimination of unfamiliar voices may underpin higher-level processing leading to the identification of familiar voices. These seven patients clearly violate such an assumption (see Warren et al., 2006 for a discussion of this hierarchical organisation of processing). This aside, the data of van Lancker and Kreiman (1987) offer a powerful demonstration of the dissociation of familiar voice recognition from unfamiliar voice discrimination.

The results of van Lancker and Kreiman (1987) were supported by three further explorations of patient groups. Van Lancker, Cummings, Kreiman and Dobkin (1988) examined the performance of six patients, providing detailed analysis of familiar and unfamiliar voice processing alongside neurological indication of the site of brain damage. Five of the six patients demonstrated a dissociation between familiar and unfamiliar voice capabilities, whilst the sixth patient showed an impairment to both capabilities. The authors described damage to temporal lobes of either hemisphere when unfamiliar voice discrimination was impaired. However, familiar voice recognition was associated more discretely with damage to right hemisphere areas, later refined to include the right inferior parietal or right temporo-parietal cortex (Gainotti, 2011), close to the right STS implicated in previous research.

In a similar vein, van Lancker, Kreiman and Cummings (1989) described the performance of left- and right-hemisphere patients relative to healthy controls on the same tests of familiar voice recognition and unfamiliar voice discrimination used by van Lancker and Kreiman (1987) previously. The data again indicated that familiar and unfamiliar voice processing capabilities dissociated from one another, and again the brain imaging data indicated that unfamiliar voice processing deficits were associated with damage to the temporal lobes of either hemisphere, whilst familiar voice processing deficits were associated with damage to the right hemisphere parietal regions.

Finally in this regard, the results of Neuner and Schweinberger (2000) are of relevance. They compared the performance capabilities of 36 brain-damaged patients drawn from a rehabilitation ward, to that of 20 healthy controls. The results suggested that face, name and voice recognition deficits all dissociated from one another in that they could be exhibited as pure deficits. More importantly, whilst four of the 36 patients displayed a voice-specific deficit, three of these patients (patients #5, #24 and #30), were impaired only on the recognition of famous celebrity voices. Indeed, their discrimination of unfamiliar voices in a same/different matching task was normal. (The fourth patient in this group (patient #1) was impaired in both capabilities.)

Together, the evidence from van Lancker and colleagues, and from Neuner and Schweinberger (2000), is compelling in demonstrating pure cases of phonagnosia in which a deficit in familiar voice recognition can occur independently of a deficit in unfamiliar voice discrimination. Alongside these group studies, three single case studies support this point. Peretz and colleagues (Peretz, Kolinsky, Tramo, Labrecque, Hublet, Demeurisse & Belleville, 1994) described two amusic patients (GL and CN) whose principal complaint, following bilateral lesions, was the loss of their former ability to recognise melodies. In addition, however, both patients showed measurable impairments in voice recognition skills. Interestingly, capabilities with familiar and unfamiliar voice processing dissociated in these two patients: GL was unable to recognise previously familiar voices, but remained able to discriminate between unfamiliar ones, whilst in contrast, CN was unable to process either set of voices. These voice processing deficits occurred in the context of intact face processing skills, defining these two cases as pure phonagnosic cases, who were differentiated in their impairments according to the familiarity of the voice.

In line with Peretz et al.’s (1994) patients, Garrido, Eisner, McGettigan, Stewart, Sauter, Hanley, Schweinberger, Warren and Duchaine (2009) described the first case of developmental or ‘congenital’ phonagnosia (KH) in which a voice processing deficit was evident in the absence of any obvious neurological damage, and was presumed to have been present from childhood. On testing, KH showed a marked inability to recognise the voices of famous celebrities, and an inability to learn and later recognise the voices of 6 new speakers as measured both on a naming test, and on an old/new test of recognition memory. In contrast, she was unimpaired in her ability to discriminate between the voices of unfamiliar speakers in three same/different tests, and was unimpaired in all tests of face recognition.

The final single case study to be considered in this section is patient AN, described by Herald, Xu, Biederman, Amir and Shilowich (2014) and again by Xu, Biederman, Shilowich, Herald, Amir and Allen (2015). Patient AN shared a diagnosis of congenital phonagnosia with patient KH discussed previously. On testing, AN showed a marked inability to pair celebrity voices with their faces or names. She also showed a deficit when trying to imagine the voices of highly familiar people, despite an intact ability to imagine non-voice sounds. These deficits with famous and familiar voices sat alongside an intact ability to discriminate between unfamiliar voices when tested using a two-alternative forced choice delayed matching task. Indeed, in this task, AN performed at a level equivalent to a set of control participants. As such, AN’s deficits were confined to familiar voice recognition, with intact unfamiliar voice discrimination, intact face recognition, and normal sensitivity to speech and non-speech sounds (Xu et al., 2015).

The patients in this section have been described at some length given the precision of the question at hand. All demonstrate that voice recognition deficits do exist, and can occur in the absence of a deficit in any other modality. Together, they may be described by the term ‘phonagnosia’ which refers to an inability to process voices in the absence of an inability when processing faces or names. However, a closer consideration of several of the patients has enabled reflection on whether impairments in famous voice recognition and unfamiliar voice discrimination may dissociate. Such a dissociation is predicted on the basis of distinct neurological areas in which familiar and unfamiliar voices are processed. Importantly, the prediction is supported through demonstration of patients who show an inability with familiar voice recognition in the absence of any deficit in unfamiliar voice discrimination, or vice versa (as summarised in Table 2). Consequently, at the level of patient data from those with a primary deficit in voice recognition, a distinction between familiar voice recognition and unfamiliar voice discrimination appears to be warranted.

(Please insert Table 2 about here)

4. Voice deficits in other clinical populations

Whilst the previous section detailed cases of patients whose primary complaint related to voice recognition, there are a number of other clinical conditions within which voice processing difficulties arise. These include schizophrenia, autism and dyslexia. The literature concerning voice processing deficits in these populations is rather sparse, nevertheless, the purpose of this section is to explore whether these additional cases may provide convergent evidence to support a distinction between familiar and unfamiliar voice processing.

4.1 Schizophrenia

A number of researchers have examined the experience of individuals with schizophrenia when recognising voices. Much of this work has revealed an impairment when judging vocal affect (Alba-Ferrera, Fernyhough, Weis, Mitchell & Hausmann, 2012a). However, work is now increasingly focussing on the additional impairment when processing vocal identity. This interest stems largely from the fact that up to 70% of individuals with schizophrenia experience auditory hallucinations (Bentall, 1990). By way of explanation, Johns, Rossell, Frith, Ahmad, Hemsley, Kuipers & McGuire (2001) suggested that such patients may exhibit a disorder of voice source identification resulting in ‘inner speech misidentification’. More recently, however, Alba-Ferrara, Weis, Damjanovic, Rowett & Hausmann (2012b) noted that auditory hallucinators often experience voices which are attributed to non-self identities, and have acoustic qualities quite unlike the patient themselves, both of which cast doubt on hallucinations being the result of a misidentification of inner speech. Nevertheless, only a limited evidence base exists to provide any systematic exploration of the perception of both real and hallucinated voices in this population.

In this regard, Chhabra, Badcock, Maybery & Leung (2012) examined the capacity of 65 individuals with schizophrenia (33 with auditory hallucinations, 32 without auditory hallucinations) alongside 32 healthy controls. The question of interest was whether individuals with schizophrenia would show basic sensitivity to vocal characteristics. Following similarity ratings between pairs of unfamiliar voices, multidimensional scaling was used to identify the dominant characteristics that individuals relied upon. In common with healthy controls, the results indicated that individuals with schizophrenia used two vocal dimensions when processing vocal identity: fundamental frequency (pitch) and formant dispersion (resonance cues). This said, the schizophrenic groups with and without auditory hallucination both made substantially less use of formant dispersion than healthy controls when discriminating between the voices. These results suggested that individuals with schizophrenia showed normal sensitivity to vocal characteristics, but they may use these characteristics differently relative to healthy controls, yielding difficulties in processing real yet unfamiliar voices.

A similar conclusion was reached by Badcock and Chhabra (2013). They examined the performance of individuals with schizophrenia, with or without auditory hallucinations, alongside a group of non-clinical auditory hallucinators (who have no diagnosis of mental illness, and attribute voices to a positive, comforting internal presence). All participants completed a sequential matching task in which two unfamiliar voices were presented, followed by a brief gap and then a probe voice. The task was to determine whether the probe voice matched with either of the study items in terms of both speaker and word. The results suggested that individuals with schizophrenia were poor at the task, regardless of their experience of auditory hallucinations. Importantly, their performance was worse when compared to both the non-clinical hallucinators, and to healthy controls.

Both studies point to an impairment amongst individuals with schizophrenia when processing vocal identity cues in unfamiliar speakers. However, neither study examined performance with familiar speakers. In this regard, Alba-Ferrara et al. (2012b) suggested that individuals with schizophrenia may also show a deficit when processing familiar voices. Using a signal-detection approach, 25 individuals with schizophrenia were asked to make a familiarity judgement to 48 famous and 48 non-famous voices. Care had been taken to ensure that the content of the clips could not reveal identity. Despite this, the patient group were markedly weaker at the task than controls in terms of hit rate and sensitivity of discrimination (A’), but this reached significance only in the patients *with* auditory hallucinations compared to healthy controls.

Thus far, the results suggest that individuals with schizophrenia exhibit measurable deficits when discriminating between unfamiliar voices. However, those who experience auditory hallucinations may experience additional difficulties when recognising familiar (famous) voices. The final paper of relevance in this section is provided by Zhang, Hao, Shi, Mou, Yao and Chen (2008) who asked 26 patients and 13 healthy controls to make familiarity judgements to personally familiar or unfamiliar voices as part of a neuroimaging study. The results supported previous empirical conclusions: individuals with schizophrenia *with* auditory hallucinations were impaired when recognising familiar voices compared to the healthy controls. Moreover, the imaging data located this deficit to lower activation in the right STG in an area corresponding to Brodman Area 22. This result corresponds well with the neuroimaging data obtained from phonagnosic patients, and with healthy individuals. Together, the examination of voice identity processing in schizophrenic patients is consistent in suggesting voice specific difficulties, with the extent of these difficulties dependant on whether the patient experienced auditory hallucination. As such unfamiliar voice discrimination can be dissociated from familiar voice recognition within the schizophrenic population. Additional work would be valuable with this patient group to confirm these initial findings.

4.2 Autism

In reviewing work with the second clinical group of interest, it is clear that the study of voice processing deficits in individuals with autism has perhaps taken a back-seat compared to the study of their face processing deficits. Nevertheless, a literature is emerging which indicates a measurable impairment in tasks involving voice processing. As with the study of schizophrenic patients above, this literature is sparse. However, it provides both behavioural and neuropsychological evidence of relevance to the current purpose.

At a behavioural level, one of the first studies of voice processing in individuals with autism was provided by Boucher, Lewis and Collis (1998). They assessed the abilities of 19 high functioning children with autism against a control group of 20 children matched on age and language ability. The children completed three tasks – a face-voice matching task, a face recognition task, and a voice recognition task - using the faces and voices of staff from their school to provide a highly familiar stimulus set. The results suggested that all participants were able to perform at above-chance levels. However, performance was significantly poorer in the autistic group than in the control group on all three tasks (see also Boucher, Lewis & Collis, 2000). Somewhat surprisingly, performance on the face recognition task did not correlate with that on the voice recognition task for the few participants who completed both tasks, suggesting deficits that may be peculiar to each stimulus type rather than being the result of common factors. This aside, the data confirmed a familiar voice recognition deficit as well as a familiar face recognition deficit in this autistic group.

