How do we recognise who is speaking?

Article *in* Frontiers in bioscience (Scholar edition) · January 2014

Source: PubMed

2 authors, including:

Samuel Robert Mathias
Yale University
42 PUBLICATIONS **90** CITATIONS

SEE PROFILE

[Frontiers in Bioscience S6, 92-109, January 1, 2014]

How do we recognise who is speaking?

Samuel R. Mathias[1,2], Katharina von Kriegstein[1,3]

[1]MPRG Neural Mechanisms of Human Communication, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1, 04103 Leipzig, Germany, [2]Center for Computational Neuroscience and Neurotechnology, Boston University, 677 Beacon Street, Boston, MA 02215, [3]Department of Psychology, Humboldt University of Berlin, Rudower Chaussee 18, 12489 Berlin, Germany

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Behavioural studies
- *3.1. Perceptual differences between speakers*
- *3.2. Recognition of unfamiliar speakers*
- *3.3. Recognition of familiar speakers*

4. Neural mechanisms of speaker recognition
- *4.1. Clinical studies*
- *4.2. Voice selectivity*
- *4.3. Processing of GPR and VTL*
- *4.4. Processing of speaker identity*
- *4.5. Unfamiliar versus familiar speakers*
- *4.6. Integration of speaker and face information*
- *4.7. Models of speaker recognition*

5. Concluding remarks
6. References

1. ABSTRACT

The human brain effortlessly extracts a wealth of information from natural speech, which allows the listener to both understand the speech message and recognise who is speaking. This article reviews behavioural and neuroscientific work that has attempted to characterise how listeners achieve speaker recognition. Behavioural studies suggest that the action of a speaker's glottal folds and the overall length of their vocal tract carry important voice-quality information. Although these cues are useful for discriminating and recognising speakers under certain circumstances, listeners may use virtually any systematic feature for recognition. Neuroscientific studies have revealed that speaker recognition relies upon a predominantly right-lateralised network of brain regions. Specifically, the posterior parts of superior temporal sulcus appear to perform some of the acoustical analyses necessary for the perception of speaker and message, whilst anterior portions may play a more abstract role in perceiving speaker identity. This voice-processing network is supported by direct, early connections to non-auditory regions, such as the visual face-sensitive area in the fusiform gyrus, which may serve to optimize person recognition.

2. INTRODUCTION

In addition to its message, natural speech conveys a wealth of information about who is speaking. The human brain effortlessly extracts this information alongside the speech message, and most normal-hearing listeners find it easy to identify a personally familiar or famous speaker, as well as being able to understand what was said. It is currently unclear how listeners accomplish such robust speaker recognition, and humans generally outperform machine algorithms on this task, particularly when listening to degraded speech (1).

In the present article, we review studies relevant to the question of speaker recognition. The scope of the article is intentionally broad in order to encompass relevant findings from multiple disciplines. It is divided into two main sections. In the first, we discuss behavioural studies that shed light on the cognitive and perceptual processes underlying speaker recognition. Some of these studies have attempted to document the sources of inter-speaker variation that lead to perceptual differences in 'voice quality'. Others have investigated which acoustic cues are important for recognition *per se*, and how the speaker

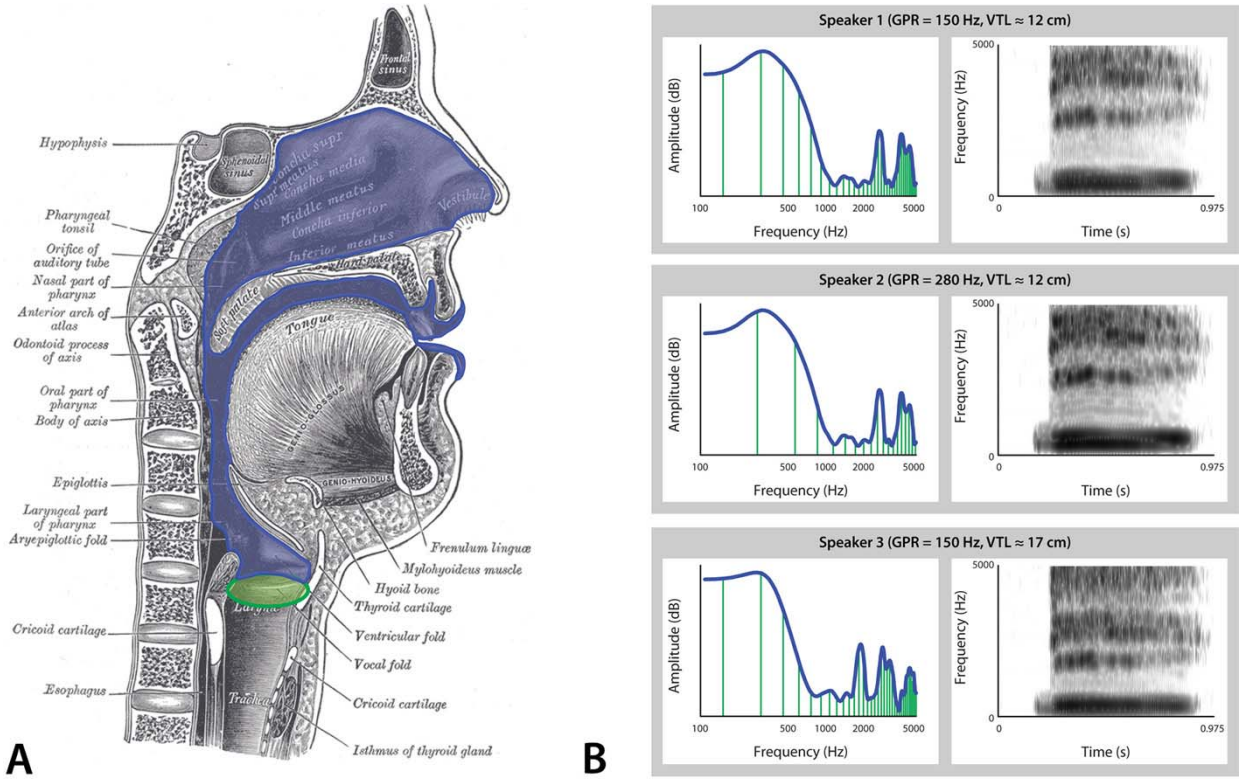


Figure 1. The contribution of glottal and vocal-tract characteristics to the perceived voice quality of a speaker. A) A sagittal slice through the head and neck with the vocal folds highlighted in green and the vocal tract highlighted in blue. B) Each row shows a cartoon spectral profile (left) and spectrogram (right) of a different synthesised speech sample. Each of the sounds was produced from the same original sound of a male speaker producing an elongated /i/ vowel, but was synthesised via STRAIGHT to simulate a speaker with a very different GPR and/or VTL.

identity is encoded. The second section discusses clinical and neuroimaging studies that have investigated the neural mechanisms underlying speaker recognition. These studies have investigated the brain networks specialised for processing voices and for recognising speakers, and have begun to explore how voice-selective brain regions interact with regions specialised for other tasks (e.g., face perception) to improve person recognition.

