British Journal of Psychology (2011), 102, 711–725

The British Psychological Society

www.wileyonlinelibrary.com

Understanding Voice Perception

Pascal Belin[1,2]*, Patricia E. G. Bestelmeyer[1], Marianne Latinus[1] and Rebecca Watson[1]

[1]Voice Neurocognition Laboratory, Institute of Neuroscience and Psychology, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK
[2]International Laboratories for Brain, Music and Sound (BRAMS), Université de Montréal & McGill University, Montreal, Quebec, Canada

Voices carry large amounts of socially relevant information on persons, much like 'auditory faces'. Following Bruce and Young (1986)'s seminal model of face perception, we propose that the cerebral processing of vocal information is organized in interacting but functionally dissociable pathways for processing the three main types of vocal information: speech, identity, and affect. The predictions of the 'auditory face' model of voice perception are reviewed in the light of recent clinical, psychological, and neuroimaging evidence.

Face and voice signals, despite the different nature of their physical structure (light reflections hitting the retina in the eye vs. pressure waves inducing vibrations of the basilar membrane in the ear), carry highly similar types of socially relevant information. Both contain linguistic information (phonemes for voice, viseme for faces, i.e., representational units used to classify speech sounds in the visual domain) but also relevant information on a range of personal biological characteristics (gender, age, size, identity, affective state, fitness . . .). From this angle, the voice can be considered as an 'auditory face'. The nature of the computational complexity imposed on the brain in processing these signals (categorization, invariance, identification . . .) is thus very similar across the two modalities, at least at higher level, relatively abstract stages of processing. Solving these similar problems using a similar neuronal implementation would seem a parsimonious principle of cerebral organization (Ellis, 1989).

We conceptualized the above notion by extending Bruce & Young's (1986) seminal model of cerebral face processing (Bruce & Young, 1986; see also Young & Bruce, 2011; Burton, Jenkins & Schweinberger, 2011) and proposed a similar functional architecture for voice processing (Belin, Fecteau, & Bedard, 2004), as already suggested by several authors (Burton, Bruce, & Johnston, 1990; Ellis, 1989).

Authors are listed in alphabetical order.
**Correspondence should be addressed to Pascal Belin, 58 Hillhead street, Glasgow G12 8QB, UK (e-mail: p.belin@psy.gla.ac.uk).*

DOI:10.1111/j.2044-8295.2011.02041.x

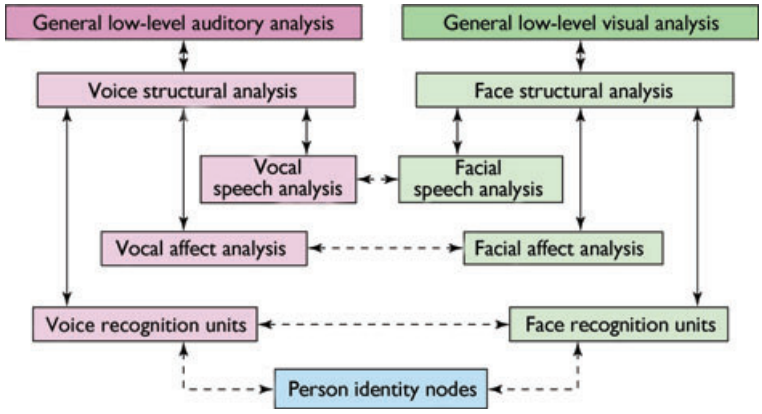

Figure 1. A model of voice perception. Reproduced from Belin *et al.* (2004). After a stage of voice structural encoding restricted to vocal sounds, three partially dissociable functional pathways are proposed to process the three main types of vocal information: speech, identity, and affect. These pathways are analogous to and interacting with equivalent functional pathways involved in facial processing.

According to the 'auditory face' model of voice processing (Figure 1), an initial low-level analysis occurs in sub-cortical nuclei and core regions of auditory cortex, after which voices are processed in a voice-specific stage of 'structural encoding'. At this stage, the three main types of vocal information are then extracted and further processed in three interacting, but partially dissociable functional pathways: (1) a pathway for analysis of speech information, involving anterior and posterior superior temporal sulcus (STS) as well as inferior prefrontal regions and pre-motor cortex predominantly in the left hemisphere; (2) a pathway for analysis of vocal affective information, involving temporo-medial regions, anterior insula, and amygdala and inferior prefrontal regions predominantly in the right hemisphere; (3) a pathway for analysis of vocal identity, involving 'voice recognition units' – probably instantiated in regions of the right anterior STS – each activated by one of the voices known to the person (Figure 1). These three functional pathways are proposed to interact with each other during normal processing. They are also proposed to interact with homologous pathways in the face-processing architecture during audio-visual face/voice integration (Campanella & Belin, 2007).

It has to be stressed that this model does not propose that all aspects of face and voice processing are exactly similar. For instance, there is evidence that faces provide more reliable identity information on familiar persons than do voices (Bredart, Barsics, & Hanley, 2009). Also, it has been suggested that whereas sex and identity information appear to be processed independently for faces, their processing might not be independent for voices (Burton, & Bonner, 2004). Nevertheless, we hope the model proposes a useful heuristic to guide research into the cerebral mechanisms of voice processing and its interactions with face processing.

Are voices special?

