Understanding the mechanisms of familiar voice-identity recognition in the human brain

Corrina Maguinness[1]*, Claudia Roswandowitz[1,2] & Katharina von Kriegstein[1,3]*

1 Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
2 International Max Planck Research School on Neuroscience of Communication, Leipzig, Germany
3 Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

*Corresponding authors: *maguinness@cbs.mpg.de; *kriegstein@cbs.mpg.de
roswandowitz@cbs.mpg.de

Abstract

Humans have a remarkable skill for voice-identity recognition: most of us can remember many voices that surround us as 'unique'. In this review, we explore the computational and neural mechanisms which may support our ability to represent and recognise a unique voice-identity. We examine the functional architecture of voice-sensitive regions in the superior temporal gyrus/sulcus and bring together findings on how these regions may interact with each other, and additional face-sensitive regions, to support voice-identity processing. We also contrast findings from studies on neurotypicals and clinical populations which have examined the processing of familiar and unfamiliar voices. Taken together, the findings suggest that representations of familiar and unfamiliar voices might dissociate in the human brain. Such an observation does not fit well with current models for voice-identity processing, which by-and-large assume a common sequential analysis of the incoming voice signal, regardless of voice familiarity. We provide a revised audio-visual integrative model of voice-identity processing which brings together traditional and prototype models of identity processing. This revised model includes a mechanism of how voice-identity representations are established and provides a novel framework for understanding and examining the potential differences in familiar and unfamiliar voice processing in the human brain.

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

1. Introduction

The human voice offers an array of social information. For example, while we listen to someone speak, we can deduce the content of the speech message, the manner in which it is spoken, e.g. with humour or sadness, and the identity of the person who is speaking. In this review, we concentrate on the remarkable ability of voice-identity recognition[1]. Although it may appear effortless, recognising the identity of a voice is a significant feat for the perceptual and cognitive system. Each voice we encounter holds the same perceptual features as the next, acoustical properties that are determined by the glottal folds and the vocal tract (Lavner et al., 2001, López et al., 2013). Thus, to represent a voice-identity in memory, the brain is tasked with extracting *and* representing often subtle differences in features across individuals (Belin et al., 2011). Furthermore, successful voice-identity recognition also involves attributing meaning to the voice, e.g. retrieving semantic knowledge about the person's identity. In the pages that follow, we examine the computational and neural processes which underpin our ability to recognise voices.

2. Voice-identity processing in the neurotypical brain

2.1. Current models of voice-identity processing

2.1.1. Prototype model of voice-identity processing

Humans have developed a remarkable expertise for representing the often subtle variations in voice properties across individuals: most of us can remember many of the voices which surround us as *unique* (see Stevenage et al., 2012, Pernet et al., 2015, Goggin et al., 1991, Mullennix et al., 2011, Perrachione et al., 2011 for factors which influence voice memorability). How might the brain accomplish this feat? One potential solution is that we might have different neurons which each dedicatedly represent, and respond only to, one unique identity (Cutzu and Edelman, 1996, Quiroga et al., 2005). However, while individual identities have been shown to elicit unique neural response patterns in the brain (Kriegeskorte et al., 2007, Natu et al., 2010, Nestor et al., 2011, Vida et al., 2016), evidence for such a discrete one-to-one neuron-to-identity mapping is lacking (Quiroga, 2017, Chang and Tsao, 2017, see Petkov and Vuong, 2013 for overview). A potentially more parsimonious explanation has been proposed by a prototype model of voice-identity processing (Lavner et al., 2001, see Rosch, 1973 for original conception of prototype coding). Under this model (Figure 1), it is proposed that each voice we encounter is represented in the brain in a multidimensional acoustical 'voice-space' (Latinus and Belin,

[1] We use the term 'voice-identity recognition' to explicitly refer to the processes of *recognising* the identity of familiar individuals by voice. We use 'voice-identity processing' as a more global term which encompasses perceptual, e.g. voice discrimination, and recognition aspects.

2011, Latinus et al., 2013, Andics et al., 2010, Petkov and Vuong, 2013). The dimensions of the space encompass the range of perceptual features which are used to identify voices. These features include, but are not limited to, the acoustic effects of the vocal tract length (VTL) and the glottal-pulse rate (GPR) (for review see Mathias and von Kriegstein (2014)[2]). At the centre of this voice-space lies the *prototype* voice. This prototype is likely built up through our prior exposure with vocal identities. It may be considered to be an average approximation of the voices we encounter or simply a "very common voice" (Lavner et al., 2001). It is argued that this prototype voice functions as a perceptual 'norm' by which other voices are represented (Latinus et al., 2013, Lavner et al., 2001). Note that prototype processing is also referred to as 'norm-based coding' in the literature (e.g. see Yovel and Belin, 2013, Latinus et al., 2013).

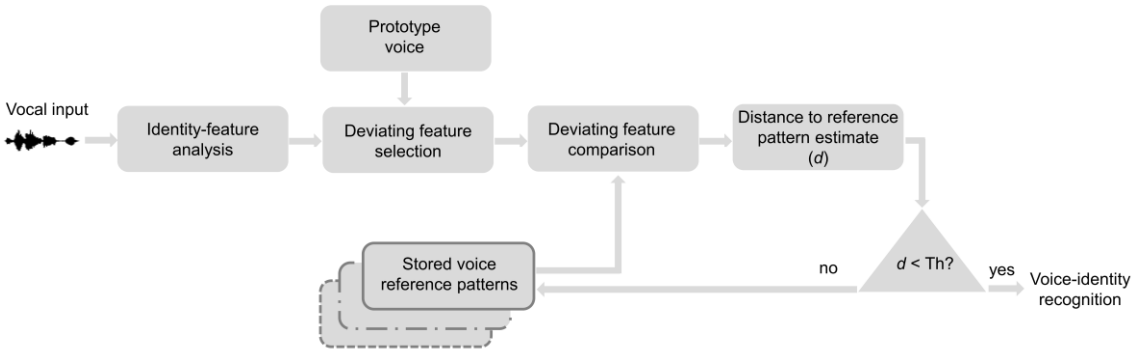


Figure 1. **A schematic representation of the prototype model of voice-identity processing** - adapted from Lavner and colleagues' original conception of the model (Lavner et al. 2001). As a first step, voice-identity features are extracted. These features are contrasted against a prototype, i.e. average voice. Features which deviate from this prototype are selected. The deviating features are then compared to stored voice reference patterns. The different outlines for the reference patterns indicate that they are likely relatively unique for each vocal identity. An estimate of the match between the selected deviating features and the stored reference patterns is then computed (i.e. *d,* distance to reference pattern). If this distance between the incoming voice and the reference pattern is lower than some perceptual threshold (Th), the analysed voice is deemed as belonging to that stored reference pattern. At this stage the voice is recognised as previously encountered i.e. the voice is recognised as familiar.

[2] Although GPR and VTL are important features for recognition, other available voice features can also support identity processing (Lavner et al., 2001, Sheffert et al., 2002, Remez et al., 1997). Listeners can use a range of these cues; this may be one of the reasons why we can recognise identity so robustly across different listening conditions when particular voice features may be altered or unavailable (e.g. Van Lancker et al., 1985a,b, Sheffert et al., 2002, Remez et al., 1997). In addition, there is evidence that listeners sensitivity to these various features can vary substantially (see e.g. Lavner et al., 2001) and that certain features may be more important for recognising particular voice-identities than others (e.g. Van Lancker et al., 1985a,b).

Thus, rather than representing each voice-identity as an absolute value in voice-space, the perceptual system may represent a unique voice-identity in terms of its *deviation* from the prototype voice (Lavner et al., 2001). These deviations may be stored as unique 'reference patterns' for each identity. The more an encountered voice deviates from this central voice, the easier it is to identify the unique reference pattern matching with that voice-identity (Lavner et al., 2001). Reference patterns likely become more robust with repeated exposure to the voice-identity. Potentially we may acquire multiple prototypes, for example there is evidence that male and female voices may be represented in terms of their *sex-specific* prototype voice (Latinus et al., 2013). Notably, the principles of the prototype model have also been successfully applied in the visual domain; with substantial evidence that face-identity is represented in this prototype manner (Leopold et al., 2005, Leopold et al., 2001, Newell et al., 1999, Rhodes and Jaquet, 2011).

Support for the prototype model of voice-identity processing comes from several complementary sources (Lavner et al., 2001, Fontaine et al., 2017, Stevenage et al., 2012, Barsics and Brédart, 2012, Mullennix et al., 2011, Latinus et al., 2013, Latinus and Belin, 2011). Here we briefly discuss evidence from behavioural investigations, before turning to the associated neural findings in Section 2.2. On a behavioural level, it has been shown that voices rated as more distinctive are more robustly recognised than their more average counterparts (e.g. Barsics and Brédart, 2012, Sørensen, 2012, Mullennix et al., 2011, for recent review see Stevenage and Neil, 2014). For example, Mullennix et al. (2011) presented listeners with previously learned average (labelled in the study as 'high-typical') or distinctive ('low-typical') voices, among a series of distractor voices which were matched for averageness/distinctiveness level to the learned voices. The authors noted that listeners showed higher false-alarm rates when recognising previously-learned average, in comparison to distinctive voices. In other words, listeners were likely to erroneously recognise *newly*-presented average voices as previously learned. This finding fits well with a prototype model of voice-identity processing. As average voices (i.e. close to prototype) are represented in the clustered centre of voice-space, confusion with similar voices is likely. In contrast, voices which deviate from typicality, i.e. distinctive voices, are represented further away from the centre in the more outer extremities of voice-space. Consequently, these voices share fewer "near neighbours" and compete less in memory with other voice-reference patterns (Stevenage and Neil, 2014, Lavner et al., 2001). Furthermore, voice samples which emphasise speaker-specific deviations, i.e. vocal caricatures, are more readily recognised as belonging to an identity than more veridical voice samples (López et al., 2013). Additional compelling evidence for prototype processing of voices also comes from voice-adaptation experiments, which have reported perceptual aftereffects which are consistent with a prototype-based processing of voice-identity (Latinus and Belin, 2011).

2.1.2. Traditional models of voice-identity processing

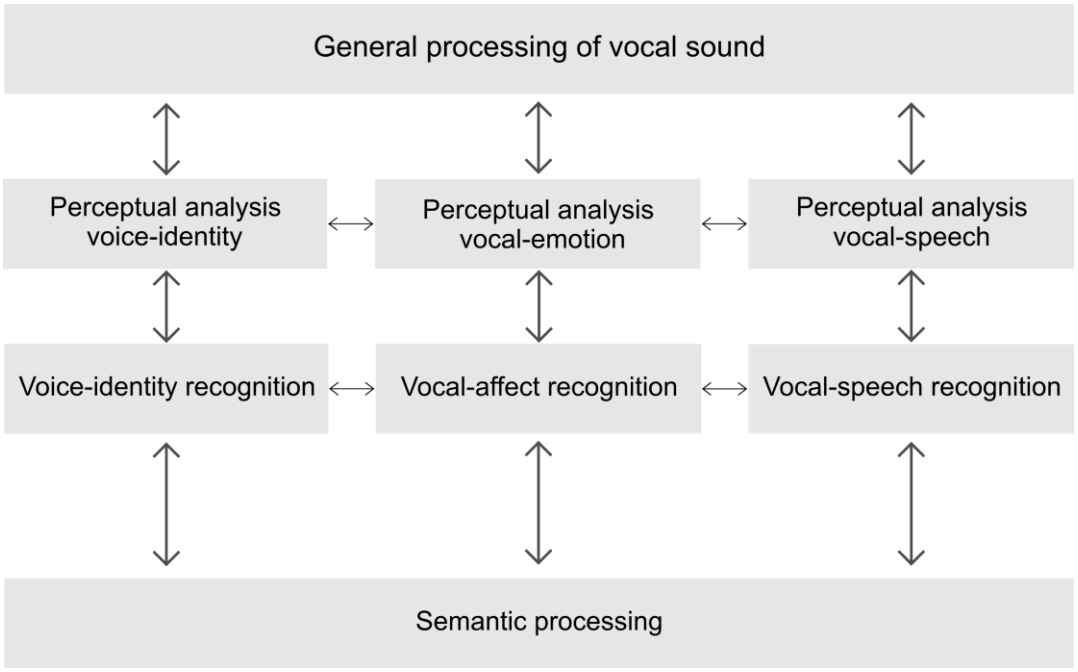


Figure 2. **A model of voice-identity processing** - originally based on a seminal model of face-identity processing proposed by Bruce and Young (1986). This model is adapted from several previous versions outlined by Ellis et al. (1997), Belin et al. (2004), Blank et al. (2014a), Neuner and Schweinberger (2000), and Roswandowitz et al. (2018a).

Traditional models of voice-identity processing conceive voice-identity processing as a multistage *sequential* process (Ellis et al., 1997, Belin et al., 2004, Neuner and Schweinberger, 2000, Roswandowitz et al., 2017, Roswandowitz et al., 2018a, for review see Blank et al., 2014b) (Figure 2, left of figure). In the model above, the vertical arrows (Figure 2, left of figure) denote the major processing pathways which are necessary for voice-identity recognition to take place. Note that while this *adapted* model assumes a sequential analysis of the voice signal, interactions between all stages of processing are also proposed (denoted by bidirectional arrows; see also Belin et al., 2004).