When considering the processing of unfamiliar voices, the work of Schelinski and colleagues (Schelinksi et al., 2016) is relevant. They combined behavioural with neuroimaging techniques and showed normal responsiveness in an adult autistic population to vocal and non-vocal sounds in the voice sensitive areas of the temporal lobes. However, the autistic group manifest dysfunctional responses during an unfamiliar voice recognition task. Specifically, they showed less activation of the right posterior STS/STG compared to healthy controls. Curiously, both the behavioural and the neuropsychological data in Schelinski et al.’s (2016) study suggested that the voice processing deficit in adults with autism was confined to the discrimination of unfamiliar voices only.

In order to compare performance with both familiar and unfamiliar voices, Schelinski, Roswandowitz and von Kriegstein (2017) again examined the performance of adults with autism. Compared to age-matched controls, those with autism showed a particularly pronounced difficulty when processing previously unfamiliar voices across three learning tasks. In contrast, the deficits shown on a familiar voice recognition task to celebrity stimuli were measurable (see Schelinski et al., 2017, Figure 2 on p. 160) but did not reach significance, potentially through greater variance in task performance when responding to famous voices. As such, these data supported the earlier conclusions of Schelinski et al. (2016). Taken together, they suggested some support for a dissociation in the processing of familiar and unfamiliar voices within an adult autistic population, with a greater deficit when processing unfamiliar than familiar voices. Such a conclusion is at odds with that of Boucher et al. (1998) when studying children with autism. Nevertheless Schelinski et al.’s (2016, 2017) work suggests that performance capabilities with familiar and unfamiliar voices may become impaired to differing degrees within a single population of autistic individuals.

Finally, a very recent paper to examine voice processing and autism is provided by Skuk, Palermo, Broemer and Schweinberger (2017). Taking a slightly different approach, they examined the association between voice processing difficulties and autistic traits in a general population. In a comprehensive set of experiments, adolescent-aged participants completed a familiarity decision task (familiar or not), and a voice identification task, to personally familiar voices drawn from the participants’ year-group at school. Participants also completed a famous face recognition task, and the autistic spectrum quotient questionnaire (AQ). The results suggested quite marked individual differences in performance levels. However, they also indicated a significant correlation between the existence of autistic traits relating to social factors, and poor performance on the voice identification task using personally familiar voices. Importantly, this emerged regardless of the familiarity of the voice. In contrast, autistic social traits were positively associated with performance in the famous face recognition task, for female listeners. As such, the data suggested a deficit in the processing of both familiar and unfamiliar voices amongst those with autistic traits.

Given that voice processing is relatively under-researched in this population, the early findings are clearly at odds with one another regarding whether deficits are confined to familiar or unfamiliar speakers. One important consideration may lie in the selection of the stimuli to form the familiar set, and studies reviewed here have varied in their use of personally familiar stimuli (Boucher et al., 1998; Skuk et al., 2017) and famous (celebrity) stimuli (Schelinski et al., 2016, 2017). What is clear, however, is that unfamiliar voice discrimination appears to dissociate from famous voice recognition when examining the performance of a single group of individuals within the same clinical category. As such, the available evidence tentatively supports a distinction between unfamiliar voice discrimination and famous voice recognition, but further work with this clinical group would be valuable.

4.3 Dyslexia

The final clinical condition to be considered in this section is dyslexia - a developmental reading disorder commonly believed to stem from impaired phonological processing (see Peterson & Pennington, 2012). Exploration of voice processing abilities has enabled researchers to determine whether this phonological processing deficit extends to the processing of the speaker as well as the speech.

Three studies are notable in this regard. First, Brachacki, Fawcett and Nicholson (1994) compared the performance of seven dyslexic adults to that of eight healthy controls. All participants were initially presented with static photographs and the corresponding voices of eight unfamiliar targets. The target voices were then presented with an equal number of distractor voices in an old/new voice recognition task conducted immediately following study, and one week later. Similarly, the target faces were presented with an equal number of distractor faces to provide an old/new face recognition task.

The results indicated that face recognition was near-perfect for both participant groups at both testing occasions. In contrast, individuals with dyslexia performed significantly worse than the healthy control group in the voice recognition task on both occasions. These results were important in establishing the existence of a voice processing deficit within dyslexia when processing unfamiliar voices.

These results have since been replicated by Perrachione, Del Tufo and Gabrieli (2011), Their dyslexic participants showed a difficulty when differentiating unfamiliar speakers regardless of their language of speech. By comparison, the control group showed a performance advantage when speakers spoke in the listeners’ native tongue. This fuelled an explanation that dyslexia may be accounted for by a specific deficit in native-language phonological processing, but that a voice processing impairment may accompany this primary deficit. Recent evidence from Perea and colleagues (Perea, Jiménez, Suárez -Coalla, Fernández, Viña & Cuetos, 2014) adds to this debate, suggesting that both children and adults with dyslexia showed poor unfamiliar voice recognition independent of language of speech. As such, they did not show the usual native language benefit of non-autistic individuals.

Taken together, three studies have been identified which demonstrate a specific deficit amongst individuals with dyslexia when recognising previously unfamiliar voices. To the author’s knowledge, no neuroimaging evidence exists to pinpoint the location of any irregularities in brain activation during voice processing tasks. Equally, no studies have been found which examine familiar voice recognition. As such, these data cannot contribute to the debate on a distinction between unfamiliar voice differentiation and familiar voice recognition. On this question, empirical testing in the context of a neuroimaging study is encouraged. Nevertheless, the available evidence suggests that a voice processing deficit may be demonstrated within a third clinical population, with difficulty noted when differentiating unfamiliar voices. Taken together, the clinical populations included within this section of the review present a rich and perhaps more readily available population of participants in whom deficits of voice processing in general, and distinctions between familiar and unfamiliar voice processing in particular, can be explored.

5. Voice processing deficits as part of a multimodal impairment

So far, voice processing has been discussed as a unimodal deficit. In this regard, the condition of phonagnosia (inability to process voices) mirrors the parallel condition of prosopagnosia (inability to process faces). Patients with either condition commonly have the capacity to compensate for a deficit in one modality through using information available in other modalities. For example, the phonagnosic patient QR (Hailstone et al., 2010) remained able to recognise familiar people by their face despite her lack of ability to recognise their voices. Similarly, the prosopagnosic patient SB described by Hoover, Demonet and Steeves (2010) was able to recognise people by their voice despite a profound inability to recognise the face.

This said, recognition deficits may also be demonstrated as part of a multimodal impairment, with combined deficits to both face and voice processing most frequently reported. In fact, Gainotti (2014, 2015) and Liu, Corrow, Pancaroglu, Duchaine & Barton (2015) have both highlighted the need to ensure the systematic testing of patients across modalities no matter what their primary complaint, in order to avoid misclassifications of diagnoses (see Gainotti’s discussion of patient M.meV described by Boudouresques, Poncet, Cherif & Balzamo (1979), later extended by Sergent & Poncet (1990). The purpose of this section of the review is to explore the incidence of multimodal impairments in person recognition in order to inform the explanations of voice processing deficits and in order to improve models of voice processing capabilities.

5.1 Clinical demonstration of multimodal impairment

Where care has been taken to test face and voice processing (and perhaps name processing as well), cases can be identified which demonstrate multimodal impairments when recognising familiar individuals. De Renzi (1986), for example, describes a patient who, following herpes simplex encephalitis, showed damage to the anterior parts of the temporal lobe. He displayed an inability to recognise familiar faces despite a preserved ability to discriminate unfamiliar faces. Additionally, however, he was unable to recognise familiar voices making his case a very clear example of multimodal impairment. Other patients with multimodal impairments are reported by Ellis, Young and Critchley (1989 – patient KS), Hanley, Young and Pearson (1989 – patient BD), Neuner and Schweinberger (2000 – patients #13 and #37), Gainotti, Barbier and Marra (2003 – patient CO), von Kriegstein, Kleinsmith and Giraud (2006 – patient SO), and Hailstone et al. (2010 – patient KL), and Gainotti (2011) provides an excellent consideration of these and additional cases.

Liu and colleagues add to this discussion through the exploration of voice processing in 10 acquired prosopagnosic patients (Liu, Pancaroglu, Hills, Duchaine & Barton, 2014). Upon investigation, all showed face recognition deficits both when processing famous and unfamiliar stimuli, as befitting their diagnosis of prosopagnosia. However, all right-hemisphere patients showed normal voice processing. In contrast, both patients with bilateral damage involving the anterior temporal lobes showed voice recognition deficits alongside their face recognition deficits, providing demonstration of multimodal impairment. The emergence of these cases supported the demonstration of multimodal impairments in the single cases noted above. However, such cases did appear to be the exception rather than the norm within Liu et al.’s (2014) group. (See also group studies provided by Hailstone, Ridgway, Bartlett, Goll, Buckley, Crutch, & Warren, 2011; Josephs, Whitwell, Vemuri, Senjem, Boeve, Knopman, Smith, Ivnik, Petersen & Jack, 2008).

These data are supported by those of Liu et al. (2015) who examined the performance of 12 developmental prosopagnosic patients. Again, a single case existed (DP035) whose voice processing was borderline, but whose voice deficit was of a similar severity to his face deficit. Interestingly, however, he was unaware of any voice deficit, highlighting the value of formal testing.

Whilst it may be tempting to conclude that patients displaying multimodal impairments have damage to a single part of the brain responsible for the integration of modalities (described as a ‘person identity node’ in models of person recognition (Burton, Bruce & Johnston, 1990)), Barton and Corrow (2016) have suggested that multimodal deficits may simply arise from multiple lesions to separate, and non-overlapping unimodal areas (see Joassin, Pesenti, Maurage, Verreckt, Bruyer & Campanella, 2011). More specifically, they suggested that ventral right temporal lesions (to the FFA area) may result in face deficits, dorsal right temporal lesions (to Belin et al.’s 2000 TVA area) may result in voice deficits, and damage encapsulating both areas may result in the multimodal deficits described above.

5.2. Evidence for cross-modal talk between face and voice areas

This co-occurrence of face and voice deficits, no matter how rare, is important to note in and of itself. However, it also suggests the possibility that multimodal deficits could arise through the absence of normal multimodal cross-talk. In fact, in describing the case of developmental prosopagnosic patient SO, von Kriegstein et al. (2006) explored the possibility that, without normal face processing, additional input to the voice area (that might otherwise occur through cross-talk) may be compromised. Motivated by this question, SO was found to be worse than controls on familiar face recognition (as per her designation as prosopagnosic). Additionally, she was also found to be significantly worse than controls on familiar voice recognition (suggestive of a multimodal impairment). Despite this, fMRI data indicated cross-modal responses in the face area when presented with a familiar and recognisable voice, and showed direct functional connectivity between the brain regions responsible for face and voice processing.

These data are of value in several regards. First, they confirmed the prediction of cross-talk between areas responsible for the unimodal processing of faces and voices. Second, they suggested that mere existence of the structural links is not sufficient for this beneficial cross-talk to result. Third, they suggested that without normal activity in the face area, voice recognition may suffer.