The 'structural' encoding stage is, by analogy with Bruce and Young, viewed as being accessed only by vocal stimuli. It is at this stage of the functional architecture that a vocal sound would be identified as such, that is, has been produced by a human vocal apparatus. From that stage onwards, irrespective of the exact nature of the information

being the attention' s focus, voice stimuli are proposed to recruit processes not activated by other, non-vocal sounds. In other words, voices are 'special' for the brain.

In the visual domain, combined evidence for the 'specialness' of faces –although the issue is still a matter of debate (Gauthier & Bukach, 2007; McKone, Kanwisher, & Duchaine, 2007) – is provided by three different experimental sources: cognitive psychology, clinical neuroscience, and neuroimaging. Briefly (the issue is reviewed at length elsewhere), cognitive psychology experiments reveal phenomena such as the face-inversion effect, or the face-composite effect, that are unique to, or more marked for, faces than other objects; clinical neuroscience describes patients with selective impairments in the identification of faces (prosopagnosia); neuroimaging techniques including functional magnetic resonance imaging (fMRI), event-related potentials (ERPs), magnetoencephalography, and depth-electrode recordings in humans, but also single-cell and local field potential recordings and fMRI in primates, highlight regions of visual and association cortex with high selectivity for faces, some consisting of mostly face-selective neurons.

Although it is not yet as strong and convincing as for faces, similar evidence for voices is accumulating. Evidence for cognitive phenomena specific to voice processing is still elusive, but converging clinical and neuroimaging evidence suggests there are indeed voice-selective cerebral processes. fMRI studies by our group and several others have demonstrated the existence of voice-selective neuronal populations (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009; Gervais *et al.*, 2004; Grandjean *et al.*, 2005; Linden *et al.*, 2011): these voice-selective regions of cortex (the 'temporal voice areas', TVA) are located bilaterally along the mid and anterior parts of superior temporal gyrus (STG)/STS (Figure 2). They show greater blood oxygenation (BOLD signal) in response to vocal sounds than to non-vocal sounds from natural sources, or acoustical controls such as amplitude-modulated noise or scrambled voices. Although it is particularly strong for speech sounds, the voice-selective response is also observed for non-speech sounds (Belin, Zatorre, & Ahad, 2002; Charest *et al.*, 2009), showing that the TVA, particularly in the right hemisphere, are not just interested in processing the linguistic content of voice.

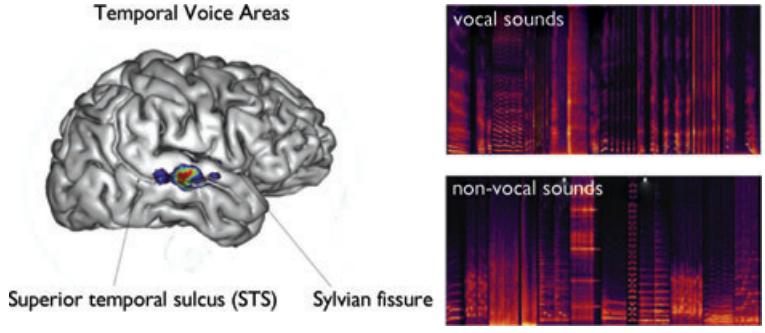

Figure 2. Voice-selective cerebral activity. The contrast of cerebral activity measured in the adult brain by functional magnetic resonance imaging (fMRI) in response to auditory stimulation with vocal versus non-vocal sounds (stimuli available at http://vnl.psy.gla.ac.uk) highlights voice selective TVA with greater activity in response to the vocal sounds. The TVA (shown here in an individual young adult subject) are mostly located along the middle and anterior parts of the superior temporal sulcus (STS) bilaterally. Reproduced from Belin & Grosbras (2010).

Recent evidence using near-infrared spectroscopy shows that voice-selective responses are already present in 7-month-old infants (Grossman, Oberecker, Koch, & Friederici, 2010), long before speech is fully developed but at a time when voice discrimination and recognition abilities are well established. This finding demonstrates that these voice-selective responses are not exclusively related to speech processing and suggests an early developmental time-course of voice processing. Similarly, voice-selective areas have recently been observed in the macaque brain (Petkov *et al.*, 2008), demonstrating that they are phylogenetically ancient (probably already present in the common ancestor of humans and macaques some 30 million years ago), and indicating that when speech appeared some tens of thousands of years ago, our ancestors were already equipped with some rudiments of a cerebral machinery for processing vocal information.

Electrophysiological techniques also suggest voice-specific cerebral activity. ERPs differences in response to vocal versus non-vocal sounds have been observed at latencies of around 320 ms (the 'voice-sensitive response' or 'VSR') (Levy, Granot, & Bentin, 2001, 2003), and, more recently, at the shorter, and more compatible with face-processing evidence, latency of around 200 ms (Charest *et al.*, 2009; De Lucia, Clarke, & Murray, 2010).