Voice-identity processing is assumed to involve an initial stage of 'structural encoding or structural analysis' of the voice, a process which is encompassed in the perceptual analysis of voice-individuating features i.e. *perceptual analysis voice-identity* (Figure 2, left of figure). It must necessarily include computations which extract the more stable features of the voice, such as the mean fundamental frequency, i.e. GPR, from the dynamic speech signal. Other features of the voice, which support speech and vocal-emotion processing are also analysed at the perceptual level but are argued to be

processed in partially distinct but interacting systems (Kreitewolf et al., 2014, von Kriegstein et al., 2010). Here we denote these cross-system interactions with horizontal bidirectional arrows. Note that previous models have assumed a *general* stage of 'structural analysis' which is common to identity, speech, and emotion processing (see e.g. Belin et al., 2004, Bruce & Young, 1986). However, given that identity, affect, and speech recognition rely on different features of the vocal signal, and that each process can be selectively impaired (see e.g. Roswandowitz et al., 2017, Luzzi et al., 2017, Lang et al., 2009, Roswandowitz et al., 2014), we have here separated these processes already at this initial stage.

At the next stage, traditional models argue that *voice-identity recognition* (Figure 2, left of figure) may be achieved via contrasting the incoming encoded voice percept with voice representations stored in 'voice recognition units' (Ellis et al., 1997). Further *semantic processing* (Figure 2, left of figure) of the voice e.g. retrieving the occupation of the speaker is accomplished via access to a multimodal person identity node (PIN) (Bruce and Young, 1986, Ellis et al., 1997), which either stores semantic information or is just a connection node (Burton et al., 1990). This PIN may be accessed via the auditory and visual system (Bruce and Young, 1986, Belin et al., 2004, Ellis et al., 1997, Neuner and Schweinberger, 2000).

2.2. Neural mechanisms of voice-identity processing

2.2.1. Voice-sensitive regions in the superior temporal gyrus/sulcus

In an extension of a pioneering study by Belin et al. (2000), Pernet et al. (2015) used functional magnetic resonance imaging (fMRI) to examine brain responses in a large group of participants (n = 218), as they passively listened to a series of vocal (speech and non-speech, e.g. laughs, coughs, utterances) or non-vocal (e.g. sounds from animals or musical instruments) sounds. The contrast between the two listening conditions, vocal versus non-vocal, revealed responses along the bilateral temporal lobes with a focus on the superior temporal gyrus/sulcus (STG/S), which were robustly seen for most participants (94%). These responses were clustered around three peaks in the posterior, middle, and anterior STG/S.

Increased responses to voices in STG/S have been consistently observed in studies which have included a wide range of control auditory stimuli and task designs (Belin et al., 2000, Belin et al., 2004, Fecteau et al., 2004, Belin et al., 2002, von Kriegstein et al., 2003, von Kriegstein and Giraud, 2004). These regions are commonly referred to as the Temporal Voice Areas (TVAs; von Kriegstein and Giraud, 2006, Campanella and Belin, 2007) and are often referred to as being not only voice-sensitive (Roswandowitz et al.,

2017, von Kriegstein and Giraud, 2004, Blank et al., 2014b, Schelinski et al., 2017), but even selective for voices (Belin et al., 2000, Belin et al., 2004, Belin et al., 2002).

2.2.2. What regions may support voice-identity processing?

It has been suggested that there is a *core-voice system* (Roswandowitz et al., 2017, Roswandowitz et al., 2018a, Roswandowitz et al., 2018b) which may function in an interactive manner to support voice-identity processing (von Kriegstein and Giraud, 2004). This system includes auditory processing regions of Heschl's gyrus (HG) and planum temporale (PT) (Warren et al., 2006, Kreitewolf et al., 2014, von Kriegstein et al., 2007, von Kriegstein et al., 2006b, Formisano et al., 2008, von Kriegstein et al., 2010), as well as TVAs in the posterior, mid, and anterior STG/S, particularly in the right hemisphere (von Kriegstein and Giraud, 2004, Belin and Zatorre, 2003, Pernet et al., 2015, Warren et al., 2006) and probably also parts of the middle temporal gyrus/sulcus (MTG/S) (von Kriegstein et al., 2003, Roswandowitz et al., 2017, Warren et al., 2006). Voice-sensitive regions along the STG/S share both functional (von Kriegstein and Giraud, 2004) and structural (Blank et al., 2011) connections with one other. The amount of responses in these regions during voice-identity recognition correlates with listeners' voice-identity recognition abilities (Schelinski et al., 2016, Schall et al., 2015).

Regions within the core-voice system are thought to serve potentially different functional roles in supporting voice-identity processing (see Table 1 for summary). The HG, PT, and the posterior STG/S (pSTG/S) have been implicated in acoustical voice-identity processing (Kreitewolf et al., 2014, von Kriegstein et al., 2007, von Kriegstein et al., 2006b, Warren et al., 2006, von Kriegstein et al., 2010). There is evidence that the analysis of acoustical voice-identity features is accomplished in HG for vocal pitch (i.e. GPR) (von Kriegstein et al., 2010) and in the pSTG/S and PT for vocal timbre (i.e. VTL) (von Kriegstein et al., 2007, von Kriegstein et al., 2006b, von Kriegstein et al., 2010). For example, von Kriegstein et al. (2010) observed that anterolateral HG was more responsive when listeners heard sequences of syllables in which speaker identity varied based on pitch cues (i.e. GPR varied), compared to when speaker identity varied based on timbre cues (i.e. VTL varied). Increased responses in the pSTG/S have been observed when listeners were exposed to sequences of syllables in which the VTL of the speaker varied, compared to sequences where the VTL was fixed. This pSTG/S sensitivity to variations in VTL appeared to be relatively human voice specific, being strongest for variations in human VTL, compared to similar variations in non-human voice stimuli such as the VTL of frogs or the acoustic scale of musical instruments (von Kriegstein et al., 2007). Warren et al. (2006) also noted involvement of the PT in acoustical voice-identity analysis, observing sensitivity in this region (as well as pSTG/S) to changes in speaker identity *and* changes in spectrotemporal properties which alter the saliency of voice-identity cues in the speech signal.

Table 1. Summary of fMRI studies reviewed in the current paper which focuses on responses in regions of the STG/S during unfamiliar and familiar voice-identity processing, in neurotypicals participants.

Voice Stimuli		Paradigm	Superior temporal gyrus/sulcus responses						Reference
			Right Hemisphere			Left Hemisphere			
Unfamiliar	*Familiar*		*pSTG/S*	*mSTG/S*	*aSTG/S*	*pSTG/S*	*mSTG/S*	*aSTG/S*	
x		Human vocal vs. non-vocal sounds [passive listening]	**x**	**x**	**x**	**x**	**x**	**x**	Belin et al. 2000; Pernet et al. 2015
x		Human vocal (VTL varied vs. fixed) vs. non-human vocal and non-vocal (VTL varied vs. fixed) sounds [listening and end of block detection]	**x**			**x**			von Kriegstein et al. 2007; von Kriegstein et al., 2010
x		Human VTL varied vs. human VTL fixed [Exp 1: speech and loudness task]	**x**			**x**			von Kriegstein et al. 2010
		Human VTL varied vs. human GPR varied [Exp 2: speech and speaker identity task]	n.s. (trend)			**x**			
x		Human vocal non-speech vs. frequency scrambled human vocal non-speech [passive listening]		**x**	**x**				Belin et al. 2002
x		Same voice ID, different syllable vs. different voice ID, same syllable [passive listening: adaptation design]			**x**				Belin and Zatorre 2003
x		Voice further vs. closer to prototype voice							Latinus et al. 2013
		[Exp 1 (manipulated voices): pure tone detection]		**x**			**x**		
		[Exp 2 (natural voices): pure tone detection]	**x**		**x**	**x**			
		[Exp 3 (manipulated voices): passive listening]		**x**			**x**		
x		Speaker vs. speech recognition [briefly-heard unfamiliar voices]			**x**				von Kriegstein et al. 2003
x	**x**	Speaker vs. speech recognition [personally familiar voices and briefly-heard unfamiliar voices]	**x**		**x**	**x**	**x**	**x**	von Kriegstein and Giraud 2004
		Speaker vs. speech recognition [briefly-heard unfamiliar voices vs. personally familiar voices]	**x**						
x	**x**	Personally familiar voices vs. unfamiliar voices [familiar/unfamiliar and left/right auditory field judgement]						**x**	Birkett et al. 2007
x	**x**	Famous vs. unfamiliar voices [Exp 1: familiar/unfamiliar judgement and name retrieval; Exp 2: familiar/unfamiliar judgement and name/semantic info. retrieval]	**x**		**x**	**x**		**x**	Bethmann et al. 2012
x	**x**	Unfamiliar voices/correct 'new' responses vs. recently-familiarised voices/correct 'old' responses [old/new judgement]	**x**						Zäske et al. 2017

The pSTG/S may play a particular role in the processing of unfamiliar voices, i.e. voices that have never been heard before or have only been recently encountered. In von Kriegstein and Giraud (2004) we examined fMRI responses while participants performed a speaker-, i.e. voice-identity, recognition task or a control speech-recognition task on the same auditory stimuli. The auditory stimuli were voices of speakers who were either personally known to the listeners, i.e. familiar acquaintances, or unfamiliar voices[3]. Participants briefly listened to the unfamiliar voices before fMRI data acquisition. We noted increased responses with a maximum in the right pSTG/S to unfamiliar, in comparison to familiar voices during the voice-identity recognition task. This same increased response profile was not observed for the control speech-recognition task (i.e. there was a task x familiarity interaction). Zäske et al. (2017) experimentally familiarised listeners with a series of voice-identities prior to scanning. Following learning, participants identified whether presented voices were from the recently-familiarised voice set or were novel, i.e. unfamiliar. Zäske and colleagues observed increased responses in the pSTG/S to voices which were correctly classified as unfamiliar, compared to those which were correctly classified as familiar (however see also Birkett et al. (2007)). Arguably, responses which have been observed in the pSTG/S may reflect the increased acoustical feature analysis which might be required when processing unfamiliar voices (Kreiman and Sidtis, 2011, Sidtis and Kreiman, 2012). Such an interpretation is further corroborated by evidence that people with autism spectrum disorder (ASD), who have difficulties with voice-identity processing at a perceptual level (Schelinski et al., 2017, Lin et al., 2016), also show atypical responses in the right pSTG/S when the task is to recognise voice-identity (Schelinski et al., 2016, see Stevenage, 2018 for an overview of voice processing in other clinical populations).

[3] We define *familiar voices* as voices of speakers who the participant has had significant prolonged exposure to either through social interactions (personal acquaintances) or media exposure (famous speakers). Usually other attributes of the speaker are known as well (e.g. the face or semantic attributes such as 'BBC Woman's Hour speaker'). In theory, familiarity with a speaker that is equitable to personally known/famous voices might also be induced by an extensive experimental training protocol. However, most current training paradigms are not equable to voice-identity learning in the natural environment. We therefore refer to voices learned as part of an experimental training protocol as *recently-familiarised voices.* We assume that a process of refining a reference pattern for such voices is still in process. In contrast, we define *unfamiliar voices* as voices that are unknown to the participant. In the case of unfamiliar voices, usually other attributes of the speaker (e.g. their face or name) are also unknown to the participant. In theory, however, a voice of a famous person whom one has encountered only via the visual modality can be unfamiliar as well. Unfamiliar voices can be novel, i.e. heard for the first time by participants during experimental testing (unfamiliar voices), or the participant has had limited exposure to them prior to testing, without an explicit training protocol (briefly-heard unfamiliar voices).

A later stage of voice-identity processing, i.e. voice-identity recognition, is assumed to be supported by the mid to anterior regions of the STG/S, i.e. mSTG/S and aSTG/S, respectively (Belin and Zatorre, 2003, Belin et al., 2002, von Kriegstein et al., 2003, Andics et al., 2010). Studies which have explicitly contrasted responses in the STG/S during voice-identity, in comparison to speech recognition, have typically observed not only the pSTG/S responses, described in the preceding paragraph, but also increased response in the anterior regions of STG/S and these aSTG/S responses seem to be particularly driven by the task of voice identification (von Kriegstein et al., 2003, von Kriegstein and Giraud, 2004). Responses in mSTG/S regions have also been reported for contrasts of voice-identity recognition tasks against speech-recognition tasks (see e.g. von Kriegstein and Giraud, 2004, Blank et al., 2011).

Converging evidence from several studies provides compelling evidence that the aSTG/S may play a prominent role in representing *unique* voice-identities (Belin and Zatorre, 2003, Schelinski et al., 2016, Schall et al., 2015, Formisano et al., 2008, Andics et al., 2010, Bethmann et al., 2012). For example, Belin and Zatorre (2003) noted that the aSTG/S was sensitive to the identity of a voice across changes in speech utterances. There were reduced responses in the aSTG/S, i.e. adaptation, when the same speaker was heard uttering different syllables. In contrast, similar adaptation effects were not observed in this region when the same syllable was uttered by different speakers. The presented voices were unfamiliar to the listeners, at least prior to the fMRI task, and participants were exposed to multiple utterances of the same speakers during scanning. Responses in the aSTG/S have also been observed during voice-identity tasks in contrast to speech tasks on the same stimulus material (von Kriegstein and Giraud, 2004, von Kriegstein et al., 2003). This was the case both for unfamiliar voices which the listener had briefly-heard prior to testing and personally familiar voices (von Kriegstein and Giraud 2004). The two studies also indicated that while the pSTG/S is responsive to spectrotemporally complex sounds independent of task, this is not the case for the aSTG/S (von Kriegstein and Giraud, 2004, von Kriegstein et al., 2003).