Interestingly, these indications of cross-talk are supported by several studies with healthy individuals. Von Kriegstein et al. (2005), for instance noted activation in the right anterior portion of the FFA when nine participants were asked to focus on the identity of voices of personally familiar individuals whose faces they knew. In other words, the presentation of a familiar voice triggered activation in the brain area responsible for processing the familiar face, suggesting direct coupling of the FFA and the STS during familiar speaker recognition. This direct coupling was confirmed in a MEG study, in which activation was recorded in the FFA when participants were presented with the voices of speakers whose faces and voices had been learned (Schall, Kiebel, Maess & von Kriegstein, 2013). Moreover, synchronised audiovisual presentation of familiar speakers led to a direct and measurable benefit very early in the processing window both when recognising voices (Robertson & Schweinberger, 2010), and when recognising faces (Schweinberger & Robertson, 2017, Expt 3), with an impact recorded just 50-80ms post-onset (Schweinberger, Kloth & Robertson, 2011).

The strongest evidence in support of cross-talk, however, comes from Blank, Anwander and von Kriegstein (2011), who used a combination of functional and diffusion magnetic resonance imaging. They asked 21 healthy individuals to complete a voice recognition task with three speakers whose faces and voices had been learned. Through application of probabilistic tractography, the results revealed direct structural links between face and voice areas within the brain, with the strongest links existing between the FFA and middle and anterior parts of the STS rather than to more posterior STS regions. This exciting evidence provides demonstration of the physical basis for multimodal cross-talk between the faces and voices of familiar individuals.

Taken together, this section of the review suggests that deficits in familiar voice processing may occur as part of a multimodal deficit in person recognition, most often involving the face and the voice. Exploration of one such multimodal patient, SO, suggests the possibility that face and voice deficits may co-occur because the normal benefit gained from multimodal cross-talk could be lost when damage is sustained to one or other unimodal brain area. This offers an interesting possibility when accounting for the multimodal face-voice deficits described within this section. Perhaps of most significance, though, is the discovery of structural as well as functional links between voice and face areas within the brain, and these are now feeding consideration of new and exciting multimodal models of person recognition. The consideration of empirical evidence that follows will allow a reflection on progress in this endeavour.

6. Empirical evidence

The previous reflection on multimodal recognition deficits, and on face-voice cross talk within functionally and structurally linked brain regions, may seem like an aside given the primary purpose of distinguishing between familiar and unfamiliar voice processing. However, its inclusion sets a context for consideration of the wealth of empirical data. When evaluating the empirical literature, a distinction between familiar and unfamiliar voice processing is stark. In fact, several researchers consider that the processes used when recognising familiar speakers are quite different to those used when discriminating between unfamiliar or once-heard speakers (Clifford, 1980; Papcun, Kreiman & Davis, 1989; Gainotti, 2011; Schweinberger & Robertson, 2017; van Lancker, Kreiman and Emmorey, 1985). More specifically, they suggest that familiar speaker recognition relies on a comparison to a unique and integrated set of cues that form a Gestalt-like pattern, whilst unfamiliar speaker discrimination relies on a perceptual pattern-matching process (see Kreiman & Sidtis, 2011). As such, empirical research with familiar voices has been dominated by a high-level, and often multimodal, approach supporting recognition, whilst empirical research with unfamiliar voices has been dominated by a consideration of the low-level acoustic features supporting discrimination. Accordingly, the previous reflection on multimodal person recognition deficits provides a useful introduction when considering the first of the empirical domains – familiar voice recognition.

6.1 Empirical evidence regarding familiar voice recognition

Developmental research examining familiar voice recognition in healthy populations has suggested that infants may enter the world primed to attend to voices. Innovative work using a variety of methods has shown that infants at just 1 month old preferred their mother’s voice over an unfamiliar female voice (see Panneton Cooper, Abraham, Berman & Staska, 1997). This early preference may reflect exposure to the mother’s voice whilst still in the womb. Indeed, measurement of the fetal heart rate through the mother’s abdomen during gestation has shown selective responding to the mother’s voice relative to an unfamiliar female’s voice in the last months before birth (Kisilevsky, Hains, Brown, Lee, Cowperthwaite, Stutzman, Swansburg, Lee, Xie, Huang, Ye, Zhang & Wang, 2009). This early predisposition to recognise familiar voices appears to dissociate from the later developmental trajectory for the processing of unfamiliar voices (Mann, Diamond & Carey, 1979). As such, humans seem to be predisposed to attend to familiar voices as a way of recognising individuals who are important to them, and they develop a facility with unfamiliar voices at a later point.

By adulthood, familiar voice recognition has been demonstrated to be relatively robust. The existence of voice adaptation after-effects (Zäske, Schweinberger & Kawahara, 2010) suggested that we develop, and then call on, high-level voice representations or prototypes when recognising familiar voices, enabling us to recognise a familiar speaker regardless of the low-level, acoustic features associated with what they are saying. Moreover, familiar voice recognition has shown some resistance to conditions that might otherwise impair performance. For instance, Hollien, Majewski and Doherty (1982) showed near-perfect recognition performance when listeners were asked to recognise the voices of speakers they were familiar with. Moreover, this fell only to 79% when the speakers tried to disguise their voices to avoid identification. By contrast, when the voices were unfamiliar, disguise provided a substantial disruption.

This said, a number of studies exist which show that familiar voice recognition is not perfect. For instance, Goldstein and Chance (1985) revealed a high level of false recognition to unfamiliar speakers when they were presented amongst the familiar voices from a fraternity group. Similarly, Ladefoged and Ladefoged (1980) showed poor recognition of the voice of P Ladefoged’s own mother when her voice was presented within an experimental context. Both failures may reflect the powerful influence of high level expectation or context in the recognition task. More recently, Skuk et al. (2013, 2017) noted that the extent of difficulty experienced when recognising personally familiar voices tends to vary according to sex of speaker and sex of listener. In this regard, not only can the recognition of a familiar voice be a difficult task, but it can also show measurable variation across listeners.

This suggestion of individual differences in voice recognition performance has been confirmed recently in a large-scale online survey (Shilowich & Biederman, 2016). This examined the capacity of 730 participants to identify fifty US celebrity voices. Even after familiarity with the celebrities had been taken account of, performance was found to vary considerably, with a substantially higher number (3.2%) scoring more poorly than might have been expected based on a normal distribution (1.1%). Interestingly, however, the participants in this study were not necessarily aware of their true voice processing capabilities, as shown through that fact that 18 of the 20 poorest performers rated their voice recognition abilities as average or above average. In common with a similar lack of awareness in clinical patients when self-assessing voice recognition abilities, it is possible that voice processing difficulties may often go unnoticed. This may reflect two facts: voices are not the only cue we rely upon for identification; and they are not usually encountered on their own.

6.1.1 The relative strength of faces and voices

Arguably, the strongest context for voice recognition occurs when the voice is presented along with its face (Ellis, Jones & Mosdell, 1997). Stevenage, Neil and Hamlin (2014) revealed how influential the facial context can be to voice recognition by using a mismatch task. Within this task, the voices and faces of celebrities formed either corresponding pairs (voice and face of the same celebrity), semantically associated pairs (voice of one celebrity and face of a semantically associated celebrity), or unrelated pairs (voice of one celebrity with the face of an unrelated celebrity). When face recognition was tested, performance was strong regardless of the identity of the co-presented voice. In contrast, voice recognition was substantially affected by the identity of the co-presented face.

This same cost/benefit pattern was noted in studies examining audiovisual integration (Schweinberger, Robertson & Kaufmann, 2007), with a benefit being noted when a voice was accompanied by its corresponding face, and a cost noted when the voice was accompanied by a non-corresponding face. Importantly, these face-voice combinations had to be presented in synchrony or near-synchrony suggesting a small temporal window within which the activation of a face may benefit the recognition of a voice (see Robertson & Schweinberger, 2010; Schweinberger & Robertson, 2017). In an interesting recent demonstration, Schweinberger and Robertson (2017) noted that the cost/benefit pattern was bidirectional and could be demonstrated when recognising the voice and when recognising the (degraded) face. Effects were, however, smaller when recognising faces, echoing the asymmetry noted by Stevenage et al. (2014) in the mismatching task. As may be expected, these cost/benefit effects were completely absent when stimuli were unfamiliar.

These demonstrations of a benefit to voice recognition in the presence of a corresponding face invited researchers to consider voice and face recognition within a multimodal person recognition framework. Testing the viability of such an approach, a number of studies sought evidence of cross-modal priming from the face to the voice and vice versa. First in this endeavour was the work of Schweinberger, Herholz and Stief (1997) who successfully demonstrated cross-modal priming of voice recognition through prior presentation of the corresponding face. Again, as may be expected, priming effects did not exist if the target was unfamiliar. This finding was replicated by Ellis et al. (1997) when the target voice immediately followed the prime face. However, cross-modal priming was demonstrated over a longer prime-test period by Stevenage, Hugill and Lewis (2012). Interestingly, in this latter study, the extent of priming of voices by faces exceeded the priming of faces by voices. Again, this asymmetry is similar to that shown by Stevenage et al. (2014) and Schweinberger and Robertson (2017), suggesting that, whilst face and voice recognition pathways may co-exist in a multimodal framework, they may not be of equivalent strength.

In this vein, a considerable literature now exists which supports the relative weakness of voices compared to faces as cues to enable recognition, and retrieval of subsequent person-related information. For instance, Hanley, Smith and Hadfield (1998) demonstrated significantly poorer recognition of celebrities, and significantly more ‘familiar only’ experiences, from the voice than the face, highlighting the difficulty of the voice recognition task. Even when voice and face recognition was equated by substantially blurring the face (Hanley & Turner, 2000), poorer recall was noted when the voice was used as a cue for the retrieval of both semantic information (Barsics & Brédart, 2012a; Hanley & Damjanovic, 2009) and episodic information (Barsics & Brédart, 2012a; Damjanovic & Hanley, 2007).

Brédart and colleagues questioned whether this poorer retrieval of information from voices than faces may have arisen because celebrity voices have been experienced less often given the dominance of images in the media. However, the same pattern of retrieval was also evident when exposure to target voices and faces was better controlled through using personally familiar (Barsics & Brédart, 2011; Brédart, Barsics & Hanley, 2009), or newly learned stimuli (Barsics & Brédart, 2012b). As a whole, this systematic body of research suggests that, whilst voice and face recognition pathways may co-exist within a multimodal person recognition framework, the voice recognition pathway is substantially weaker than the face recognition pathway (Brédart & Barsics, 2012).

6.1.2 Multimodal models of person perception and a consideration of IAC

These results have been expressed with reference to the Interactive Activation and Competition (IAC) model of Burton, Bruce and Johnston (1990). This model extended the information processing model of Bruce and Young (1986) and suggested that the successful activation of unimodal recognition units for faces (FRUs) and for voices (VRUs) triggered the activation of a supra-modal person identity mode (PIN) at which a familiarity decision was made. Successful activation of the PIN subsequently enabled access to a common pool of semantic information allowing retrieval of semantic details, and then the name.

The IAC model has played a substantial role in accounting for a considerable literature in the person perception domain, and has been adaptable enough to explain priming effects (Burton, et al., 1990; Ellis, Burton, Young & Flude, 1997), covert recognition (Young, Hellawell & De Haan, 1988), naming difficulties (Young, Hay & Ellis, 1985), nominal competitor effects (in which multiple names for the same person compete – Valentine, Hollis & Moore, 1999, Stevenage & Lewis, 2005), and reverse fan effects (in which multiple pieces of information about the same person can assist semantic retrieval – Brédart, Valentine, Calder & Gassi, 1995). More recent considerations of the data have, however, suggested that the IAC model may provide a misleading summary of the true processes. Indeed, a number of authors now contest the concept of the PIN as a supra-modal point at which familiarity is determined, and as a single gateway via which semantic and name information can be retrieved. Five lines of evidence contribute to this view. This evidence is discussed in detail by Gainotti (2007, 2011, 2014, 2015) but is summarised here given its importance.