The above evidence leaves, however, important questions unresolved: does activation of the TVA and other voice-sensitive/selective areas reflect genuine processing of vocal information, or is it simply a by-product of their particular acoustical structure? Several experiments have used a variety of acoustical control stimuli, but the possibility that these did not control for all possible (or combinations of) features cannot be excluded. Recently, Leaver and Rauschecker (2010) provided some evidence that a large part of the selectivity for voices was indeed explained by acoustical properties as this selectivity disappeared when variance accounted for by acoustical properties was included in the statistical model (Leaver & Rauschecker, 2010). Also, does voice-selective activity reflect the expertise of normal listeners with voices rather than a response to voice *per se*. Experiments testing the possible causal link between these activations and voice processing remain to be conducted, as are studies investigating possible voice-dedicated cognitive mechanisms.

Voice recognition

Voice recognition differs from the discrimination of voices from non-vocal cues in that it requires a fine tuned analysis of the vocal structure. All human voices share a similar basic organization; slight variations of acoustic parameters around a mean determine voice uniqueness, that is, the 'vocal signature' of an individual. Bruce and Young's (1986) model and Belin *et al.*'s (2004) adapted version to voices predict the existence of specific identity-related pathways (Figure 1). This prediction has received support from clinical studies (Garrido *et al.*, 2009; Hailstone, Crutch, Vestergaard, Patterson, & Warren, 2010; Van Lancker & Canter, 1982; Van Lancker, Cummings, Kreiman, & Dobkin, 1988), behavioural studies in healthy participants (Kreiman & Gerratt, 1998) as well as from a number of neuroimaging studies.

In analogy to the face literature, the term 'phonagnosia' has been introduced by Van Lancker and Canter in 1982 to describe individuals with a deficit in voice recognition. Voice recognition was impaired following damage to the right hemisphere, while left hemisphere lesions generally induced aphasia with preserved voice recognition

abilities (Van Lancker & Canter, 1982) demonstrating a double dissociation between voice-identity and speech processing. In follow-up studies of phonagnosic patients, they further revealed the dissociation between voice recognition (of familiar individuals) and voice discrimination, that is, the ability to distinguish between two unfamiliar voices (Van Lancker, Cummings, Kreiman, & Dobkin, 1988; Van Lancker & Kreiman, 1987). Voice recognition (i.e., recognition of a familiar voice) was impaired by lesion in the right parietal cortex, while a deficit in voice discrimination (i.e., perceiving that two vocal sounds are from a same unfamiliar speaker) was associated with lesion of the temporal lobe of either hemisphere.

Investigations of the interplay between the processing of vocal identity and of vocal affective information are limited, as phonagnosic patients are rarely tested on their ability to recognize vocal emotion. Yet, clinical and neuroimaging evidence tends to confirm the dissociation between identity and emotion processing of voices predicted by the voice perception model. The processing of emotional content of voices is not necessarily impaired in phonagnosic subjects (Garrido *et al.*, 2009; Hailstone *et al.*, 2010). Conversely, patients with ventro-frontal damage inducing impaired vocal emotion processing are not necessarily impaired at voice discrimination, although a non-negligible proportion (3 out of 12) present this pattern (Hornak, Rolls, & Wade, 1996).

However, despite evidence for a separate pathway for the processing of vocal identity, several reports in healthy participants challenge this view by revealing interactions between speech and identity processing. For example, voice familiarity influences subjects' responses in linguistic tasks (Nygaard & Pisoni, 1998; Pisoni, 1993). This suggests that although potentially impaired independently of one another, the functional pathways for speech and speaker identity processing are interacting during normal behaviour.

Several neuroimaging studies on healthy young adults have focused on identifying the cerebral network involved in the perception of voice identity. They provide evidence that the mid superior temporal cortex (STC), overlapping with the TVA, is involved in an acoustical processing of voices regardless of familiarity (Andics *et al.*, 2010; Charest, Pernet, Crabbe, & Belin, 2009; Latinus, Crabbe, & Belin, 2009).

The more anterior regions of the TVA extending towards the temporal pole (TP) appear to be involved in an invariant representation of voice identity, regardless of voice familiarity in both humans (unfamiliar voices [Belin & Zatorre, 2003; Formisano, De Martino, Bonte, & Goebel, 2008; Imaizumi *et al.*, 1997; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003]; familiar voices [Andics *et al.*, 2010; Nakamura *et al.*, 2001]) and macaques (Petkov *et al.*, 2008). In a recent study, looking at the processing of voices before and after voice learning, the right superior TP was found to be sensitive to acoustic information only for unfamiliar voices; its activity overall decreased for familiar voices (Latinus *et al.*, 2009). The locus of activation in the TP differs between studies using familiar and unfamiliar voices (Figure 2); we suggest that, in the right hemisphere, the superior TP is involved in an acoustic-based representation of unfamiliar voices, while the inferior part of the TP is involved in a non-verbal representation of person-related semantic knowledge (Gorno-Tempini *et al.*, 1998; Hailstone *et al.*, 2010), suggesting it could be the neural equivalent of the person identification node (PIN, Figure 1).