3. BEHAVIOURAL STUDIES

3.1. Perceptual differences between speakers

A logical starting point for investigating speaker recognition is to consider how speakers differ from one another perceptually. Such differences could be used as cues for discriminating between different speakers, judging their personal characteristics (e.g., their sex), or identifying specific individuals. The study of the sources of these inter-speaker variations comes under the moniker of 'voice quality' research (for reviews, see 2, 3, 4).

Considering the anatomy of the vocal system reveals some of the major antecedents of voice quality (Figure 1A). According to the source–filter theory of speech production (5), speech can be modelled as the action of a filter on a sound source. In voiced speech, the 'source' is a periodic waveform called the glottal pulse, created by the opening and closing of the vocal folds in the larynx (Figure 1A, green area). The 'filter' comprises the supralaryngeal vocal tract and lips, whose shape and size determine the waveform's spectral envelope and the formant frequencies (Figure 1A, blue area). Speakers manipulate their instantaneous source and filter characteristics to create the different phonetic elements of speech, but there are also relatively long-term (or 'quasi-permanent', 6) characteristics that are dependent on anatomy and are specific to the speaker.

The rate of vibration of the vocal folds, called the glottal-pulse rate (GPR), determines the speaker's fundamental frequency (f_0) and their perceived vocal pitch (Figure 1B). Several behavioural studies have shown that GPR influences the perceived similarity of unfamiliar speakers (7, 8, 9, 10, 11, 12, 13). In these studies, listeners were instructed to rate the similarity of pairs of speech stimuli produced by different speakers, and the ratings were analysed using multidimensional-scaling techniques. In each of them, listeners' similarity ratings were heavily influenced by the measured f_0 or the perceived pitch of the stimuli. This was true for different kinds of speech stimuli (isolated vowels: 7, 10, 11, 13; whole words: 8; sentences: 9, 10, 12), and when listeners heard only male speakers, only female speakers, or a mixture of both (13). Thus, GPR

appears to be an important aspect of voice quality, irrespective of the particular stimuli or speakers.

Another important aspect of voice quality is overall vocal-tract length (VTL). VTL is a filter characteristic: the spectral envelope of speech produced by a speaker with a long VTL is shifted downward in log-frequency space relative to a speaker with a short VTL (Figure 1B). Modern software has allowed researchers to manipulate VTL experimentally: vocoders such as STRAIGHT can decompose any speech signal using source–filter theory and then synthesise new stimuli with different source/filter configurations (14, 15). Psychophysical studies using STRAIGHT have shown that listeners are highly sensitive to VTL modifications, even when the values tested are outside the biologically plausible range (16, 17). Unfortunately, it is not possible to determine precisely how much influence VTL has on the perceived similarity of unfamiliar speakers from the aforementioned multidimensional-scaling studies because none of them measured their speakers' VTLs or used vocoded stimuli. However, a number of them did find that listeners' similarity ratings were influenced by the speakers' mean formant frequencies, which are determined partly by VTL, although this influence was generally weaker than that of GPR (7, 10, 13). In another study, Gaudrain and colleagues (18) presented listeners with pairs of three-vowel sequences, and asked them to judge whether '[it was] possible that both sequences were uttered by the same speaker?' The same male speaker spoke all the vowels, but his GPR/VTL was manipulated using STRAIGHT so that it differed between sequences. The authors found that listeners reported hearing different speakers with a VTL difference of at least 25%, or with a GPR difference of at least 45%. Therefore, according to this study, VTL may be perceived as a more consistent aspect of voice quality than GPR, at least when listening to isolated vowels (see also 19)

Not only are speaker-related differences in GPR and VTL perceptually salient, but listeners use these differences to infer some of the personal characteristics of a speaker. For example, GPR and VTL are important cues for discriminating speaker sex and size. Males have longer vocal tracts than females, and also tend to have slower GPRs due to their longer vocal folds (20); these differences are consistent enough for reliable sex classification based on GPR and VTL by machine-learning algorithms (21, 22). Some studies have suggested that listeners primarily use GPR when discriminating speaker sex (23, 24), but others have shown that accurate sex discrimination can be achieved even when the f_0 is modified or absent from speech (25, 26, 27, 28, 29, 30, 31, 32). More recent studies suggest that when listening to isolated vowels, both GPR and VTL contribute to the perception of speaker sex (33, 34, 35), and that there are not many 'residual' sex cues when GPR and VTL are controlled (36, 37; cf. 38). However, when listening to more complex stimuli such as whole sentences, GPR and VTL are less important because residual sex cues are available, including greater prosodic variation and faster articulation in females (22, 39, 40). Several studies have found that listeners use GPR and VTL to judge speaker size (34, 37, 41, 42, 43, 44). The role of GPR in such judgements is somewhat surprising because there is essentially no relationship between a speaker's GPR and their height, weight, or surface area, after controlling for effects of age and sex (45, 46, 47). VTL, on the other hand, is strongly correlated with speaker size (48).

Glottal and vocal-tract characteristics also have other influences on voice quality. One might describe a speaker as sounding 'breathy', for instance. Over the years, there have been many attempts to define and measure these kinds of descriptive terms. For example, in the scheme of 'vocal-profile analysis' devised by Laver (49, 50), voice quality is defined on a number of six-point descriptive scales or 'settings'. Each of these settings corresponds to an anatomical antecedent consistent with source–filter theory; for example, a breathy setting is one where the vocal folds do not close completely during vibration. Other glottal settings, according to this scheme, include laryngeal tension, whisperiness, creakiness, harshness, and falsetto. Settings related to the vocal tract include lip configuration, laryngeal position, jaw position, and tongue position. Voice-quality schemes such as Laver's (for another scheme, see 51) rely on the subjective ratings of listeners — usually expert phoneticians — rather than objective anatomical or acoustical measurements. It turns out that these schemes tend to have low inter-rater reliability, even among phoneticians (52, 53). Since listeners do not agree on how to use subjective voice-quality terms, their real-world validity is questionable. An alternative approach to defining voice quality has been to make acoustic measurements directly from the speech waveform. Voice quality is therefore defined on objective acoustic dimensions, such as 'jitter' (variation in frequency around the f_0), 'shimmer' (variation in amplitude), and harmonics-to-noise ratio (e.g., 54). This approach is also problematic, because listeners may not necessarily perceive variation along these acoustic dimensions (55). For example, several studies have instructed listeners to rate the perceived breathiness of unfamiliar speakers, and each found that breathiness was most strongly correlated with a different combination of acoustic measures (56, 57, 58, 59). In short, the lack of consistency across theoretical, acoustical, and perceptual studies of voice quality makes it very difficult to draw firm conclusions about precisely which glottal and vocal-tract characteristics, besides GPR and VTL, have a significant impact on the perceptual differences between speakers (reviewed by 2, 3).