The first line of evidence comes from the consideration of patients with person recognition disorders. Indeed, the very existence of patients who demonstrate a dissociation of face, voice and name recognition deficits and yet show compensation for the failure of one modality by relying on another, has raised questions. If familiarity is felt at the level of the PIN, and semantic retrieval proceeds via the PIN, then compromised feelings of familiarity in one modality would imply impairment at the PIN. Consequently, the establishment of familiarity together with semantic retrieval should not be possible via another modality. Nevertheless, patients with unimodal person recognition deficits are able to recognise familiar individuals, and can retrieve semantic information about them, via another modality. Familiarity decisions, and access to semantic information must therefore be located at unimodal levels before convergence at the supra-modal PIN.

The second line of evidence comes from the study of patients who show lesions to the anterior temporal lobe (ATL) of either the right or the left hemisphere. In reviewing the data from 38 patients (21 right ATL; 17 left ATL), Gainotti (2007) noted that impairments in familiarity judgements were much more common in patients with right ATL lesions than in those with left ATL lesions, and, in the former group, these impairments occurred more often to faces than to names. This again suggests that familiarity is unlikely to be determined at a common supra-modal PIN, but must instead be located at a unimodal level with some specialisation across the hemispheres.

A third line of evidence comes from the consideration of semantic information retrieval (Gainotti, 2007). If the IAC model is correct and semantic information is retrieved via a common supra-modal PIN, then semantic retrieval should be equivalent in patients who have intact familiarity judgements, regardless of the lateralisation of their lesion and regardless of the modality in which a patient is cued. In fact, the data are quite at odds with this assumption. When familiarity judgements were intact or only moderately impaired across modalities, right ATL patients nevertheless showed a greater severity of impairment when retrieving semantic information. Furthermore, this difficulty was far greater when cued by the face than the name. In contrast, most left ATL patients showed a milder impairment and a relatively equal loss of semantic information from faces and names. These data again suggest that semantic information is unlikely to be stored in a single amodal store accessed via a supra-modal PIN. Instead, qualitatively different forms of semantic information appear to be associated with each hemisphere.

The fourth line of evidence against familiarity decisions at the supra-modal PIN level comes from the empirical work using blurred faces (reviewed in Section 6.1.1 above). The approach within these studies was to use facial blurring to reduce face recognition performance down to the level of voice recognition performance. As above, under these conditions, it may be assumed that the activation of the supra-modal PIN by both (blurred) faces and voices would be equated. If this is indeed the case, there is no explanation for the fact that semantic retrieval was nevertheless weaker from voices than from faces.

The final line of evidence against familiarity decisions at the supra-modal PIN level comes from the work of von Kriegstein and colleagues, and Schweinberger and colleagues, and is summarised well by Blank et al. (2014). These research teams have provided strong evidence to indicate direct cross-talk between voice and face areas, resulting in priming (Schweinberger et al., 1997) audiovisual costs and benefits (Robertson & Schweinberger, 2010; Schweinberger et al., 2011; Schweinberger & Robertson, 2017), functional cross-talk allowing face areas to be activated by a familiar voice (von Kriegstein et al., 2005) and structural links between face and voice areas (Blank et al., 2011). In its strongest form, these lines of evidence question the need for a supra-modal PIN *at all* if information from each modality can actually be shared directly and very early on.

6.1.3: An alternative model – the voice as an ‘auditory face’

Set against this discussion is the work of Belin and colleagues (Belin, Fecteau & Bédard, 2004) who suggested that the voice may be considered as an ‘auditory face’. This model emphasized the fact that the voice could reveal more than just identity. More specifically, it could mirror the face in some regards by providing a parallel source of information regarding speech, and emotion as well as identity. This said, it is worth noting that the voice is not necessarily able to mirror the face in all regards; it cannot, for example, provide an equivalent to the directional facial cues signalled through eye gaze.

In developing the ‘auditory face’ model, Belin et al. (ibid) considered the neuropsychological evidence which identified brain sites sensitive for the processing of vocal speech, affect, and identity. They also considered the separation of vocal speech, affect and identity processing as shown in patients with selective deficits. Alongside this, however, they also incorporated the neuropsychological evidence for face-voice integration in each of the three domains of processing (speech, affect, identity). The result was a framework describing partially segregated pathways which nevertheless showed functional cross-talk to facilitate integrated processing (Belin, Bestelmeyer, Latinus & Watson, 2011; Campanella & Belin, 2007).

A considerable evidence base now exists to support the auditory face model, with a particular emphasis on the demonstration of an interaction between processing domains. For instance, Pinheiro, Rezaii, Nestor, Rauber, Spencer and Niznikiewicz (2016) showed that the identity of a speaker (self or other) influenced ERP responses when judging the affect of speech clips as positive, neutral or negative. Consequently, vocal identity was not processed independently of vocal affect. In fact, vocal identity actually assisted the perception of vocal affect.

Similarly, von Kriegstein, Smith, Patterson, Kiebel and Griffiths (2010) noted that the identity of a speaker affected the perception of speech. Participants showed brain activation in the speech area of the left posterior STG/STS which was modulated by speaker cues. Similarly, they showed brain activation in the voice area of the right posterior STG/STS during a speech recognition task. These data confirmed that speech and speaker information was also not partitioned and processed separately, but rather was combined in demonstrable ways.

The linkage of speech and identity processing from the voice was further demonstrated by Kreitewolf, Gaudrain and von Kriegstein (2014) who showed that speaker familiarity could actually enhance speech comprehension. Similarly, the results of Chandrasekaran, Chan and Wong (2011) showed an fMRI BOLD response in the left middle temporal gyrus that was sensitive to both what was said and who said it. These neuropsychological results echoed the results of behavioural studies in which a change in speaker identity affected word memory (see Mani & Schneider, 2013; Sheffert & Fowler, 1995) and the familiarity of a speaker facilitated the tracking of speech amidst noise (Johnsrude, Mackey, Hakyemez, Alexander, Trang & Carlyon, 2013). This has led to an effect known as a ‘familiar talker advantage’, in which spoken language processing was improved when the speaker was familiar (Levi, 2015; Nygaard & Pisoni, 1998). Together, these studies combined to suggest that vocal identity and vocal speech information were processed in an integrated fashion.

Taking a broader perspective, the auditory face model of Belin et al. (2004, 2011) may be reconciled with an IAC-type approach to person perception if one considers the activation of face and voice recognition units as being analogous to Belin et al.’s face and voice processing in the identity domain. However, by avoiding any explicit consideration of supra-modal convergence across faces and voices, the auditory face model assumes that familiarity is determined at a unimodal rather than a supra-modal level thereby avoiding the criticisms of the IAC approach. The model has the additional benefit of being applicable to unfamiliar as well as familiar voices in as much as speech, affect and (rapidly formed) representations of identity can be processed regardless of level of familiarity. As such, this model provides a useful transition to the consideration of empirical evidence on unfamiliar voice processing. It is to this that we now turn in the final substantive section of this review.

6.2 Empirical evidence on unfamiliar voice discrimination

The evidence base focussing on unfamiliar voice processing exists, in part, due to the requirements of forensic research examining earwitnesses responses to unfamiliar suspect voices (see Kreiman & Sidtis, 2011; Yarmey, 1995). In addition, however, there are often practical difficulties in sourcing a sufficient number of personally familiar speech clips or celebrity speech clips which are familiar to all participants (Boucher et al., 1998) and yet are free from content that may inadvertently reveal identity (Hanley & Turner, 2000). As a consequence, a considerable empirical literature exists in the area of unfamiliar voice processing.

At the outset, it is worth noting the obvious: unfamiliar voices usually do not interest the listener as much as familiar voices, and may even be considered as part of the acoustic background against which the more meaningful familiar voices ‘pop’ out (Kreiman & Sidtis, 2011). Nevertheless, some similarity of processing may exist. Indeed, the results of research using unfamiliar voices can also be interpreted within the auditory face model presented by Belin et al. (2004, 2011) inasmuch as speech, affect and identity of an unfamiliar speaker can be processed, and can be seen to interact in predictable ways. For instance, demonstrations of an interaction between speech and identity processing are evident when unfamiliar speakers talk in a non-native language (Goggin, Thompson, Strube & Simental, 1991; Perrachione & Wong, 2007). Under such conditions, the capacity to recognise the speaker becomes compromised, even when listening conditions are optimised through the use of long speech clips at study and test, and a relatively short delay between the two (Phillipon, Cherryman, Bull & Vrij, 2007). Goggin et al. (1991) accounted for this ‘other-language’ effect by suggesting **that, whilst we have the capacity to process speech, affect and identity in an unfamiliar voice, we cannot help but prioritise some aspects over others. Specifically, they suggested that we may automatically orient our attention to process speech content, at times to the detriment of other aspects of vocal processing** (Stevenage & Neil, 2014).

In a related line of work, Vitevitch (2003) used a shadowing task to focus attention onto the speech rather than the speaker. As a result, participants again oriented their attention to process speech over identity, and were again very poor at processing speaker identity. Within this study, their inability was revealed through ‘change deafness’ – a failure to detect a change of speaker mid-trial. Similarly, Fenn, Shintel, Atkins, Skipper, Bond and Nusbaum (2011) used an interactive telephone conversation to focus participants’ attention onto the speech task, and provided a similar demonstration of change deafness when the speaker changed mid-call.

6.2.1 The relative weakness of the voice versus the face as a cue to identity

Alongside the literature suggesting that voice identification is not the most important task when listening to a speaker, there is also a considerable literature suggesting that the voice is not the most important cue when trying to identify that speaker. Instead, the face rather than the voice seems to take priority when processing the identity of an unfamiliar speaker. One simple explanation for this is that voices are usually experienced along with faces, and the latter represent a robust means of determining identity. This is Stevenage, Hugill and Lewis’s (2012) differential utilisation hypothesis. The result is a highly adaptive processing framework in which attention can be directed more to the voice for speech and affect processing, and it can be directed more to the face for identity processing. This enables both perceptual channels to be optimally used whilst retaining redundancy should either become compromised. However, it also means that recognition via the voice is secondary to recognition via the face.

Demonstration of this point comes from a growing body of work examining what is known as the ‘face overshadowing effect’. This was first demonstrated by Legge, Grossman and Pieper (1984), and Armstrong and McKelvie (1996), but was named by Cook and Wilding (1997a). All studies however revealed that recognition of the briefly heard voice of a stranger is weakened if the face is present at study. This overshadowing effect may be reduced if the utterance is repeated, or if the participant is pre-exposed to the face so that attention can be shifted back to the voice when subsequently presented. However, an explicit instruction to attend to the voice was not sufficient to mitigate the overshadowing (Cook & Wilding, 2001). The authors concluded that attention to the face in an audiovisual context was automatic and unstoppable.

This primacy of the face over the voice in a recognition context has been supported by the findings of two more recent studies. The first provided evidence to suggest that whilst voice recognition was impaired by the co-presentation of a face, face recognition was not affected by the co-presentation of a voice (Stevenage, Howland & Tippelt, 2011), showing once again the importance of the face over the voice when processing identity. Moreover, this overshadowing effect was not merely the result of a competing stimulus in the visual domain. It resulted specifically through the co-presentation of a *face* alongside the voice, as shown by the fact that overshadowing was removed when the face was inverted (Tomlin, Stevenage & Hammond, 2016). As such, there is clear evidence to suggest that the visual channel is relied upon to a greater extent than the auditory channel when processing identity.