Bethmann et al. (2012) found that subjectively familiar voices (i.e. famous voices which the listeners identified as familiar to them upon listening, independent of whether they were famous or not) elicited stronger responses in the anterior regions of the temporal lobe, in contrast to subjectively unfamiliar voices (i.e. voices, including famous voices, which the listener identified as unfamiliar to them). In addition, the authors also observed increased recruitment of posterior regions of the STG/S for subjectively familiar voices. The authors argue that subjectively familiar voices may modulate responses in this region in a top-down manner. This result is contrary to the findings of von Kriegstein and Giraud (2004). von Kriegstein and Giraud (2004) argued that the pSTG/S may be more involved in the perceptual analysis of voices, because they showed increased responses in this region during the recognition of briefly-heard unfamiliar, compared to familiar voice-

identities (see Zäske et al. 2017 for similar findings). This discrepancy might be dependent on the different task designs used. While in von Kriegstein and Giraud (2004) the task was matched for the familiar and unfamiliar voices in complexity (i.e. detect target voice amongst distractors), this was not the case for the Bethmann study. Here participants were asked to judge whether the voice was from a familiar speaker or not and in case of familiarity retrieve the name or other semantic information. This means that to perform the task, participants had to identify the particular voice-identity only for the subjectively familiar voices, but not for the subjectively unfamiliar voices.

Could the mSTG/S play a potentially facilitative role in connecting the perceptual analysis of voices in pSTG/S, with representations of unique voice-identities which may be housed in the anterior STG/S? There is some evidence to suggest this may be so. The mSTG/S shares functional connections with anterior and posterior regions of the STG/S during the processing of voice-identity (von Kriegstein and Giraud, 2004) and there are also direct structural connections between these regions (Blank et al., 2011). In addition, the region might serve an intermediate computational step between perceptual analysis and identity recognition: In a recent study, Latinus and colleagues presented listeners with a series of unfamiliar voice samples which were altered to deviate in a linear fashion from a prototype voice (Latinus et al., 2013). Prototype voices were constructed by averaging many same sex voices, in a 3-dimensional voice-space which included voice-identity features of GPR (i.e. fundamental frequency), formant dispersion (i.e. the average frequency difference between formants, related to VTL), and harmonics-to-noise ratio (a measure of spectrotemporal regularity). The authors observed that voices which were manipulated to deviate more from their sex-specific prototype induced higher responses in the mSTG/S than those which were closer to their prototype (Experiment 1 and 3). In accordance with a prototype model of voice-identity processing, these deviating voices were also rated as more distinctive by listeners (Experiment 1). This might suggest that the mSTG/S may be engaged in a pivotal intermediate computational step between the processing of unique acoustical features of the voice in more posterior regions of the temporal lobe and potential identity recognition in the aSTG/S. Though, it is important to note that the mSTG/S did not appear to be sensitive to the degree to which natural, i.e. not manipulated, voices deviated from the prototype voice. Rather, listening to voices which naturally deviated more from their sex specific prototype (calculated by the authors via distance-to-mean to the sex-specific prototype voice of the stimulus set) induced increased responses in posterior and anterior regions of the STG/S (Experiment 2). Why this discrepancy between the different experiments of that study occurred is unclear. One tentative explanation could be that while the natural voices might have been far away from the prototype in terms of acoustic features, they may not have been subjectively perceived as such. Listeners in that experiment did not subjectively rate the distinctiveness of the natural voices. Therefore, it is still an open question as to whether

there is a relationship between the *perceived* distinctiveness of naturally occurring voices and responses in the voice-sensitive mSTG/S.

2.2.3. The time course of voice-identity processing

Electroencephalography (EEG) and magnetoencephalography (MEG) studies provide critical evidence concerning the *time-courses* of responses in the brain during voice-identity processing. Studies which have examined time-sensitive neural responses to voices, have reported evidence which may potentially support sequential stages of voice-identity processing (see Figure 1 and 2). For example, early voice-sensitive responses have been observed during the processing of vocal versus non-vocal sounds (~164 ms see Charest et al., 2009, ~150 ms see Capilla et al., 2013). These early responses have been localised to mid TVAs (Capilla et al., 2013). In contrast, familiarity processing (~200 ms Beauchemin et al., 2006, ~250 ms Schweinberger et al., 2011b) and the recognition of identity across novel speech utterances, (~290-370 ms Zäske et al., 2014) has been associated with later responses. The location of these later responses have been attributed to temporal electrode sites, however a more exact source localisation is lacking (Zäske et al., 2014, Schweinberger, 2001, Schweinberger et al., 2011b). One study, reported by Schall et al. (2015), has examined the time course of responses in discrete voice-sensitive regions (i.e. along posterior, mid, and anterior regions of the STG/S). Schall et al. (2015) familiarised participants via a series of training rounds with the voices and faces/occupation information of six unfamiliar male speakers. Following training, listeners were asked to recognise the identity of the recently-familiarised voices, or to recognise the content of their speech utterances, i.e. a similar design to the previously reported fMRI studies (von Kriegstein and Giraud, 2004, von Kriegstein et al., 2003). Schall et al. (2015) observed that participants who were more accurate at recognising voice-identities, also showed comparably higher aSTG/S activity at around 200 ms after voice onset; during the voice, compared to the speech recognition task. Interestingly, at this same time point the authors also noted activity in the pSTG/S, suggesting that responses in these regions may follow a similar time course. However, pSTG/S activity for the voice, in comparison to the speech, recognition task did not correlate with behavioural performance.

2.2.4. Voice-identity processing: Interactions with the extended system

The core-voice system is assumed to connect with an extended system. Potential brain candidates for the extended system include supra-modal regions encompassing the precuneus/posterior cingulate, amygdala, inferior frontal gyrus and aTL, including the temporal pole (Shah et al., 2001, Gainotti, 2015, von Kriegstein and Giraud, 2006, Andics et al., 2010, for review see Blank et al., 2014b). Reponses in these regions are often observed during voice processing. For example, when contrasting brain responses for vocal sounds versus non-vocal sounds, Pernet and colleagues observed that besides the TVAs, extra-

temporal regions including the bilateral inferior prefrontal cortex and amygdalae showed voice-sensitivity (Pernet et al., 2015). The voice area localiser used in Pernet et al. also includes voices with emotional prosody, so it is unclear what aspects of voices lead to responses in these extra-temporal regions. von Kriegstein et al. (2005) also noted that regions of the extended system were particularly engaged during the processing of familiar (personal acquaintances), in comparison with unfamiliar, voice identities, observing increased responses in the bilateral temporo-occipito-parietal, medial parietal/retrosplenial and anterior inferior temporal regions.

It is proposed that access to the extended system allows for a processing of the voice beyond a feeling of familiarity, allowing for further 'meaning' to be attributed to the voice. This *semantic processing* (Ellis et al., 1997, Roswandowitz et al., 2018a, Bruce and Young, 1986, Neuner and Schweinberger, 2000) can involve, for instance, the recall of multi-modal information characterising the person's identity such as the person's occupation or deducing where the voice was previously encountered. In traditional person-identity recognition models these processes are attributed to the PIN (see Section 2.1.2.). The existence of such a PIN remains debatable, potential candidates may include the aTL (for recent review see Blank et al., 2014b). Other roles of the extended system can be for example to evaluate one's relationship or feelings towards the identity (for a review on the extended system in face processing see Haxby et al., 2000).

3. Voice-identity processing deficits: Phonagnosia

3.1. Acquired phonagnosia: Behavioural evidence

Our understanding of the neural processes involved in voice-identity processing has also been informed through the study of phonagnosia. Phonagnosia refers to a deficit in processing identity information from the voice alone (Van Lancker and Canter, 1982). Assal et al. (1976) reported first evidence for the existence of the disorder following brain insult. The authors reported several patients with brain lesions who performed significantly worse than healthy controls when discriminating between unfamiliar voices (i.e. apperceptive phonagnosia). In 1987, Van Lancker and Kreiman were the first to show that the recognition of familiar voices can be impaired in brain-lesion patients (i.e. associative phonagnosia).

One might assume that the ability to recognise a voice as familiar is grounded on the more fundamental ability to distinguish between unfamiliar voice-identities. Interestingly, dissociation between voice-identity processing for unfamiliar and familiar voices has been reported in the literature (Assal et al., 1981, Van Lancker and Kreiman, 1987, Van Lancker et al., 1988, Van Lancker et al., 1989). Unfamiliar voice-identity

processing in the lesion literature has been typically assessed using voice-discrimination tasks. Here, listeners are presented with two sentences from unfamiliar speakers they have not previously encountered and are asked to judge whether the sentences are spoken by the same speaker, or not. Familiar voice-identity recognition, on the other hand, has been examined with both famous and personally familiar voices. In such tasks, participants are asked to judge the familiarity of the voice and/or to associate semantic information to the voice-identity e.g. recalling the name or associated face-identity (see Roswandowitz et al., 2018a for detailed overview of the test designs used in the study of phonagnosia). Group and case studies have observed evidence for brain-lesioned patients with impaired unfamiliar voice-identity discrimination and intact familiar voice-identity recognition and vice versa (Van Lancker and Kreiman, 1987; Van Lancker et al., 1988; Van Lancker et al., 1989, Luzzi et al. 2017). Such findings potentially suggest that the cognitive mechanisms underlying familiar and unfamiliar voice processing may be somewhat distinct (see Section 5. for further discussion). This might be because the representations of the voices with different familiarity are distinct or/and because the recognition of unfamiliar and familiar voices may require different task-related mechanisms. It is important to consider the differences in the tasks used to investigate unfamiliar and familiar voice-identity processing (i.e. discrimination tasks for unfamiliar voices and familiarity judgement and/or semantic association for familiar voices). If familiar and unfamiliar voice-identity processing reflect truly dissociable processes, one would predict differences in their processing even if they were tested with the same task design (e.g. see fMRI study on neurotypical participants in von Kriegstein et al. 2004, Section 2.2.2).

3.2. Acquired phonagnosia: Lesion location and behavioural specificity

Which brain structures have been associated with impaired voice-identity processing? The first studies on acquired phonagnosia provided no information about precise lesion locations. Rather they indicated that right, in contrast to left, hemispheric lesions can be associated with a deficit in unfamiliar voice-identity discrimination (Assal et al., 1976, Assal et al., 1981) and familiar voice-identity recognition (Van Lancker and Canter, 1982). Later studies provided evidence for an association between temporal as well as parietal lobe lesions and impaired voice-identity processing (Van Lancker et al., 1988, Roswandowitz et al., 2018b, Van Lancker et al., 1989).

Lesions in the bilateral temporal lobe (i.e. including anterior, mid, and posterior regions) have been associated with impaired discrimination abilities of unfamiliar voices (Van Lancker et al., 1988, Van Lancker et al., 1989). In contrast, lesions in the right inferior parietal lobe have been linked to impaired familiar voice-identity recognition (i.e. associating a face and name to a voice) (Van Lancker et al., 1988; Van Lancker et al., 1989; for summary of lesion locations in phonagnosia see Roswandowitz et al. 2018a). Van Lancker et al. (1989) observed that in 16 patients with impaired familiar voice-identity

recognition and available CT scans, 9 of these patients had a right parietal lobe lesion. However, for the additional 7 patients the lesion location was unreported. This finding of the importance of the right inferior parietal lobe for voice-identity recognition is surprising given that the neuroimaging literature has typically emphasised the involvement of the right temporal lobe in this process (see Section 2.2.). In a recent study using voxel-based lesion-behaviour mapping we addressed this discrepancy (Roswandowitz et al., 2018b). We found that lesions in the right posterior temporal lobe were associated with impaired voice-identity recognition (particularly for recently familiarised voice-identities). The lesion-behaviour association was independent of face-identity processing and discriminating between acoustical voice-identity features (VTL, GPR) suggesting a key role of the temporal lobe in selective voice-identity recognition. In contrast, lesions in the right inferior parietal lobe were linked to deficient voice-identity recognition only when voices were learned together with a face.

Recently, Luzzi et al. (2017) reported the first case of a *selective* familiar voice-identity recognition deficit (famous and personally familiar voices) following damage to the right aTL (as well as the right lenticular and caudate nuclei). The lesion in the case of MM involved the anterior right middle temporal gyrus, near the temporal pole. The lesion location was at the lateral surface of the aTL and seemed to overlap with the aSTG/S brain responses which have been observed with fMRI and MEG during voice-identity recognition in the neurotypical population (see Section 2.2.). MM's unfamiliar voice-identity discrimination was intact, likely owing to the intact posterior TVA regions which might be recruited for perceptual processing of unfamiliar and recently-familiarised voices (see Section 2.2.2.). MM also showed spared recognition of identity from the face. This is noteworthy, as most patients with aTL lesions have a multimodal person-identity recognition deficit, i.e. impairment in person recognition by voice, face, and name (Hailstone et al., 2010, Hailstone et al., 2011, Gainotti et al., 2003, Gainotti et al., 2008). These multimodal deficits might be caused by lesions to part of the *extended* system located in the aTL (von Kriegstein and Giraud 2006) or to a common lesion to both unimodal voice and face processing regions found in the aTL (Yang et al., 2016, Perrodin et al., 2012, Rajimehra et al., 2009, Collins and Olson, 2014).