It is important to point out that simply considering the anatomical differences between speakers' vocal systems cannot capture all the aspects of voice quality that are potentially useful for speaker recognition. For example, formant frequencies are influenced by factors such as accent, dialect, and individual speaking style, in addition to a speaker's anatomy. In their classic study, Peterson and Barney (60) measured the frequencies of the first two vowel formants from recordings of many speakers reading monosyllabic words (*heed*, *hid*, *head*, etc.) They found inter-speaker variations that were so large that the same vowel produced by two different speakers could be nothing alike: for instance, one speaker's *hid* could be more

similar to another's *head*. However, the authors only used speakers from the eastern seaboard of the United States; more recent measurements, which have incorporated speakers from all over the United States, have revealed even greater variability, presumably because of differences in accent and dialect (61, 62, 63). Such non-anatomical factors have been largely ignored by previous research on voice quality. However, as discussed later on, they are likely to play a critical role in the perception of speaker identity, particularly when listeners hear more complex speech stimuli, such as whole sentences.

3.2. Recognition of unfamiliar speakers

The behavioural studies discussed in the previous section all involved *discrimination*, requiring listeners to make judgments about voices, or compare pairs of stimuli in terms of some perceptual dimension. By contrast, *recognition* is the ability to identify a specific speaker from previously unheard speech. It does not necessarily follow that the acoustic cues relevant for discrimination discussed in the previous section are the same ones that are relevant for recognition, or that discrimination and recognition rely on the same kinds of perceptual and cognitive processes.

Many studies have investigated the circumstances that influence the reliability of 'ear-witness testimony', or how well listeners can pick out a target speaker they previously heard from an auditory 'line-up' (reviewed by 64, 65, 66). These studies can be thought of as investigating the processes underlying the recognition of *unfamiliar* speakers, since listeners had limited exposure to the target speakers prior to testing. These studies have found that recognition rates improve monotonically with the duration of the initial speech samples, being very poor at short durations (e.g., 6 seconds, 67), but much better at longer durations (e.g., 30 seconds, 68, 70 seconds, 66, 8 minutes, 69, see also 70, 71). Recognition rates decline as a function of the retention interval, but may remain above chance levels up to 4 weeks after initially hearing the target speaker (e.g., 67, 68, 69, 70, 71, 72, 73, 74). Listeners are best at recognition when the target speaker and the others in the line-up all have accents similar to their own (75, 76), and when all of the speech is in their native language (77, 78, 79, 80, 81). There has been some debate about whether recognition rates are influenced by interactions between the sex of the listener and of the speaker; a large-scale meta-analysis of several ear-witness experiments suggests that female listeners are significantly better at recognising female than male speakers, but that no similar advantage exists when male listeners hear male speakers (82, see also 65). Ear-witness recognition is disrupted if the target speaker deliberately disguises his or her voice between the initial and test stimuli (83, 84, 85, 86): even relatively minor changes in speaking style, such as switching from a normal to an angry tone, are enough to disrupt recognition (84).

Kreiman and colleagues have suggested that many of the findings of these studies can be explained in terms of a prototype model of speaker recognition (3, 73, 87). Prototype models are a common class of psychological model and have been used to explain the long-term encoding, categorization, and recognition of many kinds of stimulus (e.g., 88, 89, 90, 91). According to these models, the identity of a stimulus is encoded in terms of its deviations from an internal representation of a prototype or average stimulus. These models predict that stimuli are more faithfully represented if they are similar to the prototype (i.e., more typical) than if they are dissimilar (i.e., more distinctive). Since prototypes are based on prior experiences, speakers with the same accent as the listener are likely to be more similar to the prototype, and therefore easier to recognise, than those with different accents (75, 76). Moreover, if prototypes are formed at the level of specific phonemes, speaker recognition should be better in the listener's native language (77, 78, 79, 80, 81).

A behavioural study (73) provides direct evidence for the prototype model. Papcun *et al.* found that, several weeks after first hearing a target speaker, the proportion of times another speaker was labelled erroneously as the target increased over time. Importantly, this 'false-alarm' rate was higher for a distinctive speaker than a more typical-sounding speaker, suggesting that listeners' memories for speakers decay more rapidly when they are far from the prototype. Not all the results from studies on unfamiliar-speaker recognition are consistent with this assumption of the model, however. The model is difficult to reconcile with the basic finding that listeners are better able to remember highly distinctive speakers than more typical ones initially (92). Furthermore, a study by Yarmey (74) reached just the opposite conclusion to one reached by Papcun *et al.*. In this study, listeners heard either a distinctive or more typical speaker, then (up to one week later) rated the speaker's voice on several descriptive voice-quality terms. It was found that descriptions of the typical speaker became *less* reliable than descriptions of the distinctive speaker after delays in testing, suggesting that memory for the typical speaker had decayed more rapidly than memory for the distinctive speaker. Another issue with the prototype model is that so far no alternatives have been proposed or tested; it could be that a different kind of classification/recognition model provides a better explanation for the findings in the literature.

3.3. Recognition of familiar speakers

Numerous studies have investigated how well listeners recognise speakers with whom they are already familiar. Studies of this kind have tested listeners' recognition abilities when the stimuli have different durations (e.g., 93, 94, 95, 96), and when they have been modified acoustically (e.g., 94, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110). Acoustic manipulations have ranged from gross modifications affecting many potential recognition cues at once (e.g., filtering, 93, 94, 97; time reversing, 95, 101, 107; noise vocoding, 102) to more subtle manipulations that alter specific voice-quality characteristics (e.g., 98, 99, 103, 104, 105, 106). To ensure a high degree of prior familiarity with the speakers, these studies either used speakers and listeners who were already known to one another, such as classmates or work colleagues (93, 94, 95, 97, 102, 103, 105, 106, 111), used recordings of famous speakers (96, 100, 101), or included initial training in which the listeners

learned to reliably identify all of the speakers (103, 107, 108, 109, 110, 112, 113, 114, 115, 116).

Glottal and vocal-tract characteristics appear to play a moderately important role in familiar-speaker recognition (97, 98, 99, 103, 104, 105, 106). Abberton and Fourcin (98) instructed listeners to identify five female classmates from recordings of sentences made with a laryngograph, an instrument that can be used to generate stimuli reflecting the activity of the speaker's vocal folds without the filtering effect of the vocal tract. Listeners recognised the speakers from the laryngographic stimuli at well above chance levels, suggesting that glottal information alone can be used to recognise familiar speakers (see also 97, 98). However, performance was worse than with natural speech, which contains both glottal and vocal-tract information, and worse than with whispered speech, which contains vocal-tract information but little glottal information. In the most extensive study of its kind, Lavner *et al.* (105) instructed 30 listeners to identify 20 familiar male speakers — all members of the same kibbutz — from modified recordings of /a/ vowels. The modifications included shifting the frequencies of individual formants or combinations of formants to those of another speaker, shifting the whole spectral envelope (i.e., the VTL), changing the f_0, changing the shape of the glottal waveform, and creating 'hybrid' voices with the glottal characteristics of one speaker and the vocal-tract characteristics of another. Shifting the formant frequencies (particularly the higher formants, which remain relatively more stable than the lower ones across speech produced by a single speaker) and the whole spectral envelope had the largest effects on recognition. Shifting the f_0 (i.e., GPR) also had a strong effect, but changing the shape of the glottal waveform had little impact on recognition (cf. 99). Since Lavner *et al.* only used /a/ vowels, it is not clear whether their findings hold for whole words or sentences, or even different vowels.