This said, three studies exist which suggest that the combination of voice and face may actually work in conjunction with one another to facilitate performance in some tasks. In this regard, Sheffert and Olson (2004) showed a benefit to voice learning when the face was also present at study, an effect that was mirrored by Zäske, Mühl and Schweinberger (2015) towards the end of a sequence of learning trials. Similarly, Zweig, Suzuki and Grabowecky (2015) showed that the presentation of a voice may substantially facilitate the detection of its associated face in a crowd. Clearly, there is a need to resolve the conditions under which the audiovisual presentation of face and voice will be helpful and when it will be harmful. What is evident, however, is that the face and voice readily interact when processing identity cues, with the face representing the primary cue. Moreover, and as reviewed here, this primacy of the face over the voice has been observed regardless of the familiarity of the voice.

6.2.2 Interference from competitors

All preceding demonstrations in this section have shown that unfamiliar voice discrimination is handicapped by the priorities of language processing and face processing. The final example of an impairment within the unfamiliar voice processing literature actually rests on competition from other voices. Using a distraction task, Stevenage, Neil, Barlow, Dyson, Eaton-Brown and Parsons (2013) presented a target voice and a test voice in a same/different matching task. However, between target and test, participants were presented with either zero, two or four distractor voices. The results suggested that unfamiliar voice discrimination was significantly impaired as soon as any distraction was introduced. In contrast, a parallel face discrimination task showed that performance remained strong and robust no matter how much distraction was introduced.

There are several ways to interpret these findings. It is possible that voice recognition suffered in the face of distractors because of a source-monitoring error in determining which voice from memory was the target voice. It could also be that the memory for the target voice was weakened through the introduction of distractors thus reducing subsequent identification. Either way, these results start to suggest that the processing of low-level vocal properties associated with identity may be somewhat fragile. It is to this that we now turn.

6.2.3 Processing low level auditory characteristics of unfamiliar voices

The available wisdom on the processing of unfamiliar voices suggests that they may be processed relative to an average or norm, so that deviations from that norm can be identified (Papcun, Kreiman & Davis, 1989). Over time, one’s memory of an unfamiliar voice may lose the detail regarding these critical deviations from the norm, and consequently, the unfamiliar voice may become more ‘average’ in memory. This may account for the high rate of false identifications of typical-sounding voices after a delay between study and test (Mervis & Rosch, 1981). To date, there have been only a handful of studies which have examined the low-level features or dimensions used to determine a deviation from the norm.

First in this regard is the work of Kreiman and Papcun (1991) who asked participants to rate the similarity of voices following a short-term discrimination task, and following a long-term memory task for initially unfamiliar voices. Application of a multidimensional scaling analysis allowed the authors to identify the main dimensions used when processing the voices. This revealed four factors (masculinity, creakiness, variability and mood) following the discrimination task, and a simpler set of three factors (masculinity, breathiness and variability) following the memory task, suggesting that the features one attends to may change with time. Avoiding subjective labels, Baumann and Belin (2010) also examined the acoustic features used when listening to unfamiliar speakers. They revealed two dimensions of interest: fundamental frequency (pitch), and formant characteristics associated with the vocal tract (c.f., the results of Chhabra et al. (2012) with individuals with schizophrenia). Against this backdrop, one might presume that any manipulation which modified these characteristics, or affected their perception, would compromise voice processing.

Accordingly, Lavan, Scott and McGettigan (2016) recently showed that the capacity to discriminate between unfamiliar speakers substantially fell when acoustic cues were modified between study and test through switching between acted versus spontaneous laughter sounds. Similarly, Mullenix and colleagues (Mullenix, Stern, Grounds, Kalas, Flaherty, Kowalok, May & Tessmer, 2009) found that participants showed a reduced ability to discriminate between unfamiliar speakers when voice pitch was modified between study and test through the use of synthesized clips. Specifically, participants tended to select test voices that were lower in pitch than the low-pitch target, and higher in pitch than the high-pitched target, suggesting either a distortion in memory, or a lack of sensitivity to subtle changes to voice-relevant features. Additionally, recall the participants of Hollien et al. (1982) described previously, who were unable to recognise the briefly studied voices of unfamiliar speakers if, at test, their voices were intentionally disguised.

Alongside these results, the findings of Orchard and Yarmey (1995) are relevant. They tested participant memory for unfamiliar voices that were presented either whispering (to remove pitch information) or speaking normally. After a two-day delay, participants took part in a voice line-up with voices presented either in the studied format (whisper or normal) or in the opposite format. The results suggested a clear impairment in performance when voices were heard whispering, and when voices changed format between study and test. As such, the available evidence suggests that participants use a limited range of low level acoustic cues when processing the unfamiliar voice, and are easily thrown when those cues become modified or unavailable.

Finally within this series of studies, it may be anticipated that performance will be impaired when a voice clip is too brief to appreciate either fundamental frequency or formant information. This was confirmed by the work of Yarmey and Matthys (1992) and Cook and Wilding (1997b) who both showed that the identification of a stranger’s voice fell significantly as clip length became shorter. Indeed, Cook and Wilding suggested that, when trying to identify an unfamiliar speaker, the length of a clip was more important that the vowel variety within it.

6.2.4 Similarities and differences highlighted through empirical exploration of familiar and unfamiliar voice processing

As a whole, the empirical literature using unfamiliar voices has resulted in tests of unfamiliar voice discrimination or the short-term recognition of recently learned, or once-heard, voices. As a consequence, the empirical literature on unfamiliar voice processing has revealed a host of low-level acoustic features important to the processing of unfamiliar voices. These observations support the earlier suggestion that familiar voice processing involves comparison to a unique stored pattern or Gestalt, whilst unfamiliar voice processing involves a lower-level perceptual matching process.

Despite this, the literatures on familiar and on unfamiliar voice processing show some common influences through the interaction of identity and speech processing, and through the primacy of face identity over voice identity processing. The literatures also reveal a common influence of one acoustic characteristic when processing familiar and unfamiliar voices – vocal distinctiveness. Indeed, when considering unfamiliar voice processing, Yarmey (1991) showed a notable consistency when describing a distinctive voice over immediate, 24h and 1-week delays suggesting that it may be held more successfully in memory compared to a more typical voice. Similarly, Mullenix, Ross, Smith, Kuykendall, Conard and Barb (2009) showed fewer errors on a delayed old/new recognition task, and Sørensen (2012) found superior performance in a delayed line-up task, when the unfamiliar voices were distinctive rather than typical. In a nice cross-modal study, Bülthoff and Newell (2015) recently showed that face recognition was better if an unfamiliar face had been paired with a distinctive rather than a typical voice at study. The same advantage was not evident when the face had been paired with a distinctive non-vocal sound at study suggesting some benefit of the distinctive voice when generating a multimodal representation.

When considering familiar voice recognition, a distinctiveness advantage is again apparent. For instance, Foulkes and Barron (2000) found superior recognition of recorded telephone messages by members of a close friendship group when the speaker sounded distinctive. Similarly, Skuk and Schweinberger (2013) showed that distinctive personally-familiar voices were recognised more easily than typical ones. In a striking demonstration of a distinctiveness advantage, van Lancker et al. (1985) showed that some celebrity voices could even been recognised despite temporal reversal, and the authors noted that these effects may have been driven by vocal distinctiveness of the individuals themselves. Finally in this set are the results of Barsics and Brédart (2012c) who showed that semantic information retrieval was superior when the familiar speaker was more distinctive.

The importance of vocal distinctiveness when processing both familiar and unfamiliar voices is of interest in and of itself. It is likely to arise from the comparison of the low-level features of a voice relative to the norm or average of the population of voices normally encountered. Consequently, the notion of distinctiveness draws both on low-level acoustic perception (as explored with unfamiliar voices) and higher-level issues to do with a listener’s experience with a lifetime of voices (as explored with familiar voices). As such, perhaps the demonstration of a distinctiveness advantage for both familiar and unfamiliar voices is to be expected, sitting at the intersection of the two stimulus types as it does. Equally, however, it may herald a view in which the low-level processing of unfamiliar voices is not entirely separate from the higher-level processing of familiar voices. Instead, bottom-up and top-down influences may co-exist.

In summarising the empirical literature on familiar and unfamiliar voice processing, it is clear that similarities and differences may be identified. Similarities exist both in terms of the applicability of the auditory face model, the primacy of the face over the voice in identification decisions, and in terms of the influence of distinctiveness. This said, differences emerge in that the empirical research base has drawn a divide between the low-level perceptual processing associated with unfamiliar voices and the higher-level multimodal recognition associated with familiar voices. Accordingly, studies exist, and have been described, which reveal empirical effects in familiar voice processing but not in unfamiliar processing, such as the demonstration of priming effects, resistance to disguise, costs and benefits of synchronous audiovisual presentation, and the retrieval of episodic, semantic or name information. Taken together, the empirical evidence may be considered to suggest an incomplete distinction between familiar and unfamiliar processing. It is this conclusion which guides the consideration of future work to follow.

7. Conclusions

The purpose of the present review has been to bring together the available literature in four now quite mature areas of voice research. These cover the neuropsychological study of healthy individuals, the neuropsychological investigation of brain-damaged individuals, the exploration of voice recognition deficits in less commonly studied clinical conditions, and finally the empirical data on familiar and unfamiliar voice processing. The objective within this review has been to re-examine each area of literature in order to inform the question of whether unfamiliar voice discrimination is indeed distinct from familiar voice recognition.

The evidence in each area supports such a distinction between familiar and unfamiliar voice processing. However, it also suggests that this distinction may not be complete. For instance, when considering the neuropsychological study of healthy individuals, voice processing areas have been identified particularly within the right temporal cortex, with a broad posterior-to-anterior differentiation for unfamiliar and familiar voices respectively. Equally, the neuropsychological study of brain-damaged individuals with a primary deficit in person recognition suggests that damage to the areas described above lead to corresponding deficits in voice processing affecting unfamiliar and familiar voices in predictable ways.

This said, exceptions to the expected pattern of damage and deficit do exist. For example, whilst a dissociation of the capabilities with familiar and unfamiliar voices is clear and unequivocal, the areas of activation in healthy individuals, or of damage in patient groups, do not always completely mirror the posterior to anterior STS/STG separation expected (see Blank et al., 2014). Further study of these healthy participants and patient groups would be of great value, with particular care taken to test face and voice processing, and to explore performance with both familiar and unfamiliar stimuli. Echoing the comments of Gainotti (2011) and Liu et al. (2015), it is hoped that this more comprehensive approach may serve to disambiguate current findings.

When considering the voice processing capabilities of three other clinical groups for whom person recognition impairments were not a primary complaint, evidence again suggested a distinction between familiar and unfamiliar voice processing. Indeed, when examining the abilities of individuals with schizophrenia, autism, and dyslexia relative to appropriate controls, deficits were revealed in familiar voice recognition, unfamiliar voice discrimination, or both. Usefully, the examination of these populations may suggest a more readily available group of participants in whom voice recognition deficits – familiar or otherwise – may be explored. Given the relatively small number of existing studies with these populations, further work in this fruitful area is strongly encouraged.