Cases like MM highlight that phonagnosia *can* be observed in the absence of other person-recognition disorders such as prosopagnosia (see also Neuner and Schweinberger, 2000). Prosopagnosia, a visual homologue to phonagnosia, refers to a condition first documented by Bodamer (1947) in which patients display a distinct deficit in recognising identity from the face, while other visual abilities remain intact. Like phonagnosia, a developmental variant of the disorder has been reported (Lee et al., 2010, McConachie, 1976, Duchaine and Nakayama, 2005, Susilo and Duchaine, 2013). In addition, other aspects of auditory processing including language and music abilities also remained unaffected in the case of MM. A group study, which specifically examined the relation

between voice-identity and speech processing, found that aphasia in patients with left hemispheric lesions was unrelated to performance in a familiar voice-identity recognition task (Lang et al., 2009). These case and group studies indicate that voice-identity processing can reflect a unique cognitive ability, which can be selectively impaired (see Roswandowitz et al. 2018a, for review), an assumption that is central to traditional models of person recognition (see Figure 2).

3.3. Developmental phonagnosia: Behavioural evidence

Several studies have shown that phonagnosia can also occur as a developmental disorder, i.e. without apparent brain lesion (Garrido et al., 2009, Roswandowitz et al., 2014, Herald et al., 2014). The term 'congenital phonagnosia' is also used within the literature. Though, it is important to consider that this term assumes that the disorder has been present from birth. While it may be possible that phonagnosia may have a heritable component, currently the precise aetiology and onset of the disorder remains unclear.

To date, four behaviourally characterised cases and one anecdotal case of developmental phonagnosia have been documented (Garrido et al., 2009, Roswandowitz et al., 2014, Herald et al., 2014, Xu et al., 2015). Prevalence estimates suggested that anywhere within the range of 0.2 % (Roswandowitz et al., 2014) to 1 % (Xu et al., 2015), or even as high as 3.2 % (Shilowich and Biederman, 2016), of the population may be affected. This range in prevalence estimates might depend on how developmental phonagnosia is defined. For example, Roswandowitz et al. (2014) diagnostic criteria for phonagnosia encompassed the *exclusion* of any coinciding face, speech, or general auditory processing deficits. Higher prevalence estimates may be observed with less restrictive diagnostic criteria. For example, people in the autism spectrum often have voice-identity processing deficits (e.g. see Schelinski et al., 2017, Schelinski et al., 2016) and if there is no screening for this or other developmental disorders, higher prevalence estimates for voice-identity processing difficulties may be observed. Taken as a whole, the reported case studies have demonstrated that developmental phonagnosia *can* present in the absence of brain insult as a selective, modality specific, deficit in voice-identity processing[4]. The potential neural mechanisms of phonagnosia have been examined in three of the four reported developmental cases (Xu et al., 2015, Roswandowitz et al., 2017), these findings are reviewed below.

[4] Note that not all reported case studies have included control tests to examine the specificity of the disorder. For a detailed review of all control tests (and results) used in each of the case reports see Roswandowitz et al. (2018a).

3.4. Developmental phonagnosia: Neuroimaging investigations

Xu and colleagues reported on the case of AN, a 20-year-old female student who presented with a marked impairment in familiar voice-identity recognition (Herald et al., 2014, Xu et al., 2015). AN had normal face-identity recognition performance. Other assessments examining the specificity of her voice-identity recognition deficit in the auditory domain (e.g. speech or emotion processing) were not reported. Using fMRI, the authors examined responses in the bilateral TVAs as AN passively listened to a series of vocal, in contrast to non-vocal, sounds (same design as used in Belin et al., 2000, Pernet et al., 2015). This contrast revealed typical responses in the TVAs in AN. In a following fMRI experiment, aimed at targeting familiar voice-identity processing, AN and her controls were instructed to imagine a series of famous voices and non-voice sounds. Imagery trials were cued by a visually presented picture of a famous face and name or a non-human object and name combination. AN had reduced responses in the ventromedial prefrontal cortex (vmPFC), left precuneus, and left cuneus during voice, in contrast to non-voice, imagery. This finding led the authors to propose that altered function of the vmPFC, possibly driven by impaired fibre connections conveying voice information from the aTL to the vmPFC, may underpin AN's voice-identity recognition deficit. However, a connectivity analysis was not conducted to support this assumption. A meta-analysis of neuroimaging studies on person recognition has shown vmPFC involvement in multimodal famous person-identity processing (i.e., voice, face, and name), but not in identity processing of recently familiarised or personally familiar identities (Blank et al., 2014b). Therefore, it is unlikely that atypical responses in the vmPFC may fully explain AN's voice-identity processing deficit for both personally familiar and famous voices. In our view, reduced vmPFC responses could potentially be associated with AN's inability to imagine celebrities' voices, rather than being causal in her voice-identity recognition impairment (for discussion also see Roswandowitz et al., 2017).

Using fMRI, we examined two behaviourally well characterised cases of developmental phonagnosia- AS and SP (Roswandowitz et al., 2017). AS was a 32-year-old female, who showed a perceptual deficit in voice-identity processing (i.e. apperceptive phonagnosia, Lissauer, 1890, De Renzi et al., 1991, Roswandowitz et al., 2014). AS performed poorly on recognition of recently familiarised voices *and* on discrimination tasks with unfamiliar voices. However, she did not have severe problems in associating semantic information in the rare cases where she recognised a voice as familiar. In contrast to AS, SP, a 32-year old male, had intact perceptual analysis of voice-identity features (i.e. unimpaired discrimination of unfamiliar speakers' voices). Rather, SP showed a recognition impairment which was associated with a deficit in linking the voice percept with stored semantic information (i.e. associative phonagnosia, Warrington, 1975, Roswandowitz et al., 2014, Warrington and Shallice, 1984). We performed two fMRI experiments on AS and SP that had been repeatedly shown to elicit voice-sensitive

responses in neurotypicals: 1. A vocal vs. non-vocal sound experiment (same experiment as Belin et al. 2000) and 2. A voice-identity vs. speech-recognition experiment (similar design as von Kriegstein et al. 2003, Schelinski et al. 2016). Relative to controls, AS had reduced responses in regions of the core-voice system in each experiment. These regions included HG (Experiments 1 and 2) and PT (Experiment 2, HG and PT cluster of Experiment 2 extended to the right pSTG/S). This dysfunction in core-voice regions might explain AS perceptual difficulties with voice-identity recognition. SP, on the other hand, showed typical or even higher responses in the core-voice system in comparison to controls (Experiments 1 and 2). He had, however, reduced connectivity between the core-voice and extended system (Experiment 2). This may explain SP's impairment in linking voices with stored semantic information. Thus, the neuroimaging profiles of AS and SP's fit well with the nature of their respective voice-identity processing deficits (see Roswandowitz et al., 2018a for overview of core-voice and extended system in developmental phonagnosia subtypes).

4. Cross-modal interactions in voice-identity processing

4.1. Face-voice interactions in the neurotypical brain

Although most of us can recognise identity from the voice alone, in our everyday interactions voices are rarely heard in isolation. Rather, our typical communication scenarios are often face to face, where we are exposed to both the voice and the face of the speaker. These two cues, voice and face, convey concordant information to support identity processing (Smith et al., 2016a, Kamachi et al., 2003, Smith et al., 2016b, Ghazanfar et al., 2007). Both identity cues have also been shown to be represented in a similar manner in the neurotypical brain (Latinus and Belin, 2011, Valentine, 1991, Zäske et al., 2010, Leopold et al., 2005, for review see Yovel and Belin, 2013).

There is evidence that when presented together the face and voice interact to support voice-identity processing (Robertson and Schweinberger, 2010, Schweinberger et al., 2011a, Zäske et al., 2015, O'Mahony and Newell, 2012). For example, Schweinberger et al. (2007) noted that listeners were faster at judging whether a voice was from a familiar (voices of lecturers who were personally familiar to listeners) or unfamiliar speaker when the voice was presented with its corresponding (i.e. identity matched), in comparison to non-corresponding, time-synchronised facial identity. This behavioural benefit on voice-identity recognition was strongest for familiar voices, implying that previously established *multimodal* representations shape these audio-visual interactions.

Intriguingly, there is ample evidence that these audio-visual interactions in voice-identity recognition also persist when concurrent visual information is unavailable (Schelinski et al., 2014, Schall et al., 2013, von Kriegstein and Giraud, 2006, von Kriegstein

et al., 2008, von Kriegstein, 2012). For example, listeners are more accurate at recognising the identity of a speaker by voice alone, when they have been previously familiarised with the speaker's corresponding facial identity (e.g. von Kriegstein et al., 2008). This effect emerges rather rapidly following approximately two minutes of audio-visual experience with the speaker's identity (von Kriegstein et al., 2008, Schall et al., 2013, Schelinski et al., 2014). Studies observing this behavioural enhancement have employed a number of control learning conditions, including familiarising listeners with the speaker by voice alone (Sheffert and Olson, 2004) or in conjunction with other visual input including the name (von Kriegstein and Giraud, 2006) or an image depicting the occupation of the speaker (Schelinski et al., 2014, von Kriegstein et al., 2008, Schall et al., 2013). Across these manipulations voice-identity recognition remained superior for speakers who had been learned in conjunction with their facial identity (for review see von Kriegstein, 2012). This behavioural enhancement is termed the "face-benefit" (von Kriegstein et al., 2008).

While several studies have observed the face-benefit on voice-identity recognition, there is also evidence that voice-identity recognition can be *impaired* by the presence of a face during learning. It is argued that the saliency of the face may interfere with the ability to attend to the voice-identity, an effect termed 'face-overshadowing' (FOE; Cook and Wilding, 1997, Cook and Wilding, 2001). The FOE appears to be mediated by the degree of exposure to the audio-visual identity. For example, while Cook and Wilding (1997, 2001) noted that the presence of a face interfered with voice-identity recognition for speakers heard uttering a *single* 15 syllable sentence, this effect was absent when three utterances were heard. Subsequent studies have also demonstrated that the FOE can be mitigated over time (e.g. Zäske et al., 2015). These findings are in line with natural day-to-day interactions, where repeated audio-visual interactions with others are typical. Indeed, there is evidence that these audio-visual interactions may be one of the reasons we recognise personally familiar voices with such ease (see section 4.2. for further discussion).

Converging evidence from fMRI and MEG studies have demonstrated that the face-benefit for voice-identity recognition is supported by responses in the fusiform face area (FFA) (Schall et al., 2013, von Kriegstein and Giraud, 2006, von Kriegstein et al., 2008). The FFA is a visual face-sensitive region (Kanwisher et al., 1997) in the fusiform gyrus which is engaged in the processing of face-identity and facial-form cues (Grill-Spector et al., 2004, Liu et al., 2010, Kanwisher and Yovel, 2006, Axelrod and Yovel, 2015, Andrews and Ewbank, 2004, Weibert and Andrews, 2015, Ewbank and Andrews, 2008, Xu et al., 2009, Eger et al., 2004). Functional connectivity and direct structural connections between the FFA and voice-sensitive regions in anterior and middle STG/S have been reported (Schall and von Kriegstein, 2014, von Kriegstein et al., 2005, von Kriegstein and Giraud, 2006, Blank et al., 2011) and FFA responses to voices are elicited early after voice onset (i.e. 110

ms Schall et al., 2013). The direct connections likely facilitate communication between the voice-sensitive and face-sensitive regions, which has been shown to occur during the auditory-only recognition of speakers known by face (Schall and von Kriegstein, 2014, von Kriegstein et al., 2005).

There is also evidence that additional regions of the core-face processing network (Haxby et al., 2000, Haxby et al., 2002) show sensitivity to vocal input (Blank et al., 2015). These additional regions include the occipital face area (OFA), and the anterior temporal lobe face area (aTL-FA). The OFA is a region which is argued to be involved in the early perception of facial structure (Duchaine and Yovel, 2015, Pitcher et al., 2011, Haxby et al., 2000, Gauthier et al., 2000), complementing later stages of identity processing in the FFA and aTL-FA. In conjunction with the FFA, the aTL-FA has been shown to be sensitive to face-identity and may be involved in the final, potentially more abstract, stages of face-identity recognition (for recent reviews see Duchaine and Yovel, 2015, Collins and Olson, 2014). To examine the potential role of these regions in supporting voice processing, Blank et al. (2015) employed a voice-face priming paradigm. They familiarised their participants with a series of speakers, via voice and face. During fMRI data acquisition, they presented the recently-familiarised voices followed by morphed faces that matched or mismatched with respect to identity or physical properties. The authors noted that the OFA represented information about both the physical properties and identity of the voice, the FFA represented identity, and the aTL-FA represented identity information to a higher extent than information about physical properties of the voice. Functional connections between the FFA and mid and anterior STG/S voice-sensitive regions were also observed, indicating that the FFA might be the key entry region for sharing of voice and face information during an identity recognition task. Taken together, these findings suggest interactions, at multiple levels, between the voice- and face-processing systems.