Given the considerable differences between their methodologies — the use of different acoustic manipulations, different numbers of speakers and listeners, different kinds of speech stimuli, and different methods of ensuring familiarity — it is difficult to assess the relative importance of recognition cues by making direct comparisons across studies. However, one consistent finding is that familiar-speaker recognition is robust to extreme acoustic manipulation, particularly when listening to relatively long-duration speech samples. For example, Remez *et al.* (111, see also 107, 32) found that listeners could recognise their colleagues above chance levels from 'sine-wave sentences', intelligible synthetic stimuli composed of three or four time-varying sinusoids that trace the frequency contours of the formants from a real sentence. Sine-wave sentences contain very few traditional voice-quality cues — GPR and other glottal characteristics are lost completely, although VTL may be partially inferable — but retain some information regarding speaking style. By contrast, a study by Van Lancker *et al.* (100) found that listeners could sometimes recognise famous speakers from time-reversed speech, which has the same long-term spectro-temporal properties as natural speech but little information about speaking style (see also 95, 96, 107). Taken together, these results suggest that natural speech contains redundancies that allow listeners to use different sets of cues to recognise speakers, depending on the stimuli.

Another consistent finding is that the relative importance of a particular cue for familiar-speaker recognition depends on the speaker. For instance, Lavner *et al.* (106) found that some speakers were difficult to recognise from speech in which their vocal-tract characteristics were modified, but that the same manipulations hardly affected the recognition of other speakers. In a follow-up study, the authors attempted to predict via multiple regression listeners' recognition scores from direct measurements of the acoustic signals (107). They found that the regression weights for different predictors varied considerably across speakers. Similarly, van Dommelen (104) instructed listeners to identify five personally familiar female speakers from individual syllables and sentences, which were either unmodified or modified in one of three ways (overall f_0 altered, f_0 contour altered, or 'speech rhythm' altered). Recognition was poorest when the overall f_0 was altered, but this effect was strongest for the speakers with the highest and lowest original f_0 values. These results, together with those discussed in previous paragraph, strongly suggest that there is no canonical, closed set of cues along which familiar speakers are defined and recognised. Instead, they suggest that familiar-speaker recognition is a highly stimulus- and speaker-contingent process.

It is currently unclear to what extent the same principles govern the recognition of unfamiliar and familiar speakers. Some authors have proposed that they are qualitatively different processes (3, 73). However, in principle, there is no reason why both familiar and unfamiliar speakers cannot be encoded with reference to a prototype. Recently, Latinus *et al.* (116) provided evidence that the prototype model may apply to relatively familiar speakers. Studies of face perception suggest that average faces — synthetic stimuli constructed by averaging the features of many real faces — play a special role in face adaptation (for a review, see 117), suggesting that faces are represented in a multidimensional feature space, and that individual facial identities are encoded along vectors which all pass through the centre of this space. Latinus *et al.* (116) suggested that average voices play a similar role in speaker adaptation (see also 118, 119), implying that speakers are encoded relative to a prototype. Crucially, in this study, the listeners were more familiar with the target speakers than in previous studies of unfamiliar-speaker recognition or voice adaptation, having been trained to recognise them over about 6 days with various kinds of speech stimuli prior to the main experiments. However, the authors only tested for adaptation effects using isolated vowels, so it is unclear whether their findings generalise to whole words and sentences.

Further evidence that the prototype model may apply to familiar speakers is that listeners are better at recognising speakers from sentences in their native

language than a foreign language, even after extensive training (113, 114, 115). Although these results suggest that the principles governing unfamiliar- and familiar-speaker recognition may be broadly similar, the issue becomes more complicated if one considers that the *relative* familiarity of the speakers may be important. Even after prolonged amounts of speaker-recognition training, a target speaker heard only during the course of a laboratory experiment lacks the rich semantic and visual information associated with people the listeners interact with in their normal lives. It would be interesting to determine whether the results of Latinus *et al.* (116) and Perrachione *et al.* generalise to the recognition of speakers who are personally known to the listeners.

4. NEURAL MECHANISMS OF SPEAKER RECOGNITION

4.1. Clinical studies

A difficulty in recognising familiar people by their voices is called 'phonagnosia' (120). As recently reviewed by Gainotti (121), phonagnosia oftentimes co-occurs with a difficulty in recognising faces ('prosopagnosia'), usually following large lesions in the right or bilateral temporal lobes (e.g., 122, 123, 124, 125, 126). This is in contrast to the strong dissociation between prosopagnosia and naming familiar people. Prosopagnosia is normally observed in patients with right-hemisphere or bilateral lesions, and cases of phonagnosia proper (wherein the impairment cannot be explained in terms of some other difficulty) are very rare in patients with only left-hemisphere lesions (reviewed by 127, 128). By contrast, left-hemisphere lesions can lead to impairments in naming (e.g., 129). The dissociation between faces and names is consistent with the popular idea that the left temporal lobe may serve as the nexus for retrieving the verbal information associated with concrete entities (e.g., 129, 130, 131).

Although phonagnosia and proposagnosia often occur together, they are also occasionally dissociated. For example, Van Lancker and Canter (120) instructed 30 patients to identify famous white males from photographs and speech recordings. They found that five patients were poor at recognizing the speakers, but could recognize the faces normally. Five other patients exhibited prosopagnosia, and three exhibited both phonagnosia and prosopagnosia. It should be noted that this study did not rule out the possibility that there were differences in the patients' abilities on other auditory tasks. In several subsequent studies (132, 133, 134), patients not only identified famous speakers (recognition), but also judged whether pairs of speech stimuli were produced by the same or two different unfamiliar speakers (discrimination). These studies reported a double-dissociation between recognition and discrimination: in general, recognition impairments were associated with right-hemispheric parietal-lobe damage (see also 135), whereas discrimination impairments were associated with damage to either temporal lobe.