Finally, when examining the considerable literature describing empirical studies, an incomplete distinction between familiar voice recognition and unfamiliar voice discrimination is indicated. The research using familiar voices has largely studied familiar voice recognition within the context of multimodal person recognition frameworks. As a result, useful refinements to these frameworks have been suggested, especially regarding the point at which familiarity decisions are taken. The use of unfamiliar stimuli as a control set in these sorts of recognition studies has demonstrated the benefits that arise when a multimodal representation exists for a familiar target. In contrast, the research focussed on unfamiliar voices has been directed more to the identification of the low-level acoustic features supporting voice processing, and the conditions under which their use may be hampered. As such, the empirical investigation of unfamiliar voice processing has shown limited overlap with the investigation of familiar voice processing. This said, common influences do exist involving the primacy of speech perception, face perception and distinctiveness effects, and this overlap suggests valuable avenues for future consideration.

8. Avenues for Future Work

The motivation for the current review was the observation of van Lancker and Kreiman (1987) that familiar voice recognition is different to unfamiliar voice discrimination. This observation receives support here. Indeed, the suggestion of a distinction (partial or otherwise) between familiar and unfamiliar voice processing is reminiscent of the conclusion drawn by Megreya and Burton (2006) when considering face processing. As a result, it may be important to appreciate the limits of generalisability when work is conducted using familiar or unfamiliar voices as stimuli.

Whilst a certain level of clarity may come with this conclusion, it is important to consider what might change, or what questions may now be asked, as a result of this review. Perhaps the most interesting question given a distinction between the processing of unfamiliar and familiar voice is ‘how do listeners bridge this divide?’. In other words, how do unfamiliar voices become familiar?

In this regard, Kreiman and Sidtis (2011) have provided a very interesting discussion. In a well-argued thesis, they considered that familiar voice processing draws on a Gestalt-like representation for each voice, but that this reflects a unique combination of acoustic dimensions as dictated by the individual characteristics of each voice. In this sense, they suggested that it may not be possible to adequately capture a full list of individual parameters involved in the task of familiar voice processing. Similarly, they noted that, by focussing on specific brain regions of interest when localising familiar voice processing, researchers may have ignored contributions from an extensive network of brain areas including limbic regions associated with emotional processing, and parietal regions associated with the formation of associations. Given these two points, they considered that, whilst unfamiliar and familiar voice processing certainly differ, we may currently have only a partial understanding of the processing involved when a voice has become familiar. As such, the standards used to recognise vocal familiarity may, at present, be underspecified.

In terms of acquisition, Kreiman and Sidtis (2011) also considered that the development of familiarity with a voice may not reflect gradual learning through repeated exposure. Instead, they suggested that vocal familiarity may be acquired through processes more akin to one-trial learning (Lattal, 1995 in Kreiman & Sidtis, 2011) or imprinting (Lorenz, 1935 in Kreiman & Sidtis, 2011) in which the context provides a strong reinforcement to enable a near-instantaneous acquisition of information. This view recognises the importance of emotion, attention, association and memory both in behavioural responses, and in neurological activity across a network of brain areas, when processing a familiar voice. However, it leaves unspecified the detailed processes involved in binding unimodal representations together to form a multimodal representation. Nevertheless, Kreiman and Sidtis (2011) suggested that researchers may benefit from taking a much broader view whilst examining the emergent questions underlying voice acquisition.

Set within this context, examination of the process of voice acquisition would provide tremendous theoretical value, and some of the pertinent questions can now be articulated. At the neural level, insight into the process of voice acquisition would enable understanding of a shift from posterior STS involvement for unfamiliar voices, to anterior STS involvement for familiar voices (see Schweinberger, Kawahara, Simpson, Skuk & Zäske, 2014). At a patient level, it may support greater understanding of how familiar voices cannot be recognised, and new voices cannot be learned, and yet unfamiliar voices can be discriminated (patient KH). At a theoretical level, it would augment the development of multimodal recognition models by incorporating an understanding of rapid prototype formation, and the binding of unimodal inputs to create rich multimodal representations. Finally, at an experiential level, it would allow recognition of the importance of emotional context or significance when processing signals amidst noise in our acoustic environment.

In this regard, it is important to note that we are not without any evidence relating to the issue of vocal acquisition. Indeed, a handful of studies exist which have adopted a learning approach. At the neuropsychological level with healthy individuals, the use of adaptation studies has allowed demonstration of activity in the right posterior and anterior STS regions when a voice becomes familiar (Andics, McQueen, Petersson, Gál, Rudas & Vidnyánszky, 2010; Belin & Zatorre, 2003). Equally, at the behavioural level, the use of an adaptation study revealed adaptation after-effects when identifying learned stimuli (Latinus & Belin, 2011; Zäske et al., 2010). Both adaptation effects have been explained by presuming that the listener creates a voice representation or prototype for the learned voice against which the adaptation after-effects result. This said, a description of the processes involved in prototype formation is yet to be articulated.

In this vein, a description of prototype formation is difficult to formulate as the interaction between top-down and bottom-up processes contradicts a simple hierarchical framework in which low-level processing in the posterior STS gives way to higher-level processing in the anterior STS. Similarly, van Lancker and Kreiman’s (1987) seven patients capable of familiar voice recognition despite an inability at unfamiliar voice discrimination again challenges the hierarchical assumption that recognition builds on lower level perceptual matching. A one-trial learning mechanism is appealing as a way around the issue. However, without a fuller description, we may run the risk of labelling the process of prototype formation without fully articulating it.

Similarly, the evidence for cross-modal binding is also remarkably mixed. Consider for example the literature on face-voice matching which suggests that physical or personality-based attributes as determined from static faces, together with co-articulation as determined from dynamic speaking faces, may enable a perceiver to bind an unfamiliar face and voice together (see Smith, Dunn, Baguley & Stacey, 2016; Stevenage, Hamlin & Ford, 2016, for recent reviews). This binding may well rely on audio-visual presentation of faces and voices in synchrony or near-synchrony (Robertson & Schweinberger, 2010; Schweinberger et al., 2011, 2017), and it may well utilise the direct functional and structural cross-modal linkages revealed by von Kriegstein and colleagues as reviewed previously.

Nevertheless, clarity is threatened regarding the process of binding given two observations. First, consider the case of prosopagnosic patient SO who could not recognise faces and consequently could not provide the useful cross-modal facilitation to support voice processing, with the result that her voice processing also showed impairment. In contrast, note the evidence from a very under-utilised group of blind listeners who similarly could not recognise a face but whose auditory recognition abilities had improved to compensate (Hölig, Föcker, Best, Röder & Büchel, 2014). In this sense, the evidence is far from clear regarding the interaction between different modalities when creating a multimodal representation of a familiar person.

The mixed evidence on both prototype formation and cross-modal binding suggests that an answer to the question of how an unfamiliar voice becomes familiar is far from simple. Nevertheless, given the evidence within the present review supporting a distinction between unfamiliar and familiar voice processing, this seems to be an important next question for the voice processing community to consider.

Acknowledgements:

I would like to thank Rebecca Tomlin, and Greg Neil, for helpful early discussions towards the formation of ideas discussed here. This paper is dedicated to the memory of Allan – my dear friend.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Alba-Ferrara, L., Fernyhough, C., Weis, S., Mitchell, R.L.C., & Hausmann, M. (2012a). Contributions of emotional prosody comprehension deficits to the formation of auditory verbal hallucinations in schizophrenia. *Clinical Psychology Review,* 32, 244-250.

Alba-Ferrara, L., Weis, S., Damjanovic, L., Rowett, M., & Hausmann, M. (2012b). Voice identity recognition failure in patients with schizophrenia. *Journal of Nervous and Mental Disease, 200(9)*, 784-790 doi: 10.1097/NMD.0b013e318266f835

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

Armstrong, H.A., & McKelvie, S.J. (1996). Effect of face context on recognition memory for voices. *The Journal of General Psychology, 123(3)*, 259-270. doi: 10.1080/00221309.1996.9921278

Badcock, J.C., & Chhabra, S. (2013). Voices to reckon with: Perceptions of voice identity in clinical and non-clinical voice hearers. *Frontiers in Human Neuroscience, 7*, article 114, 1-9. doi: 10.3389.fnhum.2013.00114

Barsics, C., & Brédart, S. (2011). Recalling episodic information about personally known faces and voices. *Consciousness and Cognition,* *20*, 303-308. doi: 10.1016/j.concog.2010.03.008

Barsics, C., & Brédart, S. (2012a). Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology, 24(7)*, 789-795. doi: 10.1080/20445911.2012.692672

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

Barsics, C., & Brédart, S. (2012c). Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology, 24(7)*, 789-795. DOI: 10.1080/20445911.2012.692672

Barton, J.J.S., & Corrow, S.L. (2016). Selectivity in acquired prosopagnosia: The segregation of divergent and convergent operations. *Neuropsychologia, 83*, 76-87. doi: 10.1016/j.neuropsychologia.2015.09.015

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research, 74*, 110-120.

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

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Science, 8*, 129-135. doi: 10.1016/j.tics.2004.01.008

Belin, P., & Zatorre, R.J. (2003). Adaptation to speaker’s voice in right anterior temporal lobe. *Neuroreport, 14*, 2105-2109.

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

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*, 309-312.

Bentall, R.P. (1990). The illusion of reality: A review and integration of psychological research on hallucinations. *Psychological Bulletin, 107*, 217-232.

Benton, A.L., & van Allen, M.W. (1968). Impairment in facial recognition in patients with cerebral disease. *Cortex, 4*, 344-358.

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

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience, 7*, 12906-12915.

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

Bodamer J. (1947). Die Prosopagnosie. *Arch. Psychiatrie Z. Neurol, 179*, 6-54.

Boucher, J., Lewis, V., & Collis, G. (1998). Familiar face and voice matching and recognition in children with autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 39(2),* 171-181.

Boucher, J., Lewis, V., & Collis, G. (2000). Voice processing abilities in children with autism, children with specific language impairments, and young typically developing children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41(7)*, 847-857. doi: 10.1017/S00219663099006149

Boudouresques, J., Poncet, M., Cherif, A.A., & Balzamo, M. (1979). Agnosia for faces: Evidence of functional disorganization of a certain type of recognition of objects in the physical world. *Bulletin de l’Académie Nationale de Médecine,* 163, 695-702.

Brachacki, G.W.Z., Fawcett, A.J., & Nicholson, R.I. (1994). Adults with dyslexia have a deficit in voice recognition. *Perceptual and Motor Skills, 78*, 304-306.

Brédart, S., & Barsics, C. (2012). Recalling semantic and episodic information from faces and voices: A face advantage. *Current Directions in Psychological Science, 21(6),* 378-381.

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. doi: 10.1080/09541440802591821

Brédart, S., Valentine, T.R., Calder, A., & Gassi, L. (1995). An interactive activation model of face naming. *Quarterly Journal of Experimental Psychology - A, 48(2),* 466-486.

Bruce, V., & Young, A.W. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327.

Bülthoff, I., & Newell, F.N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition, 137*, 9-21. doi: 10.1016/j.cognition.2014.12.006

Burton, A.M., Bruce, V., & Johnston, R.A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81(3)*, 361-381. doi: 10.1111/j.2044-8295.1990.tb02367.x

Campanella, S., & Belin, P. (2007). Integrating Face and Voice in Person Perception. *Trends in Cognitive Science, 11(12),* 535-543.

Chandrasekaran, B., Chan, A.H.D., & Wong, P.C.M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience, 23(10)*, 2690-2700. doi: 10.1162/jocn.2011.21631

Chhabra, S., Badcock, J.C., Maybery, M., & Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia, 50(12)*, 2730-2735. doi: 10.1016/j.neuropsycholgia.2012.08.006

Clifford, B.R. (1980). Voice identification by human listeners. *Law and Human Behavior, 4(4),* 373-394.