Areas outside the TVA also show sensitivity to voice familiarity. Activity in bilateral inferior frontal cortex (IFC) is larger for unfamiliar voices than for familiar voices (Stevens, 2004; von Kriegstein & Giraud, 2004). The larger IFC activation for unfamiliar voices may reflect its involvement in processing acoustic information in previously heard voices (Andics *et al.*, 2010; Latinus *et al.*, 2009; von Kriegstein & Giraud, 2006). Greater activity for familiar voices than unfamiliar is found in right frontal, bilateral parietal cortices,

posterior STC as well as in the fusiform gyrus (FG), an area primarily described as sensitive to faces (Kanwisher, McDermott, & Chun, 1997). Activation of the bilateral parietal cortices and posterior STC, often described as a multi- or hetero-modal areas (Calvert, Campbell, & Brammer, 2000; Sestieri *et al.*, 2006), and FG are generally reported in studies using familiar voices associated with a face either because of lab training or due to familiarity itself (Andics *et al.*, 2010; von Kriegstein & Giraud, 2006). Conversely, when voice learning is achieved using voice/name association, only the IFC shows sensitivity to identity processing of voices (Latinus *et al.*, 2009). Thus, familiar voice recognition activates a range of brain areas among which the temporal cortex and IFC is likely to be involved in voice processing *per se*, while other areas appear involved in a multimodal representation of person identity or in retrieval of visual information when hearing a familiar voice.

Although progress has been made in our understanding of voice recognition and despite much research aiming at identifying acoustic parameters underlying speaker recognition, the format of voice-identity representation is still relatively unknown and the acoustic components essential to voice recognition are still unclear. A recent multidimensional scaling study proposed that voices are represented in a multidimensional 'voice space' with two main dimensions well approximated by measures of the fundamental frequency of phonation (f0) and of formant frequencies (Baumann & Belin, 2010). Many other acoustic parameters have been implicated in speaker recognition both in humans and macaques, among them the f0 range, formant structure or specific formants (Baumann & Belin, 2010; Murry & Singh, 1980), temporal dynamics (Ghazanfar & Hauser, 2001; Schweinberger, 2001), and other, less quantifiable, parameters such as accent, speech variation etc. (Belin *et al.*, 2004).

A large number of acoustic parameters appear to be implicated in speaker recognition; yet, listeners are efficient at extracting invariant features in the vocal signal used to recognize that person from novel utterances (Papcun, Kreiman, & Davis, 1989; Schweinberger, Herholz, & Sommer, 1997), suggesting that voice identity is carried by a combination of different acoustical factors. One hypothesis reconciling these different observations is that voice identity is encoded relative to a voice prototype (Bruckert *et al.*, 2010; Kreiman, Gerratt, Precoda, & Berke, 1992; Lattner & Friederici, 2003; Papcun *et al.*, 1989). This hypothesis was supported in a recent study (Latinus & Belin, 2010) showing larger auditory perceptual aftereffects (Schweinberger *et al.*, 2008; Zäske, Schweinberger, & Kawahara, 2010) for anti-voice stimuli, caricature of the average voice ($N = 16$) relative to an individual voice, than other stimuli type. This result was interpreted as indicative of a special status of the average voice in representing voice identity (Leopold, O'Toole, Vetter, & Blanz, 2001), thus providing the first evidence that voice identity is encoded in reference to a prototypical or average voice (Figure 3). Further investigations are needed in order to apprehend the acoustical information stored in that vocal prototype.

Person recognition is not only voice recognition but also face recognition. Clinical evidence showing that impairment in voice recognition is often associated with prosopagnosia, impairment in face recognition (Leopold *et al.*, 2001; Van Lancker & Canter, 1982), and recent behavioural evidence showing auditory perceptual aftereffects following the repeated presentation of a face stimulus (Zäske *et al.*, 2010) support the idea of a supramodal representation of person identity. Neuroimaging studies of faces and voices suggest that the posterior STS and the inferior TP of the right hemisphere could be involved in processing supramodal information related to person recognition, a possibility currently under investigation.

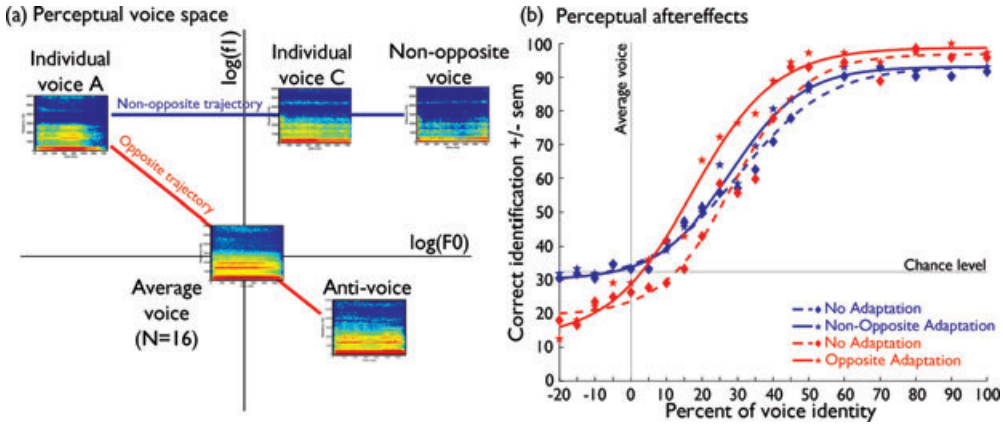

Figure 3. A prototype-based perceptual voice space. (a) Individual voices (depicted by their spectrogram) are positioned in a multidimensional voice space with first two dimensions corresponding to f0 and formant frequencies, centred on an average, prototypical voice. Anti-voice stimuli can be generated by interpolating an individual voice with the average voice. (b) Perceptual categorization of a voice-identity continuum, at baseline (dotted lines) and after adaptation (continuous lines). Anti-voice adaptors yield stronger identity aftereffects (red line) than non-opposite adaptors (blue line), a result compatible with prototype-based, but not exemplar-based, representation of vocal identity.