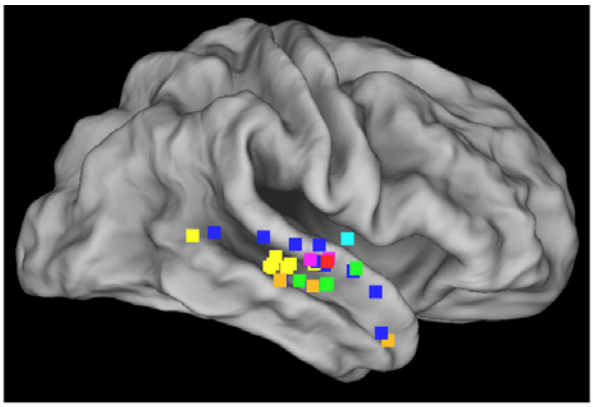
In a related study, Neuner and Schweinberger (136) instructed 36 clinical patients and controls to indicate whether faces, speech samples, or written names corresponded to famous or unknown people. Those who performed poorly on one or more of these tasks additionally performed control tasks which required them to judge whether images, sounds, or written words corresponded to living or inanimate objects. The authors also included unfamiliar face- and speaker-discrimination tasks. Although the authors observed many patterns of deficits, four patients appeared to exhibit pure phonagnosia, being worse than controls at classifying speakers (but not faces or names) as famous or unknown. Importantly, these patients were not impaired at classifying non-speech sounds, and since they performed normally on the control tasks, their impairments could not be explained in terms of anomia (impairments in naming or the recognition of names) or general auditory dysfunction. Consistent with the earlier reports (132, 133, 134), three of the four phonagnosics performed normally on the speaker-discrimination task, suggesting that their impairments were related to the recognition of speakers *per se*. Although speaker-recognition and discrimination deficits were more commonly associated with right-hemisphere damage, the patterns of lesions were complex and it is difficult to interpret them in relation to the behavioural findings in a straightforward way.

Taken together, clinical studies involving speaker recognition suggest that regions within the right hemisphere are important for recognising speakers, and that these regions are at least partially different to those involved in face and name recognition. These studies also suggest that the neural mechanisms governing speaker and face recognition either overlap or interact with one another, because cases in which phonagnosia and prosopagnosia co-occur are much more common than those in which phonagnosia occurs alone (for detailed reviews, see 121, and the article by Gainotti in the current issue).

There has been one reported case of congenital phonagnosia in the literature (137). The authors compared the performance of patient KH to controls on a range of tasks, including recognising famous speakers, learning and subsequently recognising previously unfamiliar speakers, speaker discrimination, speech-in-noise perception, environmental-sound recognition, vocal affect perception, music perception, face recognition, as well as basic auditory and neuropsychological measures. Despite having normal hearing and no known brain damage, KH was impaired on all of the speaker-related tasks. Her impairments included both recognition and discrimination, and affected the perception of famous and previously unfamiliar speakers. By contrast, KH performed normally (or slightly better than controls) on all but one of the other tasks — she was worse than controls at understanding speech in high levels of background noise.

4.2. Voice selectivity

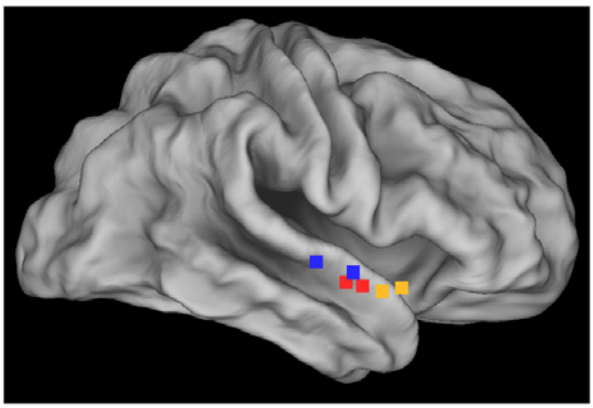
In a seminal study, Belin *et al.* investigated the processing of voices in the healthy human brain (138). In their first experiment, individuals passively listened to human vocal sounds, including speech and non-speech (e.g., coughs, cries), and other natural sounds (e.g.,



Voice selectivity

- Voices > natural sounds
- Voices > faces
- Voices > animal vocalizations
- Speech > scrambled speech
- Speech > tones and natural sounds
- Speech > noise
- Speech and laughter > natural sounds

A



Speaker identity

- Voices > scrambled voices (speech and non-speech)
- Adaptation to speaker identity
- Speaker recognition > speech recognition

B

Figure 2. A) Brain regions in the superior temporal lobe exhibiting voice selectivity. Squares represent the foci of maximal activity from several different fMRI studies that contrasted the neural activity elicited by vocal sounds with the activity elicited by other sounds. Notice that posterior, middle, and anterior portions of the STS/STG are all voice selective refs: blue: 138, 140, 142; red: 153; orange: 140; yellow: 145, 146, 147; magenta: 139, 141; green: 148, 166; cyan: 150) B) The foci of maximal activity from studies implicating the middle and anterior portions of the STS as being important for the processing of speaker identity (refs: blue: 143; red: 177, 178; orange: 148, 149)

machinery, animal vocalizations) whilst their brain activity was recorded with fMRI. The main finding was that multiple brain regions exhibited 'voice selectivity', responding more strongly to the human vocal sounds than the other sounds on average. The voice-selective regions were mostly located along the upper bank of the superior temporal sulcus (STS), with separate maxima in the posterior, middle, and anterior portions. Voice selectivity in the STS was bilateral, but appeared to be stronger in the right hemisphere. The other voice-selective maxima were in bilateral middle temporal gyrus and left planum temporale located close to STS, and the right precuneus in the parietal lobe.

Subsequent fMRI studies have since replicated the basic finding that the temporal lobe contains multiple voice-selective regions (see Figure 2A). These studies have compared the brain activity elicited by voices to the activity elicited by other natural sounds (139, 140, 141, 142), animal vocalizations (140), spectrally scrambled stimuli (143, 144, 145, 146, 147), noises (148, 149, 150, 151), tones (139, 141, 152), intelligible noise-vocoded speech (144), and faces (153). Although the precise locations of maximal activity differ across studies, it is a consistent result that multiple temporal regions usually located within the STS and/or superior temporal gyrus (STG) respond more strongly to voices than other kinds of sounds.

Further evidence of voice selectivity comes from other neuroimaging techniques. Several EEG and MEG studies have revealed that voices evoke different brain potentials to other sounds (e.g., 154, 155, 156, 157, 158, 159, 160, 161), sometimes as early as around 200 ms after stimulus onset (160, 161). A recent study using transcranial magnetic stimulation revealed that stimulating the right temporal lobes of healthy listeners impairs their abilities to classify natural sounds as either vocal or non-vocal (162). Temporal-lobe voice selectivity develops during infancy (e.g., 163, 164), and is also present in the brains of non-human animals: two recent studies have revealed an area in

the right anterior temporal lobe of the macaque brain that responds more strongly to conspecific than heterospecific vocalizations and other environmental sounds (165, 166).

A particular brain area could be voice-selective for any number of reasons, and recent work has begun to delineate the voice-selective temporal regions into their specific functional roles. Although we are far from a complete understanding of how the brain processes voices, progress in this area is rapid. A discussion of the neural mechanisms of all aspects of voice perception is beyond the scope of this article (for overviews, see 167, 168, 169); instead, we will focus now on research relevant to the question of how the healthy human brain recognises who is speaking.