Cook, S., & Wilding, J. (1997a). Earwitness Testimony 2: Voices, faces and context. *Applied Cognitive Psychology, 11,* 527-541.

Cook, S., & Wilding, J. (1997b). Earwitness Testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11(2)*, 95-111.

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the Face Overshadowing Effect. *British Journal of Psychology, 92*, 617-629.

Damjanovic, L., & Hanley, J.R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory and Cognition, 35*, 1205-1210.

De Renzi, E. (1986). Current issues in prosopagnosia, in Ellis, H.D., Jeeves, M.A., Newcombe, F., & Young, A. (Eds.), *Aspects of Face Processing*. Martinus Nisoff, Dordrecht, pp 243-252.

Ellis, A.W., Burton, A.M., Young, A., & Flude, B.M. (1997). Repetition priming between parts and wholes: Tests of a computational model of familiar face recognition. *British Journal of Psychology, 88*, 579-608.

Ellis, A.W., Young, A.W., & Critchley, E.M.R. (1989). Loss of memory for people following temporal lobe damage. *Brain, 112,* 1469-1483. doi: 10.1093/brain/112.6.1469

Ellis, H.D., Jones, D.M., & Mosdell, N. (1997). Intra- and Inter-Modal Repetition Priming of Familiar Faces and Voices. *British Journal of Psychology, 88*, 143-156.

Fenn, K.M., Shintel, H., Atkins, A.S., Skipper, J.I., Bond, V.C., & Nusbaum, H.C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *Quarterly Journal of Experimental Psychology, 64(7),* 1442-1456.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of close social network. *Forensic Linguistics: The International Journal of Speech, Language and the Law* 7, 180-198.

Gainotti, G. (2007). Different patterns of famous people recognition disorders in patients with right and left anterior temporal lesions: A systematic review. *Neuropsychologia, 45*, 1591-1607. doi: 10.1016/j.neuropsychologia.2006.12.013

Gainotti, G. (2011). What the study of voice recognition in normal subjects and brain-damaged patients tells us about models of familiar people recognition. *Neuropsychologia, 49*, 2273-2282. doi: 10.1016/j.neuropsychologia.2011.04.027

Gainotti, G. (2014). The neuropsychology of familiar person recognition from face and voice. *Psychologica Belgica, 54(3)*, 298-309. doi: <http://dx.doi.org/10.5334/pb.at>

Gainotti, G. (2015). Implications of recent findings for current cognitive models of familiar people recognition. *Neuropsychologia, 77*, 279-287. doi: 10.1016/j.neurosychologia.2015.09.002

Gainotti, G., Barbier, A., & Marra, C. (2003). Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. *Brain, 126(4)*, 792-803. doi: 10.1093/brain/awg092

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger. S.R., Warren, J.D., & Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia, 47*, 123-131. doi: 10.1016/j.neuropsychologia.2008.08.003

Goggin, J.P., Thompson, C.P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory and Cognition, 19*, 448-458.

Goldstein, A.G., & Chance, J.E. (1985). Voice recognition: The effects of faces, temporal distribution of ‘practice’ and social distance. Paper presented at the annual meeting of the Midwestern Psychology Association, Chicago, IL.

Hailstone, J.C., Crutch, S.J., Vestergaard, M.D., Patterson, R.D., & Warren, J.D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia, 48,* 1104-1114.

Hailstone, J.C., Ridgway, G.R., Bartlett, J.W., Goll, J.C., Buckley, A.H., Crutch, S.J., & Warren, J.D. (2011). Voice processing in dementia: a neuropsychological and neuroanatomical analysis. *Brain, 134(9)*, 2535-1547. doi: 10.1093/brain/awr205

**Hanley, J.R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person’s name and occupation from their voice than from their blurred face. *Memory, 17*, 830-839.**

Hanley, J.R., Smith, S.T., & Hadfield, J. (1998). I recognise you but I can’t place you. An investigation of familiar-only experiences during tests of voice and face recognition. *Quarterly Journal of Experimental Psychology, 51A(1)*, 179-195.

Hanley, J.R., & Turner, J.M. (2000). Why are familiar-only experiences more frequent for voices than for faces? *Quarterly Journal of Experimental Psychology, 53A(4)*, 1105-1116.

Hanley, J.R., Young, A.W., & Pearson, N.A. (1989). Defective recognition of familiar people. *Cognitive Neuropsychology, 6*, 179-210. doi: 10.1080/02643298908253418

Herald, S.B., Xu, X., Biederman, I., Amir, O., & Shilowich, B.E. (2014). Phonagnosia: A voice homologue to prosopagnosia. *Visual Cognition, 22(8)*, 1031-1033. doi: 10.1080/13506285.2014.960670

Hölig, C., Föcker, J., Best, A., Röder, B., & Büchel, C. (2014). Brain systems mediating voice identity processing in blind humans. *Human Brain Mapping, 35*, 4607-4619. doi: 10.1002/hbm.22498

Hollien, H., Majewski, W., & Doherty, E.T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10(2),* 139-148.

Hoover, A.E.N., Demonet, J.F., & Steeves, J.K.E. (2010). Superior voice recognition in a patient with acquired prosopagnosia and object agnosia. *Neuropsychologia, 48(13)*, 3725-3732. doi: 10.1016/j.neuropsychologia.2010.09.008

Imaizumi, S., Mori, K., Kiritani, R., et al. (1997). Vocal identification of speaker and emotion activates different brain regions. *NeuroReport, 8*, 2809-2812.

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R., & Campanella, S. (2011). Cross-modal interactions between human faces and voices involved in person recognition. *Cortex, 47*, 367-376. doi: 10.1016/j.cortex.2010.03.003

Johns, L.C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., & McGuire, P.K. (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychological Medicine, 31*, 705-715.

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. doi: 10.1177/0956797613482467

Josephs, K.A., Whitwell, J.L., Vemuri, P., Senjem, M.L., Boeve, B.F., Knopman, D.S., Smith, G.E., Ivnik, R.J., Petersen, R.C., & Jack, C.R. Jr. (2008). The anatomic correlates of prosopagnosia in semantic dementia. *Neurology, 71(2)*, 1628-1633. doi: 10.1212/01.wnl.0000334756.18558.92

Kisilevsky, B.S., Hains, S.M.J., Brown, C.A., Lee, C.T., Cowperthwaite, B., Stutzman, S.S., Swansburg, M.L., Lee, K., Xie, X., Huang, H., Ye, H.-H., Zhang, K., & Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development, 32*, 59-71. doi: 10.1016/jinfbeh.2008.10.002

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Chichester: Wiley-Blackwell.

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

Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics,* 49,43-51.

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

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

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

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

Legge, G.E., Grossman, C., & Pieper, C.M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning Memory and Cognition, 10*, 298-303.doi: 10.1037/0278-7393.10.2.298

Levi, S. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language, 42(4)*, 843-872. doi: 10.1017/S0305000914000506

Liu, R.R., Corrow, S.L., Pancaroglu, R., Duchaine, B., & Barton, J.J.S. (2015). The processing of voice identity in developmental prosopagnosia. *Cortex, 71*, 390-397. doi: 10.1016/j.cortex.2015.07.030

Liu, R.R., Pancaroglu, R., Hills, C.S., Duchaine, B., & Barton, J.J.S. (2014). Voice recognition in face-blind patients. *Cerebral Cortex, 26*, 1473-1487. doi: 10.1093/cercor/bhu240

Mani, N., & Schneider, S. (2013). Speaker identity supports phonetic category learning. *Journal of Experimental Psychology: Human Perception and Performance, 39(3)*, 623-629. doi: 10.1037/a0030402

Mann, V.A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology, 27*, 153-165.

Mathiak, K., Menning, H., Hertrich, I., Matiak, K.A., Zvyagintsev, M., & Ackermann, H. (2007). Who is telling what from where? A functional magnetic resonance imaging study. *Neuroreport, 18(5)*, 405-409. doi: 10.1097/WMR/0b-13e328013cec4

Megreya, A.M., & Burton, A.M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34(4)*, 865-876.

Mervis, C.B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Mullenix, J.W., Ross, A., Smith, C., Kuykendall, D., Conard, J., & Barb, S. (2009). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology, 25(1)*, 29-34. doi: 10.1002/acp.1635

Mullenix, J.W., Stern, S.E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., May, E., & Tessmer, B. (2009). Earwitness memory: Distortions for voice pitch and speaking rate. *Applied Cognitive Psychology, 24(4)*, 513-526. doi: 10.1002/acp.1566

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., & Kojima, S. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia, 39*, 1047-1054.

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

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

Orchard, T., & Yarmey, A.D. (1995). The effects of whispers, voice sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9(3),* 249-260.

Panneton Cooper, R., Abraham, J., Berman, S., & Staska, M. (1997). The development of infants’ preference for motherese. *Infant Behavior and Development, 20(4),* 477-488.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85*, 913-925.

Perea, M., Jiménez, M., Suárez-Coalla, P., Fernández, N., Viña, C., & Cuetos, F. (2014). Ability for voice recognition is a marker for dyslexia in children. *Experimental Psychology, 61(6)*, 480-486. doi: 10.1027/1618-3169/a000265

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain, 117*, 1283-1301.

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

Perrachione, T.K., & Wong, P.C. (2007). Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex. *Neuropsychologia, 45,* 1899-1910.

Peterson, R.L., & Pennington, B.F. (2012). Developmental dyslexia. *Lancet, 379*, 1997-2007.

Phillipon, A.C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, threat, deliberate strategies and indirect measures. *Applied Cognitive Psychology, 21*, 539-550. doi: 10.1002/acp.1296

Pinheiro, A.P., Rezaii, N., Nestor, P.G., Rauber, A., Spencer, K.M., & Niznikiewicz, M. (2016). Did you or I say pretty, rude or brief? An ERP study of the effects of speaker’s identity on emotional word processing. *Brain and Language, 153,* 38-49. doi: 10.1016/j.bandl.2015.12.003

Rämä, P., & Courtney, S.M. (2005). Functional topography of working memory for face or voice identity. *NeuroImage, 24*, 224-234. doi: 10.1016/j.neuroimage.2004.08.024

Robertson, D.M.C., & Schweinberger, S.R. (2010). The role of audio-visual asynchrony in person recognition. *The Quarterly Journal of Experimental Psychology, 63(1)*, 23-30.

Schall, S., Kiebel, S.J., Maess, B., & von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage, 77*, 237-245.

Schall, S., Kiebel, S.J., Maess, B., & von Kriegstein, K. (2015). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience, 27(2)*, 280-291. doi: 10.1162.jocn\_a\_00707

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social, Cognitive and Affective Neuroscience, 2016,* 1812-1822. doi: 10.1093/scan/nsw089

Schelinski, S., Roswandowitz, C., & von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism Research, 10(1)*, 155-168. doi:10.1002/aur.1639.

Schweinberger, S.R., Herholz, A., & Stief, V. (1997). Auditory Long-term Memory: Repetition Priming of Voice Recognition. *Quarterly Journal of Experimental Psychology, 50A(3)*, 498-517.