Affect perception

Belin *et al.*'s (2004) adaptation of the Bruce and Young model to voice perception also proposes a functional pathway dedicated to the processing of vocal affect. Evidence for such a pathway comes from behavioural, neuropsychological, and neuroimaging studies. Test materials examining the properties of affective processing typically involve recordings of speech with different emotional intonation. The problem with these stimuli is the possible interaction between affective and semantic content carried by speech prosody as well as these stimuli being language specific so that they cannot be compared cross-culturally. To minimize this interaction, studies have used meaningless sentences composed of pseudo-words that are spoken in various emotional tones or by using non-linguistic verbalizations such as laughter or screams of fear. The Montreal Affective Voices is a database of such non-linguistic verbalizations and consists of 90 vocal expressions of anger, disgust, fear, pain, sadness, pleasure, surprise, and happiness (Belin, Fillion-Bilodeau, & Gosselin, 2008; http://vnl.psy.gla.ac.uk).

Adaptation paradigms have been helpful in furthering our understanding of how sensory signals are coded and organized in the brain (Webster, Kaping, Mizokami, & Duhamel, 2004) and have lent support to a dedicated functional pathway dealing with affective information conveyed in the face and voice. Adaptation refers to a process during which continued stimulation results in a biased perception towards opposite features of the adapting stimulus. Research using adaptation has revealed neural populations tuned to respond to specific stimulus attributes by isolating and subsequently distorting the perception of these attributes (Bestelmeyer, Rouger, DeBruine, & Belin, 2010; Grill-Spector *et al.*, 1999; Winston, Henson, Fine-Goulden, & Dolan, 2004). Bestelmeyer *et al.* (2010) were the first to examine whether the processing of vocal affect is malleable by means of adaptation to angry and fearful non-linguistic vocalizations. Adaptation to angry vocalizations caused voices drawn from an anger–fear morphed

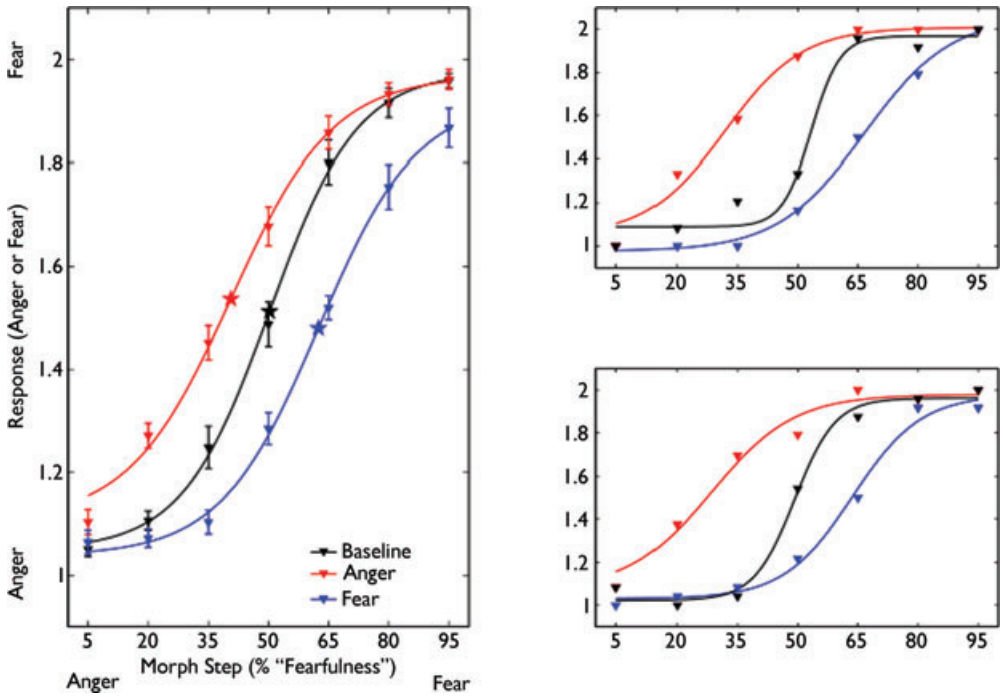

Figure 4. Adaptation of perceived voice affect. Psychophysical functions for three adaptation conditions: baseline (black), adaptation to anger (red), and fear (blue). Left enlarged graph displays the grand average of all participants; the two graphs to the right display individual participants. Stars coloured correspondingly for each condition indicate the PSE. Error bars represent standard error of the mean (SEM). Reproduced from Bestelmeyer *et al.* (2010).

continuum to be perceived as less angry and more fearful, while adaptation to fearful vocalizations elicited opposite aftereffects (Figure 4). These adaptation effects could not be solely explained by low-level adaptation to acoustical characteristics of the adaptors but were due to higher-level adaptation of neural representations of vocal affect (Bestelmeyer *et al.*, 2010). This study also shows that vocal affect perception can be isolated using adaptation and suggests that the existence of a pathway dedicated to vocal affect is possible.