4.2. Processing of GPR and VTL

As discussed earlier, a speaker's long-term glottal and vocal-tract characteristics are two major aspects of voice quality, and both can be used for speaker recognition. A speaker's GPR is perceived as vocal pitch, and numerous fMRI and MEG studies have investigated the neural correlates of pitch perception using synthetic laboratory sounds, such as complex tones and periodic noises (for a review, see 170). These studies have mostly found that bilateral antero-lateral Heschl's gyrus (HG) responds more strongly to pitched sounds than non-pitched sounds (e.g., 171, 172, 173). Two fMRI studies have explored the neural correlates of GPR perception by presenting listeners with sequences of syllables that were all synthesised from one original speaker (174, 175). In one of these studies (176), listeners heard sequences of voiced (i.e., pitched) syllables and sequences of 'whispered' syllables in which the speaker's glottal-source waveform was replaced by noise. It was found that voiced syllables activated antero-lateral HG more than whispered syllables, suggesting that vocal pitches are processed by the same neural mechanisms as the pitches of other sounds. Greater activity was observed in a more anterior region adjacent to antero-lateral HG when the authors compared the activity elicited by sequences of voiced syllables in which the speaker's GPR was randomly shifted from one syllable to the next than when the speaker's GPR always stayed the same. The latter result is consistent with studies showing that melodies elicit greater activity in anterior regions than sequences of same-pitch sounds (e.g., 171).

Several fMRI studies have also explored the neural correlates of VTL perception using synthesised syllables (174, 175, 176). Two studies found that bilateral portions of the posterior STS/STG (along with other areas) responded more strongly to VTL-varying syllable sequences than VTL-fixed sequences (175, 176). In one of these two studies (175), listeners also heard sequences of synthesised musical instruments and bullfrog vocalizations, whose spectral envelopes were either fixed or varied randomly within a sequence. It was found that posterior STS/STG responded more strongly to VTL variation in the vowels than to variation in the spectral envelopes of the non-speech sounds, whereas portions of the anterior temporal lobe and intra-parietal sulcus responded to variation in spectral envelope irrespective of sound category. These results suggest that posterior STS/STG is specialised for processing spectral-envelope fluctuations of the kind present in speech.

4.3. Processing of speaker identity

There is converging evidence suggesting that the anterior region of the STS plays a more abstract role in the processing of speaker identity per se. In a follow-up to their original study, Belin *et al.* (143) presented listeners with sequences of speech sounds, non-speech vocal sounds, and spectrally scrambled versions of the same stimuli (the same frequency components, but with their amplitudes and phases randomised). They found that bilateral posterior, middle, and anterior STS, along with bilateral primary auditory cortex, all responded more strongly to unscrambled speech than scrambled speech (see also 148, 146, 147). Importantly, only two foci in the right middle/anterior STS responded more strongly to unscrambled non-speech vocal sounds than to their scrambled versions (see Figure 2B). This result suggests that these regions responded to the presence of a voice rather the presence of basic acoustic features or linguistic content. However, these regions also responded more strongly to unscrambled speech than unscrambled non-speech. Speech typically conveys more speaker information than non-speech; therefore, a speculative explanation for the latter result is that the middle/anterior STS regions responded most strongly to the category of stimulus that contained the richest amount of information regarding the speaker.

Further support for the idea that the right anterior STS processes speaker identity rather than the speech message comes from a study using fMR adaptation (177). Listeners heard sequences of syllables in which either the syllable or the speaker randomly varied from one syllable to the next. Activity in the right anterior STS was reduced when the speaker repeated and the syllable varied than when the syllable repeated and the speaker varied; in other words, this area adapted to the presentation of different speech sounds from the same speaker. Another study (178) found that repetition of the speaker led to significant reductions in activity regardless of linguistic content in the middle/anterior superior temporal region, but this effect was bilateral rather than restricted to the right hemisphere. The authors also reported speaker-identity adaptation in the left inferior frontal gyrus.

If right anterior STS is important for encoding speaker identity, it should play an important role when listeners are required to recognise the speaker rather than the speech message. In the first of a series of related fMRI studies (148), listeners heard random sequences of sentences spoken by six different unfamiliar speakers, and performed two different tasks. In one task (the speech task), they were instructed to respond each time a sentence was the same as the very first sentence in the sequence, regardless of the speaker. In the other (the speaker task), they were instructed to respond each time a sentence was spoken by the same speaker as the first sentence, regardless of the speech message. The authors found greater activity in the right anterior STS (and also right precuneus) during

the speaker task than during the speech task. The activity in these areas must have related to task demands because listeners heard exactly the same stimuli during both tasks. The anterior STS showed another interesting response pattern during the experiment: it was no more active during the speech task than during a control condition in which the listeners heard speech-envelope noises (replicated in a subsequent study, 149). This result provides a further indication that the activity in the anterior STS related to the processing of voice identity rather than being an obligatory response to stimulus characteristics.

Further support for the idea that right anterior STS processes speaker identity comes from an fMRI study using multivariate pattern analysis. Formisano *et al.* (179) repeatedly presented listeners with three vowels spoken by three speakers (one female) and attempted to classify individual trials based on the fMRI responses. Successful speaker classification relied on a relatively small set of brain regions which were mostly located in the right hemisphere. Consistent with the studies discussed above, the greatest concentration of informative voxels was in right anterior STS.

Other fMRI studies have also implicated a role of the anterior temporal lobe in voice-identity processing, but have used relatively complex designs which make the precise function of this region difficult to decipher. Warren *et al.* (180) presented listeners with syllable sequences from the same or different speakers, which were noise-vocoded using 1, 6, or 32 filter channels. The authors found that activity in right anterior STS (along with bilateral posterior STS, right posterior STG, and left middle temporal gyrus) correlated positively with the number of filter channels but did not differ depending on whether the syllables were from the same or different speakers (cf. 177, 178). However, whether the anterior STS responded to the speaker-related aspects of the stimuli in this study is unclear because the acoustic manipulations, whilst affecting how 'voice-like' the stimuli sounded, also affected other things such as their intelligibility. Two recent studies combined stimuli generated from voice-morph continua with prior behavioural training on a speaker-recognition task to investigate the perception of speaker identity (181, 182). The first (183) reported that bilateral anterior temporal regions displayed 'identity sensitivity'. However, the study used an unusual paradigm in which listeners learned (and then re-learned) to associate portions of a voice-morph continuum with an arbitrary face; therefore, none of the speech stimuli used in this study had identities in the traditional sense. The second study used a related design, but did not find sensitivity to speaker identity in the temporal lobes (184).