4.2. Face-voice interactions in familiar voice-identity recognition

Cross-modal interactions may support the recognition of familiar voices. Indeed, responses in the FFA have also been reported during the auditory-only recognition of speakers who are personally familiar to the listener, i.e. where audio-visual face-to-face communication is typical (von Kriegstein et al., 2005, von Kriegstein et al., 2006a). The influence of face information on voice-identity recognition is also supported by evidence that familiar voice-identity recognition is often poorer in individuals with developmental prosopagnosia, relative to their neurotypical controls (von Kriegstein et al., 2006a, Jones and Tranel, 2001, see Maguinness and von Kriegstein, 2017 for recent review). Developmental prosopagnosics show a selective impairment in voice-identity recognition for visually familiar speakers, i.e. speakers known by face (Jones and Tranel, 2001, von Kriegstein et al., 2006a, von Kriegstein et al., 2008). In contrast, their voice-identity recognition for visually unfamiliar speakers, e.g. speakers known by voice alone, is

unimpaired (von Kriegstein et al., 2008, Liu et al., 2015). This provides additional compelling evidence for the integral role of visual face mechanisms in voice-identity recognition (see Maguinness and von Kriegstein, 2017 for recent review).

5. Re-examining current models of voice-identity processing

5.1. Is voice-identity processing for voices of different familiarity the same sequential process?

Traditional models of voice-identity processing (see Figure 2) have been influential in providing a framework for our understanding of how voice-identity processing is achieved. However, several findings reported in this review conflict with an assumption that is at the core of these models, i.e. that voice-identity processing proceeds in a sequential manner from perceptual stages to identity recognition stages. For example, findings from the clinical studies (reviewed in Section 3), indicate that individuals with brain lesions who are impaired on discrimination of unfamiliar voices, a process which likely relies heavily on *perceptual* voice analysis, can nevertheless recognise a *familiar* voice and vice versa (Van Lancker and Kreiman, 1987, Van Lancker et al., 1988, Van Lancker et al., 1989). This points to the idea that familiar and unfamiliar voice-identity processing may at least partially dissociate in the human brain. More evidence against a strict sequential processing was reported in Section 2.2.3. Here we highlighted that responses in pSTG/S (measured via MEG) occur at a similar time point (~200ms) to behaviourally relevant responses in the aSTG/S, during a voice-identity recognition task (Schall et al., 2015). Schall and colleagues observed these parallel responses in STG/S regions for voices with which participants were recently familiarised. If voice-identity processing proceeded in a strictly sequential manner one might expect early responses in the pSTG/S, a region implicated in the perceptual processing of the voice (see Section 2.2.2.), followed by later responses in the aSTG/S.

5.2. An integrative model of voice-identity processing

How might one explain that recognition of familiar voices does not always rely on the more basic perceptual processes involved in discrimination of unfamiliar voices?[5] The

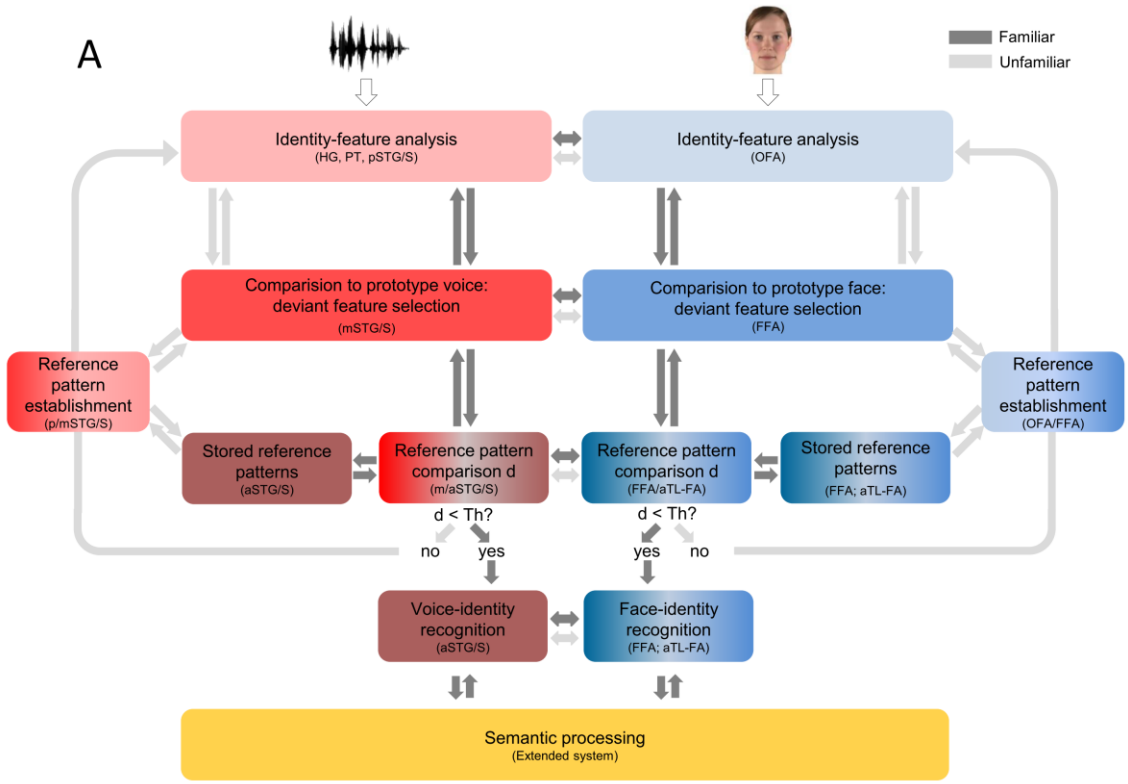
[5] Note that previous models of face- and voice-identity processing have been often geared towards explaining familiar face or voice identification. They do not explicitly integrate differences between unfamiliar and familiar voice and face processing (Bruce & Young, 1986, Haxby et al., 2000, Burton et al., 1990; Ellis et al., 1997, Kreiman & Sidtis, 2011, Belin et al., 2004, Lavner et al. 2001). However, many authors have highlighted the potential for differences in the processing of familiar and unfamiliar identities. Most notably that unfamiliar faces and voices rely more heavily on a detailed perceptual analysis which is tightly bound to the incoming stimulus, e.g. view-dependent or feature-based analysis, compared to their familiar counterparts (Bruce & Young, 1986, Kreiman & Sidtis, 2011).

findings may be potentially understood if one extends the prototype model of voice-identity processing (see Section 2.1.1.) and merges it with the existing traditional models of voice-identity processing (see Section 2.1.2.). In the following section, we attempt to bring together these models and thereby explain the potential partial dissociation between voice-identity recognition of familiar voices and perceptual processing of unfamiliar voices. Due to the intricate link between voice-identity and face-identity processing and the potential equivocal computations involved, we extend this model to include interactions between voice- and face-processing streams. For both modalities, we include brain regions which may be potentially involved in the respective levels of identity processing. Although the attribution of these specific brain regions is still tentative, we point to them to inform experimentally testable hypotheses on how identity processing might be achieved in the human brain.

5.2.1. Identity-feature analysis and prototype processing

A central aspect of voice-identity processing is the perceptual level of processing, where identity-features are extracted and merged to create a coherent voice percept i.e. *identity-feature analysis* (Figure 3A, left of figure). Features of the voice supporting speech and vocal-emotion processing are also analysed at the perceptual level. Evidence reviewed in Sections 3.2. and 3.3. supports the argument that these features may be processed in partially distinct systems (see Roswandowitz et al. 2018a for further discussion). Perceptual voice-identity processing is likely supported by the pSTG/S, the PT, and anterolateral HG (Figure 3B, pink patches) of the core-voice system (see Sections 2.2.2. and 3.4.). On the right of Figure 3A we show the parallel perceptual analysis of face-individuating features. The OFA (Figure 3B, light-blue patch), a region implicated in the early processing of facial structure cues (Duchaine and Yovel, 2015, Pitcher et al., 2011, Haxby et al., 2000, Gauthier et al., 2000), is also sensitive to physical and identity cues from the voice (Blank et al., 2015).

The extracted features may then be contrasted against a stored prototype(s) or average voice (Lavner et al., 2001), i.e. *comparison to prototype voice* (Figure 3A, left of figure). The acoustical deviations between the incoming merged voice percept and the prototype voice are computed (Lavner et al., 2001). Based on findings by Latinus et al. (2013), we suggest that this process may be supported by mid regions of the STG/S in the core-voice system (Figure 3B, red patch). There is also substantial evidence that individual faces may be contrasted against an internal prototype face (Figure 3A, right of figure) (Leopold et al., 2005, Leopold et al., 2001, Newell et al., 1999, Rhodes and Jaquet, 2011), possibly supported by processing in the FFA (Loffler et al., 2005) (Figure 3B, blue patch). The mSTG/S and aSTG/S share connections with the FFA (Blank et al., 2011), potentially these regions may exchange information at this level of processing. There is behavioural and neuroimaging evidence to suggest such potential interactions (von Kriegstein et al.,



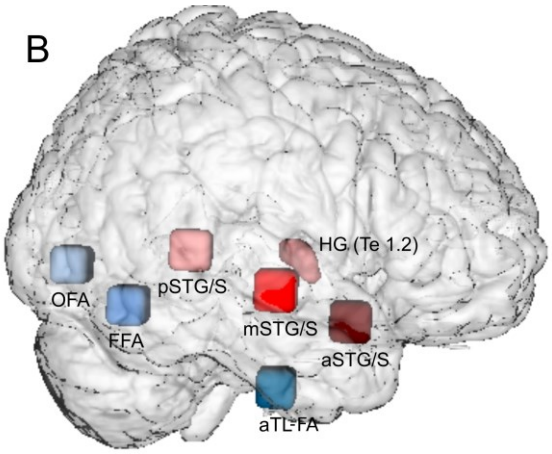


Figure 3. **A) Audio-visual integrative model** which describes the processes involved in voice-identity processing. This model incorporates aspects of previous models proposed by Bruce and Young (1986), Ellis et al. (1997), Neuner and Schweinberger (2000), Lavner et al. (2001), Belin et al. (2004), von Kriegstein et al., 2005, Blank et al. (2014b), and Roswandowitz et al., (2018a). The colour codes in (A) for the different levels of identity processing refer to the potential brain regions associated with the respective process in (B). The relevant brain region(s) for each level of processing are listed in the coloured boxes. Boxes in (A) with transitioning colours denote when more than one region has been implicated in the literature for a level of processing, or also potential *interactions* between the respectively coloured brain regions in(B). Dark grey arrows denote processing for familiar identities, light grey for unfamiliar identities. The bidirectional horizontal arrows denote cross-modal interactions. These cross-modal interactions may be stronger for familiar compared to unfamiliar identities. **B) Brain regions in the core-voice system** which may support voice-identity processing (red colours) and in the core-face system which may support face-identity processing (blue colours). *Voice-identity processing*: Pink patches: The regions are centred on the peak MNI-coordinates

reported in von Kriegstein et al. (2007) for VTL processing in pSTG/S (60, -34, 4). HG is visualised based on the probabilistic anatomical map of the HG Te 1.2 according to the Juelich Histological Atlas (Morosan et al., 2001, Eickhoff et al., 2005). Te 1.2 refers to the anterolateral part of HG as defined by Morosan et al. (2001). Anterolateral HG was responsive to GPR-change induced speaker-identity changes in von Kriegstein et al. (2010). Red patch: The mSTG/S is centred on the peak coordinates reported by Latinus et al., 2013 for prototype processing in Experiment 3 (63, -9, -6). Bordeaux patch: The aSTG/S is centred on the peak coordinates reported by von Kriegstein et al. (2003) for voice-identity recognition (54, 12, -15). *Face-identity processing:* The regions are centred on the peak coordinates from studies localising the respective region. Light blue patch: OFA (Gauthier et al., 2000 peak co-ordinates: 30, -77, -4). Blue patch: FFA (Kanwisher et al., 1997 peak co-ordinates: 40, -56, -15). Dark blue patch: aTL-FA (Axelrod and Yovel, 2013 peak co-ordinates: 34, -10, -39). All co-ordinates are reported in x,y,z orientation in MNI space. HG = Heschl's gyrus; STG/S = superior temporal gyrus/sulcus; OFA = occipital face area; FFA = fusiform face area; TL-FA: temporal lobe face area; a = anterior; m = mid; p = posterior.

2005, von Kriegstein et al., 2008, Blank et al., 2015, Bülthoff and Newell, 2017, Bülthoff and Newell, 2015). How multidimensional face- and voice-spaces interact remains an open question. However, given the behavioural findings (Bülthoff and Newell, 2017, Bülthoff and Newell, 2015) and the observation that the FFA and STG/S voice-sensitive regions are directly connected in the neurotypical brain (Blank et al., 2011), we assume that multidimensional voice- and face-spaces may be modulated by cross-modal input. This would potentially negate the need for a multimodal face-voice space.