Evidence from neuropsychology suggests that lesions in the right hemisphere are more detrimental to the recognition of vocal affect than lesions to the left side of the brain (Hornak *et al.*, 1996; D. Van Lancker & Sidtis, 1992). For example, Heilman, Scholes, and Watson (1975) report a clear double dissociation in which patients with damage to the right hemisphere had difficulty judging the emotion expressed in a sentence while content perception was unaffected. In contrast, patients with left hemisphere damage were unable to judge the content of the sentence, but their perception of the affect expressed in the sentence was unaffected (Heilman *et al.*, 1975). Similarly, early research using fMRI on emotional prosody classification has shown that the right hemisphere is particularly involved (Buchanan *et al.*, 2000; Morris, Scott, & Dolan, 1999; Rama *et al.*, 2001). Recent studies confirm the right liberalized activity in mid temporal gyrus (MTG) and STG (Ethofer, Anders, Erb *et al.*, 2006; Grandjean *et al.*, 2005; Mitchell, Elliott, Barry, Cruttenden, & Woodruff, 2003; Wildgruber *et al.*, 2004). This activation seems

relatively independent of attentional demands (Ethofer, Anders, Wiethoff *et al.*, 2006) and low-level acoustic features such as frequency and amplitude of the sounds (Grandjean *et al.*, 2005).

The idea that vocal emotional comprehension is specific to the right hemisphere is, however, clearly oversimplified. Additional neuroimaging studies have painted a more complex picture in which a more distributed, bilateral neural network is engaged when processing emotional prosody. Although the activity elicited in response to emotional prosody is often stronger on the right, bilateral TVA are typically active during the processing of affective compared to neutral vocalizations. In fact, Ethofer *et al.* (2009) have shown recently that each tested emotional category (anger, sadness, neutral, relief, joy) was encoded in spatially distinct parts of the voice-sensitive areas (Ethofer *et al.*, 2009). In addition to the voice-sensitive areas, regions not included in auditory cortex such as bilateral orbitofrontal cortices and inferior frontal cortices are active during the processing of emotional prosody and respond particularly during emotion classification tasks compared to orthogonal tasks (Ethofer, Anders, Wiethoff *et al.*, 2006; Imaizumi *et al.*, 1997; Wildgruber *et al.*, 2004; Wildgruber, Pihan, Ackermann, Erb, & Grodd, 2002). Some studies also demonstrate that the processing of vocal affect involves subcortical structures such as the basal ganglia (Pell & Leonard, 2003) and amygdala (Fecteau, Belin, Joanette, & Armony, 2007; Leitman *et al.*, 2010; Morris *et al.*, 1999; Phillips *et al.*, 1998; Sander & Scheich, 2005).

Very similar to Bruce and Young's (1986) face perception model Schirmer and Kotz (2006) integrate the areas reported by previous studies from neuropsychology and neuroimaging by proposing a three-step model for the understanding of emotional prosody. The first stage consists of a low-level acoustic analysis in bilateral auditory cortices. These areas then project to STC for more complex processing in which emotionally salient information is synthesized into an emotional 'Gestalt' or acoustic object. STS and STG then feed into frontal areas for higher-order cognition (e.g., evaluative judgements of emotional prosody). This model, largely based on research utilizing emotional speech, is an important and informative step towards understanding the functional pathway dedicated to vocal affect perception (Schirmer & Kotz, 2006).

Face/voice audio-visual integration

Understanding of both facial and vocal information plays a crucial role in our interpersonal interactions. Although anatomically distinct, voice- and face-processing areas usually act in parallel and are assumed to communicate, facilitating our social responses. Indeed, almost from birth both faces and voices pervade our perceptual experience, and communication between them is crucial for the acquisition of both linguistic and social skills. Integrating these two sources of information is advantageous as it allows our brain to exploit redundancies between face and voice and combine non-redundant, complementary cues to maximize information gathered from the two modalities (Calvert, 2001; Campanella & Belin, 2007). However, despite its importance, audio-visual integration of person-related information has received comparably little attention in comparison to information processing from separate modalities: it is only recently that investigation has focused on integrative mechanisms, particularly with regards to paralinguistic processing.

Belin *et al.*'s (2004) model of voice perception provides a way of understanding vocal processing (Figure 1) in which each of the pathways is not only be seen as working in parallel with the others, but also communicating with the homologous pathways

in the face-processing network to allow integration of speech, affect, and identity information. Undoubtedly, the main focus within the field of audio-visual integration research has been with regards to speech perception. It appears we integrate speech information from the face and voice before we are a year old (Kuhl & Meltzoff, 1982). Integration of facial and vocal speech is associated behaviourally with both facilitation (speech intelligibility enhancements when the speaker's face is visible) and interference (decreases in identification performance in incongruent conditions). In the 'McGurk effect', an incongruent phoneme–viseme pairing can be strong enough to provoke an illusory percept (McGurk & MacDonald, 1976).

However, face/voice integration involves more than processing of speech information: the face and voice are also rich in paralinguistic information, such as identity and emotion. Clear evidence suggests that healthy individuals are able to combine facial and vocal information in order to decide upon the identity of a person. Visual information has been found to aid recognition of the voice of the same individual, indicating a cross-modal facilitation effect comparable to that of audiovisual speech integration (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Schweinberger, Robertson, & Kaufmann, 2007). With regards to affect perception, various studies have demonstrated behaviourally that congruent affective information expressed in the face and voice facilitates the categorization of such information (e.g., Collignon *et al.*, 2008; de Gelder & Vroomen, 2000; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007). Specifically, these authors have observed faster categorization in bimodal, as opposed to unimodal conditions.