The findings discussed above are in several respects reminiscent of those regarding the role of the fusiform face area (FFA) during face perception. First, several studies have shown that the FFA adapts to facial identities under certain circumstances (reviewed in 183). Second, FFA activity is task-modulated: it is stronger in response to mixed images containing both faces and houses when the subjects attend to the faces than when they attend to the houses (184). Third, it is currently thought that the FFA does not respond to any specific facial configurations or features, but rather processes facial identity in a holistic manner (see 183); this is somewhat similar to the observations made earlier that posterior rather than anterior superior temporal regions are sensitive to specific voice-quality features such as VTL. A final similarity lies in the processing of typical and distinctive faces and voices. Distinctive faces (those that deviate considerably from the average face) elicit stronger FFA activity than more typical ones (185). Similarly, an fMRI study that compared the neural responses to typical male and female speakers to those of atypical speakers (the same speakers, but whose GPRs were shifted to values more typical of the opposite sex) found that atypicality elicited stronger activity in an area very close to anterior STS (186, see also 181). Taken together, these findings suggest that the right anterior STS and the FFA might perform similar functions in the recognition of speakers and faces, respectively.

4.4. Unfamiliar versus familiar speakers

To date, only a few studies have investigated the neural correlates of speaker familiarity. An early study (187) found increased activity in several brain regions (left frontal pole, right temporal pole, right entorhinal cortex, and left precuneus) in response to sequences of sentences spoken by a mixture of familiar and unfamiliar speakers than to sequences of sentences that were all spoken by unfamiliar speakers. However, the findings were confounded because the familiar speakers were never heard alone. Nevertheless, more recent studies have implicated similar brain regions when contrasting the voices of familiar speakers against those of unfamiliar speakers (e.g. right temporal pole, right amygdala and para-/hippocampus, right precuneus/retrosplenial cortex; 188). Moreover, another study, which contrasted the joint responses to familiar faces and familiar speakers to unfamiliar faces and unfamiliar speakers, found significant activity in retrosplenial cortex (153). In addition to these brain areas, a number of studies have reported greater activity in the fusiform gyrus in response to familiar speakers (188, 189, 190, 191; see section 4.5). Activity in the fusiform gyrus appears to be specific to speaker recognition, because it occurs when listeners perform a speaker-recognition task but not when they perform a speech-recognition task. A study with a similar design (192) from a different laboratory failed to find increased activity in response to familiar speakers in the fusiform gyrus, possibly because it included relatively few subjects (11). Finally, a recent study presented listeners with famous and unknown speakers in an event-related fMRI design (193). They focused their analysis on several regions of interest along the STG/STS, and found clusters that responded more strongly to famous than to unfamiliar speakers when the listeners indicated whether those voices were familiar or unfamiliar.

To our knowledge, only one study has reported results concerning the neural mechanisms associated with listening to *unfamiliar* speakers. von Kriegstein and Giraud (148) instructed their listeners to perform a speaker task or

a speech task (see 149) whilst listening to sequences of sentences spoken by unfamiliar speakers or familiar work colleagues. The authors found that the right posterior STS (along with regions within the frontal lobes, angular gyrus, and amygdala) exhibited stronger responses to the unfamiliar than the familiar speakers when the listeners performed a speaker-recognition task, but not when they performed a speech task. It was argued that the activity in this region might have reflected the acoustic analyses necessary to perform the tasks, since the speaker task was more difficult with the unfamiliar speakers than with the familiar speakers.

4.5. Integration of speaker and face information

In everyday situations, voices and faces are usually perceived together. It is well known that the brain routinely combines information from both modalities for the purposes of improving speech intelligibility, and the neural correlates of this kind of integration have been studied extensively (reviewed by 194). What is relatively less clear is how auditory and visual information are combined for person recognition. Several behavioural studies have shown that listeners can successfully match a sentence spoken by an unfamiliar speaker to a silent video of the same speaker, and vice versa, even when the spoken and articulated sentences are different (195, 196, 197, 198, 199, 200). The perception of facial *movements* appears to be a critical factor in these experiments, since listeners are usually unable to match speech from unfamiliar speakers to static images of their faces (e.g., 196). Schweinberger and colleagues reached similar conclusions in a series or related studies (201, 202, 203; see also the article by Schweinberger and Robertson in the current issue). In these experiments, listeners were instructed to classify sentences as being spoken by familiar or unfamiliar speakers whilst viewing static images or time-synchronised videos of faces. Viewing static images and videos of faces improved classification when the identities of the voices and faces matched, whilst videos (but not static images) of faces disrupted classification when they did not match. These results demonstrate that voices and faces (particularly moving faces) convey complimentary and consistent speaker-related information that listeners can combine for recognition, but it is not clear at what stage of neural processing this combination occurs.

Traditionally, models of face perception have assumed that face recognition begins with an analysis of the visual sensory input, followed by a stage of processing in which complementary information from different modalities is combined together (see section 4.6). This characterization implies that supra-modal integration occurs only at a relatively late stage of the person-recognition process and simply enriches the representation of the person with associated biographical and other sensory knowledge. However, there is evidence to suggest that auditory and visual information are combined for person recognition more directly and at a much earlier stage of processing. Over several studies, von Kriegstein and colleagues (188, 189, 190, 191) compared the neural mechanisms of recognising visually familiar speakers to those of recognising visually unfamiliar speakers. The 'visually familiar' speakers were either personally known to the listeners (work colleagues, 188, 190), or listeners were trained to recognise their voices together with their faces (189, 191). The 'visually unfamiliar' speakers were either unknown (189), or listeners were trained to recognise them together with an arbitrary visual symbol (189, 190, 191). All of these studies included several different control conditions, such as those in which the listeners recognised the speech message rather than the speakers using the same stimuli. A consistent finding was that the recognition of the visually familiar speakers activated the FFA (as defined by an independent face-area localiser scan). Moreover, functional connectivity analyses showed that only when recognising visually familiar speakers, there were strong correlations in the time course of activity between the FFA and the voice-selective regions in the middle/anterior STS. The functional connectivity was not present to the same extent when listeners recognised the visually unfamiliar speakers or when they performed a speech task. The results suggest that when listeners attempt to recognise a visually familiar speaker by voice, face-recognition regions are recruited and communicate with the voice-recognition regions.

The recruitment of face regions during speaker recognition does not necessarily mean that visual information is integrated at an early stage of voice processing: the functional connectivity between voice and face regions could reflect co-modulation by a later supra-modal region rather than direct communication. Evidence against this explanation comes from two recent studies (204, 205). In the first (204), the authors used fMRI to localise the voice-selective regions (posterior, middle, and anterior STS) and the FFA individually in each listener, and then used diffusion MRI to investigate the structural connections between these regions. They found structural white-matter connections between the individually localised voice-sensitive regions and the FFA. These connections were stronger between the middle and anterior STS (the areas implicated in the processing of speaker identity, 138, 148, 149, 177, 179) and the FFA than between the posterior STS and the FFA. The second study used MEG to investigate at what time point the FFA is recruited during the recognition of visually familiar speakers (205). The authors found evidence that the recruitment occurred early during sensory processing, around 110 ms after the onset of the acoustic stimuli. These findings strongly suggest that the functional correlations observed between these areas during speaker recognition are the result of direct communication rather than indirect co-modulation via a third region.