Following comparison to the prototype, the deviating voice features are extracted, i.e. *deviant feature selection*, and can be passed on for further analysis. These features may be compared to internal 'stored reference patterns' which are unique to each voice-identity (Lavner et al., 2001), i.e. *reference pattern comparison*. The distance between the stored and incoming pattern is computed, i.e. *d* (Figure 3A, left of figure). If there is a sufficient match, 'd' is lower than some perceptual threshold (Th), a sense of familiarity is generated, i.e. *voice-identity recognition* (Figure 3A, left of figure). Note that damage at this stage of processing would produce a deficit in voice-identity recognition, while leaving perceptual analysis of the voice-identity intact (see Section 3.2.). We have previously termed this 'familiarity-associative phonagnosia' (Roswandowitz et al. 2018a).

However, for unfamiliar voices there is not yet a matching stored reference pattern. We suggest that this is the point where unfamiliar and familiar voice-identity processing might start to dissociate. The familiar voice can be recognised (voice-identity recognition, Figure 3A, left of figure). In contrast, for the unfamiliar voice a reference pattern needs to be established, i.e. *reference pattern establishment* (Figure 3A, left of figure). Depending on the amount of training and the distinctiveness of the voice, there may only be a relatively incomplete reference pattern available for recently-familiarised voices. This pattern may not be sufficient to robustly recognise the voice-identity. This

reference pattern will be refined with continued exposure to the voice-identity, becoming more robust through a continued process of reference pattern establishment.

5.2.2. Unfamiliar and recently-familiarised voices: Reference pattern establishment

Reference pattern establishment likely involves an iterative loop with the identity-feature analysis and comparison to the prototype voice. In the model, this is denoted by the loop between these processing levels for unfamiliar voice-identity processing (left of Figure 3A, light-grey arrows). We refer here to this as a *perceptual voice-identity processing loop*. Note that the initiation of this loop will most often occur when the incoming voice does not match with a stored reference pattern, i.e. 'd' > Th. However, the loop may also begin at the outset of processing, i.e. without having to pass through the level of *reference pattern comparison (d)*. For example, this could occur when one knows one is encountering an unfamiliar voice. Evidence from the lesion literature, examinations in individuals with ASD, and the neurotypical population have revealed the pSTG/S as an important region for the processing of unfamiliar voices or recently-familiarised voices (Roswandowitz et al., 2018b, von Kriegstein et al., 2004, Schelinski et al., 2017). Increased functional connectivity between the pSTG/S, a region implicated in perceptual voice-identity processing, and the mSTG/S, a region implicated in prototype processing (Latinus et al., 2013), has been observed during unfamiliar, compared to familiar voice-identity recognition (von Kriegstein and Giraud, 2004). This suggests that interactions between these regions might be particularly relevant for processing unfamiliar voices. Unfamiliar and recently familiarised voices may require multiple iterations through the 'perceptual voice-identity processing loop' in order to generate a robust representation, i.e. reference pattern of the new voice-identity. The number of loop iterations will likely be modulated by the perceived averageness or distinctiveness of the voice-identity. For example, a voice which is further from the prototype voice, i.e. distinctive, may require less iterations than one which is more average. The availability of the face during learning may also modulate this process (denoted by the horizontal arrows indicating interactions between the posterior and mSTG/S voice regions and OFA and FFA). Recently-familiarised voices are better recognised when they have been learned by face, compared to other control learning conditions where the listener is exposed to the voice-identity for an *equable* amount of time (von Kriegstein et al., 2008, Sheffert and Olson, 2004, Schall et al., 2015, Schelinski et al., 2014). A more robust reference pattern may therefore emerge more readily for face-learned speakers' voices. In general, reference patterns are likely built up through multiple iterations during learning. An incomplete reference pattern may initially be laid down (e.g. with recently-familiarised voices). With continued exposure a unique voice reference pattern will eventually be stored for the new voice-identity, joining other *stored reference patterns*. In Figure 3A, this is denoted by the interacting arrows between *'reference pattern establishment'* and *'stored reference patterns'*. The need for a reference pattern establishment might explain the lesion study findings of *impaired* perceptual

discrimination of unfamiliar voice-identities and intact familiar voice-identity recognition i.e. apperceptive phonagnosia (see Sections 3.1. and 3.4.). One can explain this under the model with two potentially complementary scenarios. First, apperceptive phonagnosia might be caused by lesions affecting regions that are responsible for the reference pattern establishment (see Figure 3A). Potentially, there may be dedicated sub-regions within the pSTG/S and/or connectivity patterns with mSTG/S. These might be involved in the computations necessary for discriminating and learning unfamiliar voices. Second, potentially the initial levels of processing could be partly intact - enough for one run-through for familiar voice-identity recognition, but not enough for the reiterative mechanisms required for unfamiliar voices or recently-familiarised voices. These are two scenarios, but it is an open testable hypothesis whether these preliminary levels of processing are entirely essential for recognising voice-identities for which a representation has already been established.

It is important to consider that interactions between the two systems, i.e. perceptual voice-identity processing and voice-identity recognition, are likely to occur throughout development of voice-identity representations. For example, similar dissociations reported in the lesion literature of impaired unfamiliar voice-identity discrimination, with intact familiar voice-identity recognition have been observed in individuals with ASD (Schelinski et al., 2017). Importantly, these individuals also have a deficit in *learning* new voice-identities i.e. laying down new voice reference patterns. In addition, the parallel responses observed in pSTG/S and aSTG/S during a voice-identity recognition task for *recently-familiarised* voices (Schall et al., 2015) may reflect important interactions between the two systems when refining new voice-identity representations in neurotypical processing. Therefore, it is likely that some interaction between these processes is necessary to establish new, but perhaps not to maintain, voice representations.

5.2.3. Recognising familiar voices

Recognition of familiar voices requires comparing the incoming voice to stored reference patterns i.e. computing the distance ('d') between the stored and incoming patterns (Figure 3A, left of figure). Currently it is unclear whether or where this computation occurs in the neurotypical brain. It is possible that such a process may be supported by the mid STG/S (red patch, Figure 3B) and possibly also the aSTG/S (bordeaux patch, Figure 3B) (Bethmann and Brechmann, 2014, von Kriegstein and Giraud, 2004, Schall et al., 2015, Belin and Zatorre, 2003, von Kriegstein et al., 2003). It has been proposed that individual voice identities may be represented in the aSTG/S (Formisano et al., 2008, Schall et al., 2015). However, it is an open question what the nature of these representations is, e.g. whether the aSTG/S communicates with more posterior regions,

acting as a potential hub for stored reference patterns. Alternatively, it could represent the output from this process, i.e. voice-identity recognition.

If a voice is familiar, the speaker is also often known by face (von Kriegstein et al., 2005, von Kriegstein and Giraud, 2006, Maguinness and von Kriegstein, 2017). Several studies have demonstrated sensitivity in the FFA to face-identity (right of figure 3A) (Loffler et al., 2005, Grill-Spector et al., 2004, Kanwisher and Yovel, 2006, Axelrod and Yovel, 2015, Vida et al., 2016, Nestor et al., 2011). Unique blood oxygenation level dependent response patterns have been observed in the FFA in response to different facial identities (Nestor et al., 2011, Natu et al., 2010). It is therefore possible that the FFA may be engaged in contrasting the incoming face input with stored reference patterns which are unique for each face-identity. Potentially the output of this process may engage the aTL-FA. The mSTG/S and aSTG/S also shares direct structural connections with the FFA (Blank et al., 2011). These cross-modal interactions support the recognition of familiar voices and recently-familiarised voices that have been learned together with faces (von Kriegstein et al., 2008, see von Kriegstein, 2012 for review). The aTL-FA (dark blue patch, Figure 3B) is also sensitive to *identity* information in voices (Blank et al., 2015). The aTL-FA, might have similar roles for face-identity representations as the aSTG/S has for voice-identity representations (for reviews see Duchaine and Yovel, 2015, Collins and Olson, 2014). Note, that it is also possible to discriminate responses in the aTL to unique face identities (Kriegeskorte et al., 2007). Sensitivity of the aTL-FA to voice input suggests that interactions between voice-identity and face-identity processing regions may occur at multiple levels (Blank et al., 2015).

After the voice has been recognised as familiar the 'meaning' of the voice can be accessed, i.e. semantic processing, through interactions with an extended system. The extended system can also be accessed via the visual modality (Haxby et al., 2000, Haxby et al., 2002). Disruption of the propagation of signals *between* the core-voice and the extended system could result in an inability to link semantic information to the voice-identity (Roswandowitz et al., 2014, 2017, see Section 3.4). Damage to the extended system, rather than altered connectivity *between* the systems, would likely result in a multi-modal (i.e. not voice-selective) person-identity recognition disorder.

6. Concluding remarks and future considerations

Converging evidence from fMRI and MEG in neurotypicals, cases of developmental and acquired phonagnosia, and ASD demonstrate the important role that regions of the temporal lobe play in voice-identity processing (von Kriegstein et al., 2003, Roswandowitz et al., 2018b, Belin and Zatorre, 2003, Schelinski et al., 2016, Roswandowitz et al., 2017, Andics et al., 2010, Assal et al., 1981, Van Lancker et al., 1988, Van Lancker et al., 1989, von

Kriegstein and Giraud, 2004, Schall et al., 2015). Although it is not yet fully understood how and at what time point these regions interact to support voice-identity processing, the reviewed findings suggest that interactions between and responses in these regions may vary as a function of the familiarity of the voice.

The reviewed findings from both behavioural and neuroimaging methods highlight that prototype processing may play a fundamental role in the computational processes involved in the representation of unique voice-identities in the brain. This raises an interesting question; might some individuals who are impaired in voice-identity recognition show altered or impaired prototype voice processing? Although no study has to date explicitly addressed this question one can look to research which has examined face-space, the visual homologue to auditory voice-space (Valentine, 1991), in developmental prosopagnosics. Perhaps surprisingly, these studies demonstrate that developmental prosopagnosics often perceive faces in a manner which is consistent with prototype identity processing (Nishimura et al., 2010, Susilo et al., 2010, but see Palermo et al., 2011). This suggests that the recognition deficit in developmental prosopagnosia may be linked to atypicalities at later stages of processing (Susilo and Duchaine, 2013, Nishimura et al., 2010), potentially where the outputted deviant features in face-space are to be linked to stored reference patterns. Future studies may examine this in developmental phonagnosics, particularly in relation to potential differences in prototype processing in apperceptive or associative variants of the disorder (Roswandowitz et al., 2014, 2017). To date, voice-space in neurotypicals has been largely investigated via controlled manipulations, or examination, of voice stimuli which altered a *sample* of the voice features which listeners can use to support recognition (Latinus et al., 2013). Further studies may continue to address how voice-space is defined in neurotypicals, e.g. whether it is flexibly updated with exposure to new voice-identities (i.e. prototype may shift in line with exposure) and may explore the potential that this space may interact directly with identity information from other sensory modalities, i.e. the face (Bülthoff and Newell, 2015, Bülthoff and Newell, 2017). The potential for such cross-modal interactions is possible, given the sensitivity of face-processing regions to voice input (von Kriegstein et al., 2008, von Kriegstein and Giraud, 2006, von Kriegstein et al., 2006a, Schall et al., 2013, Blank et al., 2015).

The reviewed findings suggested that familiar and unfamiliar voices may be represented differently in the brain. Our audio-visual integrative model can explain these findings and generates testable predictions on how the human brain accomplishes the feat of learning and recognising the identity of others.

Acknowledgements

We would like to thank two anonymous reviewers for their careful reading and insightful comments on previous versions of this manuscript. This work was funded by a Max Planck Research Group grant to KvK. The authors declare no competing financial interests.

References

AGLIERI, V., WATSON, R., PERNET, C., LATINUS, M., GARRIDO, L. & BELIN, P. 2017. The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods,* 49**,** 97-100.

ANDICS, A., MCQUEEN, J. M., PETERSSON, K. M., GAL, V., RUDAS, G. & VIDNYANSZKY, Z. 2010. Neural mechanisms for voice recognition. *NeuroImage,* 52**,** 1528-40.

ANDREWS, T. & EWBANK, M., P 2004. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage,* 23**,** 905-913.

ASSAL, G., AUBERT, C. & BUTTET, J. 1981. Asymetrie cerebrale et reconnaissance de la voix. *Review Neurology (Paris),* 137**,** 255-268.

ASSAL, G., ZANDER, E., KREMIN, H. & BUTTET, J. 1976. Discrimination des voix lors des lesions du cortex cerebral. *Archives Suisses de Neurologie, Neurochirurgie et de Psychiatrie,* 119**,** 307-315.

AXELROD, V. & YOVEL, G. 2013. The challenge of localizing the anterior temporal face area: A possible solution. *NeuroImage,* 81**,** 371-380.

AXELROD, V. & YOVEL, G. 2015. Successful decoding of famous faces in the fusiform face area. *PLoS ONE,* 10**,** e0117126.

BARSICS, C. & BRÉDART, S. 2012. Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology,* 24**,** 789-795.

BEAUCHEMIN, M., DE BEAUMONT, L., VANNASING, P., TURCOTTE, A., ARCAND, C., BELIN, P. & LASSONDE, M. 2006. Electrophysiological markers of voice familiarity. *European Journal of Neuroscience,* 23**,** 3081-3086.