A central question within the field of audio-visual integration is whether face-/voice-identity integration requires a supramodal stage of cortical processing (that may correspond to 'person identity nodes (PINs)' (Ellis, Jones, & Mosdell, 1997)) or is mediated by crosstalk between 'unimodal' auditory- and visual-processing systems. This alternative is indicated in our model by the presence of both direct links between the face- and voice-identity pathways and indirect links via a supramodal stage of processing. Indeed, studies have suggested a number of structures (e.g., STS, amygdala, superior colliculus) may work as supramodal, multimodal processors, and there is evidence that unimodal sensory cortices have integrative mechanisms that respond to a supramodal stage of processing (Joassin, Maurage, Bruyer, Crommelinck, & Campanella, 2004; Joassin *et al.*, in press). Other studies have indirectly demonstrated direct crosstalk, via fMRI evidence of increased functional coupling between the Fusiform Face Area (FFA) and the voice-selective TVA of right mid-STS during (unimodal) familiar speaker recognition (von Kriegstein, Kleinschmidt, & Giraud, 2005). However, recent reports support the existence of a cerebral integrative network composed of both unimodal and multimodal regions, which sustain different aspects of integration such as sensory inputs processing, attention, and memory. In such a network, hetoromodal areas work in parallel and influence each other, as opposed to being the last stage in a multimodal 'hierarchical framework' (Noppeney, Ostwald, & Werner, 2010). In particular, the posterior part of the STS (pSTS) has emerged as a structure that plays a key role in integrating face and voice information. This amodal 'convergence' zone receives projections from the sensory cortex and has increased activity for bimodal, in comparison to unimodal, presentation of stimuli. Activity in this area also differs between congruent and incongruent information presentation (e.g., Jones & Callan, 2003).

Conclusion

Available evidence from cognitive psychology, clinical neuroscience, and neuroimaging thus largely supports the notion of similar and interacting functional architectures for

the cerebral processing of socially relevant information in faces and voices. Some of the predictions of this model (e.g., dissociation between receptive aphasia and phonagnosia) have been tested, but several others (e.g., direct communication between face- and voice-selective areas during identity processing) remain to be tested. A number of additional questions emerge and require further investigations: how does the voice-processing network develop, and what is the balance of genetically programmed versus environmental factors in this development? Do analogous functional architectures exist in the brain of other primates, or mammals, and what is their degree of similarity with the human one? Does the proposed architecture for processing face and voice information bring useful new knowledge potentially usable in the growing industry of automated person perception? We hope the proposed framework, inspired by the pioneering work of Bruce and Young (1986) and of others (Ellis, 1989; Ellis *et al.*, 1997), will provide a useful heuristic in addressing these outstanding questions.

References

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *Neuroimage*, *52*, 1528–1540.

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

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129–135.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The 'Montreal Affective Voices': A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavioural Brain Research*, *40*, 531–539.

Belin, P., & Grosbras, M. H. (2010). Before speech: Cerebral voice processing in infants. *Neuron*, *65*, 852–858.

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.

Bestelmeyer, P., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, *117*, 217–223.

Bredart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology*, *21*, 1013–1021.

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

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., *et al*. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*, 116–120.

Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., & Jäncke, L. (2000). Recognition of emotional prosody and verbal components of spoken language: An fMRI study. *Cognitive Brain Research*, *9*, 227–238.

Burton, A. M., & Bonner, L. (2004). Familiarity influences judgments of sex: The case of voice recognition. *Perception*, *33*, 747–752.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, 361–380.

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

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, (12) 1110–1123.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*, 535–543.

Charest, I., Pernet, C., Crabbe, F., & Belin, P. (2009). *Investigating the representation of voice gender using a continuous carry-over fMRI design*. Paper presented at the Human Brain Mapping Conference, San Francisco, June 2009.

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., *et al*. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10*, 127.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., *et al*. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289–311.

De Lucia, M., Clarke, S., & Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience*, *30*, 11210–11221.

Ellis, A. W. (1989). Neuro-cognitive processing of faces and voices. In A. W. Young & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 207–215): Elsevier Science Publishers B.V: North-Holland, The Netherlands.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*, 143–156.

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., *et al*. (2006). Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping*, *27*, 707–714.

Ethofer, T., Anders, S., Wiethoff, S., Erb, M., Herbert, C., Saur, R., *et al*. (2006). Effects of prosodic emotional intensity on activation of associative auditory cortex. *Neuroreport*, *17*, 249–253.

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current Biology*, *19*, 1028–1033.

Fecteau, S., Belin, P., Joanette, Y., & Armony, J. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *Neuroimage*, *36*, 480–487.

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.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., *et al*. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia*, *47*, 123–131.

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, *103*, 332–330.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Barthelemy, C., *et al*. (2004). Abnormal voice processing in autism: A fMRI study. *Nature Neuroscience*, *7*, 801–802.