Schweinberger, S.R., Kawahara, H., Simpson, A.P. Skuk, V.G., & Zäske, R. (2014). Speaker Perception. *WIREs Cognitive Science, 5*, 15-25. doi: 10.1002/wcs.1261

Schweinberger, S.R., Kloth, N., & Robertson, D.M.C. (2011). Hearing facial identities: Brain correlates of face-voice integration in person identification. *Cortex, 47(9)*, 1026-1037. doi: 10.1016/j.cortex.2010.11.011

Schweinberger, S.R., & Robertson, D.M.C. (2017). Audiovisual integration in familiar person recognition. *Visual Cognition,* doi: 10.1080/13506285.2016.1276110

Schweinberger, S.R., Robertson, D., & Kaufmann, J.M. (2007). Hearing facial identities: *Quarterly Journal of Experimental Psychology, 60*, 1446-1456.

Schweinberger, S.R., Walther, C., Zäske, R., & Kovács, G. (2011). Neural correlates of adaptation to voice identity. *British Journal of Psychology, 102*, 748-764.

Sergent, J. & Poncet, M. (1990). From covert to overt recognition of faces in a prosopagnosic patient. *Brain, 113*, 989-1004.

Sheffert, S.M., & Fowler, C.A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 34*, 665-685.

Sheffert, S.M., & Olson, E. (2004). Audio-visual speech facilitates voice learning. *Perception and Psychophysics, 66(2)*, 352-362.

Shilowich, B.E., & Biederman, I. (2016). An estimate of the prevalence of developmental phonagnosia. *Brain and Language, 159*, 84-91. doi: 10.1016/j.bandl.2016.05.004

Skuk, V.G., Palermo, R., Broemer, L., & Schweinberger, S.R. (2017). Autistic traits are linked to individual differences in familiar voice identification. *Journal of Autism and Developmental Disorders*. doi: 10.1007/s10803-017-3039-y

Skuk, V.G., & Schweinberger, S.R. (2013). Gender differences in familiar voice identification. *Hearing Research,* 296, 131-140.

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

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. doi: 10.1558/ijsll.v19i2.145

Stevenage, S.V., Hamlin, I., & Ford, B. (2016). Distinctiveness helps when matching static faces and voices. *Journal of Cognitive Psychology,* *29(3)*, 289-304. DOI: 10.1080/20445911.2016.1272605

Stevenage, S.V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness Recognition. *Applied Cognitive Psychology, 25(1)*, 112-118.

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

Stevenage, S.V., & Lewis, H.G. (2005). By which name should I call thee? The consequences of having multiple names. *Quarterly Journal of Experimental Psychology, 58A(8)*, 1447-1461. doi: 10.1080/02724980443000692

Stevenage, S.V., & Neil, G.J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54 (3),* 266-281. doi: <http://dx.doi.org/10.5334/pb.ar>

Stevenage, S.V., Neil, G.J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological Research, 77(2)*. 167-175. doi: 10.1007/s00426-012-0450-z

Stevenage. S.V., Neil, G.J., & Hamlin, I. (2014). When the face fits: Recognition of celebrities from matching and mismatching faces and voices. *Memory, 22(3),* 284-294. doi: http://dx.doi.org/10.1080/09658211.2013.781654

Tomlin, R., Stevenage, S.V., & Hammond, S. (2016). Putting the pieces together: Revealing face-voice integration through the facial overshadowing effect. *Visual Cognition,* doi: 10.1080/13506285.2016.1245230

Valentine, T., Hollis, J., & Moore, V. (1999). The nominal competitor effect: When one name is better than two. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 749–754). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

van Lancker, D.R., & Canter, G.J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1*, 185-195.

van Lancker, D.R., Cummings, J.L., Kreiman, J., & Dobkin, B.H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209.

van Lancker, D.R., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25*, 829-834.

van Lancker, D.R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11(5)*, 665-674. doi: 10.1080/01688638908400923

van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backwards voices. *Journal of Phonetics, 13*, 19-38.

Vitevitch, M.S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29(2)*, 333-342. doi: 10.1037/0096-1523.29.2.333

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*, 48-55. doi: 10.1016/S0926-6410(03)00079-X

von Kriegstein, K., & Giraud, A.L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948-955.

von Kriegstein, K., Kleinschmidt, A., & Giraud, A.L. (2006). Voice recognition and cross-modal responses to familiar speakers’ voices in prosopagnosia. *Cerebral Cortex, 16*, 1314-1322.

von Kriegstein, K., Kleinschmidt, A. Sterzer, P., & Giraud, A.L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*, 367-376.

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

Warren, J.D., Scott, S.K., Price, C.J., & Griffiths, T.D. (2006). Human brain mechanisms for the early analysis of voices. *Neuroimage, 31(3)*, 1389-1397. doi: 10.1016/j.neuroimage.2006.01.034

Yarmey, A.D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Sciences Society, 31(4)*, 421-428.

Yarmey, A.D. (1995). Earwitness speaker identification. *Psychology, Public Policy and Law, 1(4)*, 792-816.

Yarmey, A.D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology*, 6(5), 367-377.

Young, A.W., Hay, D.C., & Ellis, A.W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognising people. *British Journal of Psychology, 76,* 495-523.

Young, A.W., Hellawell, D., & De Haan, E.H.F. (1988). Cross domain semantic priming in normal subjects and a prosopagnosic patient. *Quarterly Journal of Experimental Psychology - A, 40*, 561-580.

Xu, X.K., Biederman, I., Shilowich, B.E., Herald, S.B., Amir, O., & Allen, N.E. (2015). Developmental phonagnosia: Neural correlates and a behavioral marker. *Brain and Language, 149*, 106-117.

Zäske, R., Mühl, C., & Schweinberger, S.R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLOSOne*, November 20th, doi: 10.1371/journal-pone.0143151

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

Zäske, R., Volberg, G., Kovács, G., & Schweinberger, S.R. (2014). Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience, 34(33),* 10821-10831. doi: 10.1523/jneurosci.0581-14.2014

Zhang, Z.J., Hao, G.R., Shi, J.B., Mou, X.D., Yao, Z.J., & Chen, N. (2008). Investigation of the neural substrates of voice recognition in Chinese schizophrenic patients with auditory verbal hallucinations: an event-related functional MRI study. *Acta Psychiatrica Scandinavica, 118*, 272-280.

Zweig, L.J., Suzuki, S., & Grabowecky, M. (2015). Learning face-voice pairings facilitate visual search. *Psychonomic Bulletin & Review, 22(2)*, 429-436. doi: 10.3758/s13423-014-0685-3

Table 1: Summary of neuroimaging studies on healthy individuals, arranged in alphabetical order within unfamiliar/familiar voice types.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors | Imaging | Familiarity of Stimuli | Task | Site of Activity |
| Belin et al. (2000) | fMRI | Unfamiliar | Passive listening to vocal vs non-vocal stimuli | TVA: Superior bank of right STS;  Posterior STG; Posterior to anterior STS |
| Belin et al. (2002) | fMRI | Unfamiliar | Passive listening to non-speech vocal sounds vs scrambled sounds | Anterior right STS |
| Imaizumi et al. (1997) | PET | Unfamiliar | Speaker discrimination | Bilateral temporal poles |
| Lattner et al. (2005) | fMRI | Unfamiliar | Naturalness decision | Right posterior STG |
| Mathiak et al. (2007) | fMRI | Unfamiliar | Speaker discrimination | Bilateral temporal poles |
| Nakamura et al. (2001) | PET | Friends/self vs unfamiliar | Familiarity decision | Left frontal pole and right temporal pole |
| Rämä & Courtney (2005) | fMRI | Unfamiliar | Encode and later recognise unfamiliar voices | Posterior STG, and posterior and middle STS, mainly in right hemisphere |
| Schall et al. (2015) | MEG | Unfamiliar | Voice recognition task | Right STS |
| Warren et al. (2006) | fMRI | Unfamiliar | Passive listening to unfamiliar voices | Posterior STS mainly in right hemisphere |
|  |  |  |  |  |
| Belin & Zatorre (2003) | fMRI | Learned | Adaptation study through passive listening | Right anterior STS |
| Bethmann & Brechmann (2014) | fMRI BOLD | Celebrity | Discrimination of vocal vs non-vocal sounds | Anterior Middle Temporal Gyrus (MTG) |
| Latinus et al. (2011) | fMRI BOLD | Learned | Adaptation study | Superior temporal pole when voices are unfamiliar. Anterior STS and convexity of inferior frontal cortices when voices are familiar. |
| Von Kriegstein et al. (2003) | fMRI | Learned | Voice recognition task | Anterior STS in right hemisphere |
| Von Kriegstein & Giraud (2004) | fMRI | Familiar vs unfamiliar | Voice recognition task | Anterior right STS (familiar and unfamiliar voices). Posterior right STS (unfamiliar voices) |
| Von Kriegstein et al. (2005) | fMRI | Personally familiar vs unfamiliar | Voice recognition task | Differential activation of bilateral anterior STS/STG and temporal poles |
| Zäske et al. (2014) | ERP | Learned | Voice Adaptation task | Suppressed beta-band oscillations (16-17Hz) at central and right temporal sites at 290-370ms post onset. |

Table 2: Summary of patient studies in alphabetical order detailing type of deficit and site of lesion where available.

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | Patient Details | Deficit | Site of Lesion |
| IDENTIFICATION OF VOICE-SPECIFIC DEFICITS | | | |
| Hailstone et al. (2010) | QR: Progressive phonagnosic | Voice recognition deficit with spared face recognition. | Right anterior temporal lobe, extending back to STS |
| Van Lancker & Canter (1982) | 30 right- or left-hemisphere patients | Face recognition deficit.  Voice recognition deficit.  Face and voice recognition deficits. | 1 right-hemisphere patient  1 right-hemisphere patient  3 right-hemisphere patients |
| DISSOCIATION OF FAMILIAR/UNFAMILIAR VOICE DEFICITS | | | |
| Garrido et al. (2009) | KH: Developmental phonagnosic | Familiar voice recognition deficit and impaired ability to learn new voices alongside spared unfamiliar voice discrimination and spared face recognition. | - |
| Herald et al. (2014) and Xu et al. (2015) | AN: Developmental phonagnosic | Familiar voice recognition deficit and familiar voice imagery deficit alongside spared unfamiliar voice discrimination and spared face recognition. | - |
| Neuner & Schweinberger (2000) | 3/36 rehabilitation patients  1/36 rehabilitation patient | Familiar voice recognition deficit with spared unfamiliar voice discrimination.  Deficit for both familiar and unfamiliar voice processing. | 1 left- and 2 right-hemisphere patients  1 right-hemisphere patient |
| Peretz et al. (1994) | GL: Amusic patient  CN: Amusic patient | Familiar voice recognition deficit alongside spared unfamiliar voice discrimination.  Familiar voice recognition deficit and unfamiliar voice discrimination deficit | Bilateral lesions  Bilateral lesions |
| van Lancker & Kreiman (1987) | 3/45 patients  7/45 patients | Celebrity voice recognition deficit alongside spared unfamiliar voice discrimination.  Unfamiliar voice discrimination deficit alongside spared celebrity recognition. | Right hemisphere lesions  Right hemisphere lesions |
| van Lancker et al. (1988) | 5/6 patients  1/6 patient | Dissociation of familiar and unfamiliar voice processing deficits.  Deficit for both familiar and unfamiliar voice processing. | Unfamiliar deficit – bilateral temporal lobes  Familiar deficit – right inferior parietal lobe, right temporo-parietal lobe |
| van Lancker et al. (1989) | Patient study | Dissociation of familiar and unfamiliar voice processing deficits. | Unfamiliar deficit – bilateral temporal lobes  Familiar deficit – right parietal lobe |