BELIN, P., BESTELMEYER, P. E. G., LATINUS, M. & WATSON, R. 2011. Understanding voice perception. *British Journal of Psychology,* 102**,** 711-725.

BELIN, P., FECTEAU, S. & BEDARD, C. 2004. Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences,* 8**,** 129-135.

BELIN, P. & ZATORRE, R. J. 2003. Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport,* 14**,** 2105-2109.

BELIN, P., ZATORRE, R. J. & AHAD, P. 2002. Human temporal-lobe response to vocal sounds. *Cognitive Brain Research,* 13**,** 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.

BETHMANN, A. & BRECHMANN, A. 2014. On the definition and interpretation of voice selective activation in the temporal cortex. *Frontiers in Human Neuroscience,* 8**,** 499.

BETHMANN, A., SCHEICH, H. & BRECHMANN, A. 2012. The Temporal Lobes Differentiate between the Voices of Famous and Unknown People: An Event-Related fMRI Study on Speaker Recognition. *PLoS ONE,* 7**,** e47626.

BIRKETT, P. B., HUNTER, M. D., PARKS, R. W., FARROW, T. F., LOWE, H., WILKINSON, I. D. & WOODRUFF, P. W. 2007. Voice familiarity engages auditory cortex. *Neuroreport,* 18**,** 1375-1378.

BLANK, H., ANWANDER, A. & VON KRIEGSTEIN, K. 2011. Direct Structural Connections between Voice- and Face-Recognition Areas. *The Journal of Neuroscience,* 31**,** 12906-12915.

BLANK, H., KIEBEL, S. J. & VON KRIEGSTEIN, K. 2014a. How the human brain exchanges information across sensory modalities to recognize other people. *Hum Brain Mapp,* 36**,** 324-339.

BLANK, H., KIEBEL, S. J. & VON KRIEGSTEIN, K. 2015. How the human brain exchanges information across sensory modalities to recognize other people. *Hum Brain Mapping,* 36**,** 324-339.

BLANK, H., WIELAND, N. & VON KRIEGSTEIN, K. 2014b. Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience and Biobehavioral Reviews,* 47**,** 717–734.

BODAMER, J. 1947. Die Prosop-Agnosie. *Arch Psychiatr Nervenkr Z Gesamte Neurol Psychiatr,* 118**,** 6-53.

BRUCE, V. & YOUNG, A. 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.

BÜLTHOFF, I. & NEWELL, F. N. 2017. Crossmodal priming of unfamiliar faces supports early interactions between voices and faces in person perception. *Visual Cognition*.

BURTON, A. M., BRUCE, V. & JOHNSTON, R. A. 1990. Understanding face recognition with an interactive activation model. *British Journal of Psychology,* 81 (Pt 3)**,** 361-80.

CAMPANELLA, S. & BELIN, P. 2007. Integrating face and voice in person perception. *Trends Cogn Sci.,* 11**,** 535-543.

CAPILLA, A., BELIN, P. & GROSS, J. 2013. The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cereb Cortex,* 23**,** 1388-95.

CHANG, L. & TSAO, D. Y. 2017. The code for facial identity in the primate brain. *Cell,* 169**,** 1013-1028.

CHAREST, I., PERNET, C. R., ROUSSELET, G. A., QUINONES, I., LATINUS, M., FILLION-BILODEAU, S., CHARTRAND, J. P. & BELIN, P. 2009. Electrophysiological evidence for an early processing of human voices. *BMC Neurosci,* 10**,** 127.

COLLINS , J. A. & OLSON, I. R. 2014. Beyond the FFA: The role of the ventral anterior temporal lobes in face processing. *Neuropsychologia,* 61**,** 65-79.

COOK, S. & WILDING, J. 1997. Earwitness testimony .2. Voices, faces and context. *Applied Cognitive Psychology,* 11**,** 527-541.

COOK, S. & WILDING, J. 2001. Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology,* 92**,** 617-629.

CUTZU, F. & EDELMAN, S. 1996. Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences,* 93**,** 12046–12050.
DE RENZI, E., FAGLIONI, P., GROSSI, D. & NICHELLI, P. 1991. Apperceptive and associative forms of prosopagnosia. *Cortex; a journal devoted to the study of the nervous system and behavior,* 27**,** 213-221.
DUCHAINE, B. & NAKAYAMA, K. 2005. Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience,* 17**,** 249-261.
DUCHAINE, B. & YOVEL, G. 2015. A revised neural framework for face processing. *Annual Review of Vision Science,* 1**,** 393-416.
EGER, E., SCHYNS, P. G. & KLEINSCHMIDT, A. 2004. Scale invariant adaptation in fusiform face-responsive regions. *NeuroImage,* 22**,** 232-242.
EICKHOFF, S. B., STEPHAN, K. E., MOHLBERG, H., GREFKES, C., FINK, G. R., AMUNTS, K. & ZILLES, K. 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage,* 25**,** 1325-35.
ELLIS, H. D., JONES, D. M. & MOSDELL, N. 1997. Intra- and inter-modal repetition priming of familiar faces and voices. *Br.J.Psychol.,* 88 (Pt 1)**,** 143-156.
EWBANK, M. P. & ANDREWS, T. J. 2008. Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. *NeuroImage,* 40**,** 1857-1870.
FECTEAU, S., ARMONY, J. L., JOANETTE, Y. & BELIN, P. 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. *NeuroImage,* 23**,** 840-8.
FONTAINE, M., LOVE, S. A. & LATINUS, M. 2017. Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology,* 8**,** 1180.
FORMISANO, E., DE MARTINO, F., BONTE, M. & GOEBEL, R. 2008. "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science,* 322**,** 970-973.
GAINOTTI, G. 2015. Implications of recent findings for current cognitive models of familiar people recognition. *Neuropsychologia* 77**,** 279-287.
GAINOTTI, G., BARBIER, A. & MARRA, C. 2003. Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. *Brain : a journal of neurology,* 126**,** 792-803.
GAINOTTI, G., FERRACCIOLI, M., QUARANTA, D. & MARRA, C. 2008. Cross-modal recognition disorders for persons and other unique entities in a patient with right fronto-temporal degeneration. *Cortex,* 44**,** 238-48.
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.
GAUTHIER, I., TARR, M. J., MOYLAN, J., SKUDLARSKI, P., GORE, J. C. & ANDERSON, A. W. 2000. The fusiform "face area" is part of a network that processes faces at the individual level. *J Cogn Neurosci,* 12**,** 495-504.
GHAZANFAR, A. A., TURESSON, H., K, MAIER, J. X., VAN DINTHER, R., PATTERSON, R. D. & LOGOTHETIS, N. K. 2007. Vocal-tract resonances as indexical cues in rhesus monkeys. *Current Biology,* 17**,** 425-430.
GOGGIN, J., THOMPSON, C., STRUBE, G. & SIMENTAL, L. 1991. The role of language familiarity in voice identification. *Memory & Cognition,* 19**,** 448-458.

GRILL-SPECTOR, K., KNOUF, N. & KANWISHER, N. 2004. The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience,* 7**,** 555-562.

HAILSTONE, J. C., CRUTCH, S. J., VESTERGAARD, M. D., PATTERSON, R. D. & WARREN, J. D. 2010. Progressive associative phonagnosia: a neuropsychological analysis. *Neuropsychologia,* 48**,** 1104-14.

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**,** 2535-2547.

HAXBY, J. V., HOFFMAN, E. A. & GOBBINI, M. I. 2000. The distributed human neural system for face perception. *Trends Cogn Sci,* 4**,** 223-233.

HAXBY, J. V., HOFFMAN, E. A. & GOBBINI, M. I. 2002. Human neural systems for face recognition and social communication. *Biol Psychiatry,* 51**,** 59-67.

HERALD, S. B., XU, X., BIEDERMAN, I., AMIR, O. & SHILOWICH, B. E. 2014. Phonagnosia: A voice homologue to prosopagnosia. *Visual Cognition,* 22**,** 1031-1033.

JONES, R. D. & TRANEL, D. 2001. Severe developmental prosopagnosia in a child with superior intellect. *Journal of Clinical and Experimental Neuropsychology,* 23**,** 265-273.

KAMACHI, M., HILL, H., LANDER, K. & VATIKIOTIS-BATESON, E. 2003. "Putting the face to the voice": matching identity across modality. *Curr.Biol.,* 13**,** 1709-1714.

KANWISHER, N., MCDERMOTT, J. & CHUN, M. M. 1997. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience,* 17**,** 4302-4311.

KANWISHER, N. & YOVEL, G. 2006. The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* 361**,** 2109-2128.

KREIMAN, J. & SIDTIS, D. 2011. Voices and listeners: Toward a model of voice perception. *Acoustics Today* 7**,** 7-14.

KREITEWOLF, J., GAUDRAIN, E. & VON KRIEGSTEIN, K. 2014. A neural mechanism for recognizing speech spoken by different speakers. . *NeuroImage,* 91**,** 375-385.

KRIEGESKORTE, N., FORMISANO, E., SORGER, B. & GOEBEL, R. 2007. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences,* 104**,** 20600-20605.

LANG, C. J., KNEIDL, O., HIELSCHER-FASTABEND, M. & HECKMANN, J. G. 2009. Voice recognition in aphasic and non-aphasic stroke patients. *Journal of neurology,* 2009/04/09.

LATINUS, M. & BELIN, P. 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology,* 2**,** 175.

LATINUS, M., MCALEER, P., BESTELMEYER, P. E. & BELIN, P. 2013. Norm-based coding of voice identity in human auditory cortex. *Current Biology,* 23**,** 1075-80.

LAVNER, Y., ROSENHOUSE, J. & GATH, I. 2001. The prototype model in speaker identification by human listeners. *IJST,* 4**,** 63-74.

LEE, Y., DUCHAINE, B., WILSON, H. R. & NAKAYAMA, K. 2010. Three cases of developmental prosopagnosia from one family: Detailed neuropsychological and psychophysical investigation of face processing. *Cortex,* 46**,** 949-964.

LEOPOLD, D. A., RHODES, G., MÜLLER, K. & JEFFERY, L. 2005. The dynamics of visual adaptation to faces. *Proceedings of the Royal Society of Biological Sciences,* 272**,** 897-904.

LEOPOLD, D. A., TOOLE, A. J. O., VETTER, T. & BLANZ, V. 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience,* 4**,** 89-94.

LIN, I. F., AGUS, T. R., SUIED, C., PRESSNITZER, D., YAMADA, T., KOMINE, Y., KATO, N. & KASHINO, M. 2016. Fast response to human voices in autism. *Nature Scientifc Reports,* 19**,** 26336.

LISSAUER, H. 1890. Ein Fall von Seelenblindheit nebst einem Beitrage zur Theorie derselben. *Archiv für Psychiatrie und Nervenkrankheiten,* 21**,** 222-270.

LIU, J., HARRIS, A. & KANWISHER, N. 2010. Perception of face parts and face configurations: An FMRI study. *Journal of Cognitive Neuroscience,* 22**,** 203-211.

LIU, R. R., CORROW, S. L., PANCAROGLU, R., DUCHAINE, B. & BARTON, J. 2015. The processing of voice identity in developmental prosopagnosia. *Cortex***,** 390-397.

LOFFLER, G., YOURGANOV, G., WILKINSON, F. & WILSON, H. R. 2005. fMRI evidence for the neural representation of faces. *Nature Neuroscience,* 8**,** 1386-1390.

LÓPEZ, S., RIERA, P., ASSANEO, M. F., EGUÍA, M., SIGMAN, M. & TREVISANA, M. A. 2013. Vocal caricatures reveal signatures of speaker identity. *Scientific Reports,* 3**,** 3407.

LUZZI, S., COCCIA, M., POLONARA, G., REVERBERI, C., CERAVOLO, G., SILVESTRINI, M., FRINGUELLI, F., BALDINELLI, S., PROVINCIALI, L. & GAINOTTI, G. 2017. Selective associative phonagnosia after right anterior temporal stroke. *Neuropsychologia*.

MAGUINNESS, C. & VON KRIEGSTEIN, K. 2017. Cross-modal processing of voices and faces in developmental prosopagnosia and developmental phonagnosia *Visual Cognition,* 25**,** 644-657.

MATHIAS, S. R. & VON KRIEGSTEIN, K. 2014. How do we recognise who is speaking? *Frontiers in Bioscience,* 6**,** 92-109.

MCCONACHIE, H. R. 1976. Developmental prosopagnosia. A single case report. *Cortex,* 12**,** 76-82.

MOROSAN, P., RADEMACHER, J., SCHLEICHER, A., AMUNTS, K., SCHORMANN, T. & ZILLES, K. 2001. Human Primary Auditory Cortex: Cytoarchitectonic Subdivisions and Mapping into a Spatial Reference System. *NeuroImage,* 13**,** 684-701.

MULLENNIX, J. W., ROSS, A., SMITH, C., KUYKENDALL, K., CONARD, J. & BARB, S. 2011. Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology,* 25**,** 29-34.

NATU, V. S., JIANG, F., NARVEKAR, A., KESHVARI, S., BLANZ, V. & O'TOOLE, A. J. 2010. Dissociable neural patterns of facial identity across changes in viewpoint. *Journal of Cognitive Neuroscience,* 22**,** 1570-1582.