Ghazanfar, A. A., & Hauser, M. D. (2001). The auditory behaviour of primates: A neuroethological perspective. *Current Opinion in Neurobiology*, *11*, 712–720.

Gorno-Tempini, M. L., Price, C. J., Josephs, O., Vandenberghe, R., Cappa, S. F., Kapur, N., *et al*. (1998). The neural systems sustaining face and proper-name processing. *Brain*, *121*, 2103–2118.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., *et al*. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, *8*, 145–146.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*, 187–203.

Grossman, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron*, *65*, 852–858.

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warren, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia*, *48*, (4) 1104–1114.

Heilman, K. M., Scholes, R., & Watson, R. T. (1975). Auditory affective agnosia. Disturbed comprehension of affective speech. *Journal of Neurolology, Neurosurgergy, Psychiatry*, *38*, 69–72.

Hornak, J., Rolls, E. T., & Wade, D. (1996). Face and voice expression identification inpatients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, *34*, 247–261.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., *et al*. (1997). Vocal identification of speaker and emotion activates different brain regions. *Neuroreport*, *8*, 2809–2812.

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., & Campanella, S. (2004). When audition alters vision: An event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience Letter*, *369*, 132–137.

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

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *Neuroreport*, *14*, 1129–1133.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003) 'Putting the face to the voice': Matching identity across modality. *Current Biology*, *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.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *Neuroimage*, *37*, 1445–1456.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of Acoustical Society of America*, *104*, 1598–1608.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Hearing Research*, *35*, 512–520.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*, 1138–1141.

Latinus, M., & Belin, P. (2010). *Auditory aftereffects reveal prototype-based coding of voice identity* Paper presented at the Cognitive Neuroscience Meeting, Montreal.

Latinus, M., Crabbe, F., & Belin, P. (2009). fMRI investigations of voice identity perception. *Neuroimage*, *47*(S1), S156.

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing–evidence from event-related brain potentials. *Neuroscience Letter*, *339*, 191–194.

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience*, *30*, 7604–7612.

Leitman, D. I., Wolf, D. H., Ragland, J. D., Laukka, P., Loughead, J., Valdez, J. N., *et al*. (2010). 'It's not what you say, but how you say it': A reciprocal temporo-frontal network for affective prosody. *Frontiers in Human Neuroscience*, *4*, 19.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*, 89–94.

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport*, *12*, 2653–2657.

Levy, D. A., Granot, R., & Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology*, *40*, 291–305.

Linden, D. E., Thornton, K., Kuswanto, C. N., Johnston, S. J., van de Ven, V., & Jackson, M. C. (2011). The brain's voices: Comparing nonclinical auditory hallucinations and imagery. *Cerebral Cortex*, *31*, 330–337.

McGurk, H., & MacDonald, J. D. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*, 8–15.

Mitchell, R. L., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*, *41*, 1410–1421.

Morris, J. S., Scott, S. K., & Dolan, R. J. (1999). Saying it with feeling: Neural responses to emotional vocalizations. *Neuropsychologia*, *37*, 1155–1163.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *Journal of Acoustical Society of America*, *68*, 1294–1300.

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

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, *30*, 7434–7446.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, *60*, 355–376.

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

Pell, M. D., & Leonard, C. L. (2003). Processing emotional tone from speech in Parkinson's disease: A role for the basal ganglia. *Cognitive, Affective and Behaviorol Neuroscience*, *3*, 275–288.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience*, *11*, 367–374.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., *et al*. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society B*, *265*, 1809–1817.

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

Rama, P., Martinkauppi, S., Linnankoski, I., Koivisto, J., Aronen, H. J., & Carlson, S. (2001). Working memory of identification of emotional vocal expressions: An fMRI study. *Neuroimage*, *13*, 1090–1101.

Sander, K., & Scheich, H. (2005). Left auditory cortex and amygdala, but right insula dominance for human laughing and crying. *Journal of Cognitive Neuroscience*, *17*, 1519–1531.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*, 24–30.

Schweinberger, S. R. (2001). Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia*, *39*, 921–936.

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., *et al*. (2008). Auditory adaptation in voice perception. *Current Biology*, *18*, (9) 684–688.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*, 453–463.

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

Sestieri, C., Di Matteo, R., Ferretti, A., Del Gratta, C., Caulo, M., Tartaro, A., *et al*. (2006). 'What' versus 'where' in the audiovisual domain: An fMRI study. *Neuroimage*, *33*, 672–680.

Stevens, A. A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, *18*, 162–171.

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

Van Lancker, D., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: All errors are not created equal. *Journal of Speech and Hearing Research*, *35*, 963–970.

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

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

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research Cognitive Brain Research*, *17*, 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. (2005). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, *16*, 1314–1322.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*, 557–561.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., *et al*. (2004). Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. *Cerebral Cortex*, *14*, 1384–1389.

Wildgruber, D., Pihan, H., Ackermann, H., Erb, M., & Grodd, W. (2002). Dynamic brain activation during processing of emotional intonation: Influence of acoustic parameters, emotional valence, and sex. *Neuroimage*, *15*(4), 856–869.

Winston, J. S., Henson, R. N., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, *92*, 1830–1839.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*, 959–974. doi:10.1111/j.2044-8295.2011.02045.x

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research*, *268*, 38–45.

Received 16 November 2010; revised version received 28 January 2011