NESTOR, A., PLAUT, D. C. & BEHRMANN, M. 2011. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences,* 108**,** 9998-10003.

NEUNER, F. & SCHWEINBERGER, S. R. 2000. Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition,* 44**,** 342-66.

NEWELL, F. N., CHIRORO, P. & VALENTINE, T. 1999. Recognizing unfamiliar faces: the effects of distinctiveness and view. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology,* 52**,** 509-534.

NISHIMURA, M., DOYLE, J., HUMPHREYS, K. & BEHRMANN, M. 2010. Probing the face-space of individuals with prosopagnosia. *Neuropsychologia,* 48**,** 1828-1841.
O'MAHONY, C. & NEWELL, F. N. 2012. Integration of faces and voices, but not faces and names, in person recognition. *British Journal of Psychology,* 103**,** 73-82.
PALERMO, R., RIVOLTA, D., WILSON, C. E. & JEFFERY, L. 2011. Adaptive face space coding in congenital prosopagnosia: typical figural aftereffects but abnormal identity aftereffects. *Neuropsychologia,* 49**,** 3801-3812.
PERNET, C. R., MCALEER, P., LATINUS, M., GORGOLEWSKI, K. J., CHAREST, I., BESTELMEYER, P. E., WATSON, R. H., FLEMING, D., CRABBE, F., VALDES-SOSA, M. & BELIN, P. 2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage,* 119**,** 164-74.
PERRACHIONE, T. K., DEL TUFO, S. N. & GABRIELI, J. D. 2011. Human voice recognition depends on language ability. *Science,* 333**,** 595.
PERRODIN, C., KAYSER, C., LOGOTHETIS, N. K. & PETKOV, C. I. 2012. Voice cells in the primate temporal lobe. *Current Biology,* 21**,** 1408-1415.
PETKOV, C. I. & VUONG, Q. C. 2013. Neuronal coding: The value in having an average voice. *Current Biology,* 23**,** R521-R523.
PITCHER, D., WALSH, V. & DUCHAINE, B. 2011. The role of the occipital face area in the cortical face perception network. *Experimental Brain Research,* 209**,** 481-493.
QUIROGA, R. Q. 2017. How do we recognize a face? *Cell,* 169**,** 975-977.
QUIROGA, R. Q., REDDY, L., KREIMAN, G., KOCH, C. & FRIED, I. 2005. Invariant visual representation by single neurons in the human brain. *Nature,* 435**,** 1102–1107.
RAJIMEHRA, R., YOUNG, J. C. & TOOTELLA, R. 2009. An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 1995–2000.
REMEZ, R. E., FELLOWES, J. M. & RUBIN, P. E. 1997. Talker identification based on phonetic information. *J.Exp.Psychol.Hum.Percept.Perform.,* 23**,** 651-666.
RHODES, G. & JAQUET, E. 2011. Aftereffects reveal that adaptive face-coding mechanisms are selective for race and sex. *In:* ADAMS, R. B., AMBADY, N., NAKAYAMA, K. & SHIMOJO, S. (eds.) *The science of social vision.* Oxford: UK: Oxford University Press.
ROBERTSON, D. M. C. & SCHWEINBERGER, S. R. 2010. The role of audiovisual asynchrony in person recognition. *Quarterly Journal of Experimental Psychology,* 63**,** 23-30.
ROSCH, E. H. 1973. Natural categories. *Cognitive Psychology,* 4**,** 328–350.
ROSWANDOWITZ, C., MAGUINNESS, C. & VON KRIEGSTEIN, K. 2018a. Deficits in voice-identity processing: acquired and developmental phonagnosia. *In:* BELIN, P. & FRUEHHOLZ, S. (eds.) *Oxford Handbook of Voice Perception.* Oxford, UK: Oxford University Press.
ROSWANDOWITZ, C., KAPPES, C., OBRIG, H. & VON KRIEGSTEIN, K. 2018b. Obligatory and facultative brain regions for voice-identity recognition. *Brain,* 141**,** 234–247.
ROSWANDOWITZ, C., MATHIAS, S. R., HINTZ, F., KREITEWOLF, J., SCHELINSKI, S. & VON KRIEGSTEIN, K. 2014. Two cases of selective developmental voice-recognition impairments. *Current Biology,* 24**,** 2348-2353.
ROSWANDOWITZ, C., SCHELINSKI, S. & VON KRIEGSTEIN, K. 2017. Developmental phonagnosia: Linking neural mechanisms with the behavioural phenotype. *NeuroImage,* 155**,** 97-112.

SCHALL, S., KIEBEL, S. J., MAESS, B. & VON KRIEGSTEIN, K. 2013. Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage,* 77**,** 237-45.
SCHALL, S., KIEBEL, S. J., MAESS, B. & VON KRIEGSTEIN, K. 2015. Voice identity recognition: functional division of the right STS and its behavioral relevance. *J Cogn Neurosci,* 27**,** 280-91.
SCHALL, S. & VON KRIEGSTEIN, K. 2014. Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. *PloS ONE,* 9**,** e86325.
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,* 11**,** 1812-1822.
SCHELINSKI, S., RIEDEL, P. & VON KRIEGSTEIN, K. 2014. Visual abiltites are important for auditory-only speech recognition: Evidence from autism spectrum disorder. *Neuropsychologia,* 65**,** 1-11.
SCHELINSKI, S., ROSWANDOWITZ, C. & VON KRIEGSTEIN, K. 2017. Voice identity processing in autism spectrum disorder. *Autism Research,* 10**,** 155-168.
SCHWEINBERGER, S. R. 2001. Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia,* 39**,** 921-36.
SCHWEINBERGER, S. R., KLOTH, N. & ROBERTSON, D. M. C. 2011a. Hearing facial identities: brain correlates of face-voice integration in person identification. *Cortex,* 47**,** 1026-1037.
SCHWEINBERGER, S. R., ROBERTSON, D. & KAUFMANN, J. M. 2007. Hearing facial identities. *Quaterly Journal of Experimental Psychology,* 60**,** 1446-1456.
SCHWEINBERGER, S. R., WALTHER, C., ZÄSKE, R. & KOVÁCS, G. 2011b. Neural correlates of adaptation to voice identity. *British Journal of Psychology,* 102**,** 748-764.
SHAH, N. J., MARSHALL, J. C., ZAFIRIS, O., SCHWAB, A., ZILLES, K., MARKOWITSCH, H. J. & FINK, G. R. 2001. The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. *Brain,* 124**,** 804-15.
SHEFFERT, S. M. & OLSON, E. 2004. Audiovisual speech facilitates voice learning. *Perception and Psychophysics,* 66**,** 352-62.
SHEFFERT, S. M., PISONI, D. B., FELLOWES, J. M. & REMEZ, R. E. 2002. Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance,* 28**,** 1447 %@ 1939-1277.
SHILOWICH, B. E. & BIEDERMAN, I. 2016. An estimate of the prevalence of developmental phonagnosia. *Brain and Language,* 159**,** 84-91.
SIDTIS, D. & KREIMAN, J. 2012. In the beginning was the familiar voice: personally familiar voices in the evolutionary and contemporary biology of communication. *Integr Psychol Behav Sci,* 46**,** 146-159.
SMITH, H. M. J., DUNN, A. K., BAGULEY, T. & STACEY, P. C. 2016a. Concordant Cues in Faces and Voices: Testing the Backup Signal Hypothesis. *Evolutionary Psychology,* 14**,** 1-10.
SMITH, H. M. J., DUNN, A. K., BAGULEY, T. & STACEY, P. C. 2016b. Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics,* 78**,** 868-879.

SØRENSEN, M., H 2012. Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language & the Law,* 19**,** 145-158.

STEVENAGE, S. & NEIL, G. 2014. Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica,* 54**,** 266-281.

STEVENAGE, S. V. 2018. Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, 116, 162-178.

STEVENAGE, S. V., CLARKE, G. & MCNEILL, A. 2012. The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology,* 24**,** 647-653.

SUSILO, T. & DUCHAINE, B. 2013. Advances in developmental prosopagnosia research *Current Opinion in Neurobiology,* 23**,** 423-429.

SUSILO, T., MCKONE, E., DENNETT, H., DARKE, H., PALERMO, R., HALL, A., PIDCOCK, M., DAWEL, A., JEFFERY, L., WILSON, C. E. & RHODES, G. 2010. Face recognition impairments despite normal holistic processing and face space coding: evidence from a case of developmental. *Cognitive Neuropsychology,* 27**,** 636-664.

VALENTINE, T. 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A : Human Experimental Psychology,* 43**,** 161-204.

VAN LANCKER, D. & KREIMAN, J. 1987. Voice discrimination and recognition are separate abilities. *Neuropsychologia,* 25**,** 829-834.

VAN LANCKER, D., KREIMAN, J. & EMMOREY, K. 1985a. Familiar Voice Recognition: Patterns and Parameters. Part I: Recognition of Backward Voices. *Journal of Phonetics,* 13**,** 19-38.

VAN LANCKER, D., KREIMAN, J. & WICKENS, T. D. 1985b. Familiar Voice Recognition: Patterns and Parameters. Part II: Recognition of Rate-Altered Voices. *Journal of Phonetics,* 13**,** 39-52.

VAN LANCKER, D. R. & CANTER, J. G. 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. & CUMMINGS, J. 1989. Voice perception deficits: neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology,* 11**,** 665-674.

VIDA, M. D., NESTOR, A., PLAUT, D. C. & BEHRMANN, M. 2016. The Spatiotemporal Dynamics of Similarity-based Neural Representations of Facial Identity. *Proceedings of the National Academy of Sciences,* 114**,** 388-393.

VON KRIEGSTEIN, K. 2012. A multisensory perspective on human auditory communication. *In:* MURRAY, M. M. & WALLACE, M. T. (eds.) *The Neural Bases of Multisensory Processes.* Boca Raton (FL): CRC Press/Taylor & Francis.

VON KRIEGSTEIN, K., DOGAN, O., GRUTER, M., GIRAUD, A. L., KELL, C. A., GRUTER, T., KLEINSCHMIDT, A. & KIEBEL, S. J. 2008. Simulation of talking faces in the human brain improves auditory speech recognition. *Proc.Natl.Acad.Sci.U.S.A,* 105**,** 6747-6752.

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.

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. & GIRAUD, A. L. 2006. Implicit multisensory associations influence voice recognition. *PLoS Biology,* 4**,** e326.

VON KRIEGSTEIN, K., KLEINSCHMIDT, A. & GIRAUD, A. L. 2006a. 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., PATTERSON, R. D., IVES, D. T. & GRIFFITHS, T. D. 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Curr.Biol.,* 17**,** 1123-1128.

VON KRIEGSTEIN, K., SMITH, D. R., PATTERSON, R. D., KIEBEL, S. J. & GRIFFITHS, T. D. 2010. How the human brain recognizes speech in the context of changing speakers. *J Neurosci,* 30**,** 629-38.

VON KRIEGSTEIN, K., WARREN, J. D., IVES, D. T., PATTERSON, R. D. & GRIFFITHS, T. D. 2006b. Processing the acoustic effect of size in speech sounds. *Neuroimage.,* 32**,** 368-375.

WARREN, J., SCOTT, S., PRICE, C. & GRIFFITHS, T. 2006. Human brain mechanisms for the early analysis of voices. *Neuroimage,* 31**,** 1389-1397.

WARRINGTON, E. K. 1975. The selective impairment of semantic memory. *Quaterly Journal of Experimental Psychology,* 27**,** 635-657.

WARRINGTON, E. K. & SHALLICE, T. 1984. Category Specific Semantic Impairments. *Brain,* 107**,** 829-854.

WEIBERT, K. & ANDREWS, T. J. 2015. Activity in the right fusiform face area predicts the behavioural advantage for the perception of familiar faces. *Neuropsychologia,* 75**,** 588-596.

XU, X., 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.

XU, X., YUE, X., LESCROART, M. D., BIEDERMAN, I. & KIM, J. G. 2009. Adaptation in the fusiform face area (FFA): Image or person? *Vision Research,* 49**,** 2800-2807.

YANG, H., SUSILO, T. & DUCHAINE, B. 2016. The Anterior Temporal Face Area Contains Invariant Representations of Face Identity That Can Persist Despite the Loss of Right FFA and OFA. *Cerebral Cortex,* 26**,** 1096-1107.

YOVEL, G. & BELIN, P. 2013. A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences,* 17**,** 263-271.

ZÄSKE, R., AWWAD SHIEKH HASAN, B. & BELIN, P. 2017. It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex,* 94**,** 100-112.

ZÄSKE, R., MÜHL, C. & SCHWEINBERGER, S. R. 2015. Benefits for voice learning caused by concurrent faces develop over time. *PLoS ONE,* 10**,** e0143151.

ZÄSKE, R., SCHWEINBERGER, S. R. & KAWAHARA, H. 2010. Voice aftereffects of adaptation to speaker identity. *Hearing Research,* 268**,** 38-45.
ZÄSKE, R., VOLBERG, G., KOVACS, G. & SCHWEINBERGER, S. R. 2014. Electrophysiological correlates of voice learning and recognition. *The Journal of Neuroscience,* 34**,** 10821-10831.