4.6. Modelling speaker recognition

In a now classic theoretical paper, Bruce and Young (206) proposed a theoretical model of face recognition. The model assumed that after basic visual processing, faces go through a stage of 'structural encoding', which produces a set of abstracted representations of the face (e.g., view-centred, expression-independent). These representations are sent to three modules, which perform the analysis of facial speech, emotion expression, and face identity, respectively. The

face-identity module was characterised in terms of 'face-recognition units' (FRUs) that each code for a specific face that is familiar to the viewer. Finally, the FRUs connect to another module containing supra-modal 'person-identity nodes' (PINs), which retrieve biographical knowledge and information from other sensory modalities about the individual. Bruce and Young's original model has proven enormously influential since its inception, and although a number of its elements have been refined and revised, most subsequent models have not deviated from its core tenets — namely, structural encoding, FRUs, and PINs (e.g., 207, 208, 209, 210; for a recent overview, see 211).

Subsequent work has expanded Bruce and Young's (206) model to incorporate a processing pathway for voices (168, 169, 209, 212). In particular, Ellis *et al*. (212) suggested that speech goes through a set of processing stages mirroring those of faces, including initial structural encoding followed by an extraction of the speaker's identity via 'voice-recognition units' (VRUs). Based on its importance in tasks involving speaker recognition and its ability to adapt to speaker identity, the right anterior STS is a good candidate for the neural locus of the VRUs. One problem with this hypothesis, however, is that the right anterior STS has not been consistently shown to respond differently to unfamiliar and familiar speakers, as recent work suggests that it should (see the articles by Gainotti and Hanley in the current issue). If there are indeed VRUs within the right anterior STS, and if there are corresponding FRUs within the FFA, the empirical findings described in section 4.5 suggest that these units communicate with each other directly in addition to projecting to the same supra-modal brain regions (i.e., the PINs; 188, 189, 190, 191, 204, 205).

5. CONCLUDING REMARKS

How we recognise who is speaking remains an open question. Aspects of a speaker's anatomy lead to perceptible differences in the voice quality, which can be used to discriminate between speakers or make judgments about their personal characteristics. Among the most salient of these voice-quality characteristics are GPR and VTL. GPR and VTL appear to play an important role when listeners recognise familiar speakers, but this is highly dependent on the type of stimuli (e.g., vowels, words, or sentences) and on the speaker's themselves. It is not fully understood how we encode representations of speaker identity, or how this process differs for unfamiliar and familiar speakers. Clinical studies involving patients with predominantly right-hemisphere brain damage suggest that damage to these areas can impair familiar-speaker recognition whilst leaving other abilities, such as voice discrimination and face recognition, relatively intact. Functional neuroimaging in healthy listeners has revealed that the processing of voices relies on a predominantly right-lateralised network of temporal brain regions, which share direct and early connections with regions that deal with face perception. Future progress is likely to be expedited by the parallel development of psychological and neuroanatomical models of speaker recognition.

6. REFERENCES

1. A Schmidt-Nielsen, T Crystal: Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Digit Signal Process* 10, 249–266 (2000)

2. J Kreiman, D Van Lancker-Sidtis, B Gerratt. Perception of voice quality. In: The Handbook of Speech Perception. Eds: DB Pisoni, RE Remez Malden, Massachusetts (2007)

3. J Kreiman, B Gerratt: Perceptual assessment of voice quality: Past, present, and future. *Perspectives on Voice and Voice Disorders* 20, 62–67 (2010)

4. J Kreiman, D Van Lancker-Sidtis. Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception. Chichester, United Kingdom (2013)

5. G Fant. Acoustic Theory of Speech Production. The Hague, Netherlands (1960)

6. D Abercrombie. Elements of General Phonetics. Chicago, Illinois (1967)

7. H Matsumoto, S Hiki, T Sone, T Nimura: Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Trans Acoust* 21, 428–436 (1973)

8. B Walden, A Montgomery, G Gibeily, R Prosek, D Schwartz: Correlates of psychological dimensions in talker similarity. *J Speech Hear Res* 21, 265–275 (1978)

9. S Singh, T Murry: Multidimensional classification of normal voice qualities. *J Acoust Soc Am* 64, 81–87 (1978)

10. T Murry, S Singh: Multidimensional analysis of male and female voices. *J Acoust Soc Am* 68, 1294–1300 (1980)

11. J Kreiman, B Gerratt, K Precoda, G Berke: Individual differences in voice quality perception. *J Speech Hear Res* 35, 512–520 (1992)

12. M Gelfer: A multidimensional scaling study of voice quality in females. *Phonetica* 50, 15–27 (1993)

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

14. H Kawahara, I Masuda-Katsuse, A Cheveigné: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun* 27, 187–207 (1999)

15. H Kawahara, T Irino. Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In: Speech Separation by

Humans and Machines. Ed; P Divenyi Boston, Massachusetts (2004)

16. D Smith, R Patterson, R Turner, H Kawahara, T Irino: The processing and perception of size information in speech sounds. *J Acoust Soc Am* 117, 305–318 (2005)

17. D Ives, D Smith, R Patterson: Discrimination of speaker size from syllable phrases *J Acoust Soc Am* 118, 3816–3822 (2005)

18. E Gaudrain, S Li, V Ban, R Patterson: The role of glottal pulse rate and vocal tract length in the perception of speaker identity. Interspeech 2009, 152–155 (2009) (available at http://www.researchgate.net/publication/221478138_The_role_of_glottal_pulse_rate_and_vocal_tract_length_in_the_perception_of_speaker_identity)

19. H Kuwabara, T Takagi: Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Commun* 10, 491–495 (1991)

20. I Titze: Physiologic and acoustic differences between male and female voices. *J Acoust Soc Am* 85, 1699–1707 (1989)

21. J Bachorowski, M Owren: Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J Acoust Soc Am* 106, 1054–1063 (1999)

22. J Hillenbrand, M Clark: The role of f(0) and formant frequencies in distinguishing the voices of men and women. *Atten Percept Psychophys* 71, 1150–1166 (2009)

23. R Coleman: A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *J Speech Hear Res* 19, 168–180 (1976)

24. S Whiteside: Identification of a speaker's sex: A study of vowels. *Percept Motor Skills* 86, 579–584 (1998)

25. F Ingemann: Identification of the speaker's sex from voiceless fricatives. *J Acoust Soc Am* 44, 1142–1144 (1968)

26. M Schwartz: Identification of speaker sex from isolated, voiceless fricatives. *J Acoust Soc Am* 43, 1178–1179 (1968)

27. M Schwartz, H Rine: Identification of speaker sex from isolated, whispered vowels. *J Acoust Soc Am* 44, 1736–1737 (1968)

28. W Brown, S Feinstein: Speaker sex identification utilising a constant laryngeal source. *Folia Phoniat* 29, 240–248 (1976)

29. M Lass, K Hughes, M Bowyer, L Waters, V Bourne: Speaker sex identification from voiced, whispered, and filtered isolated vowels. *J Acoust Soc Am* 59, 675–678 (1976)

30. R Coleman: Male and female voice quality and its relationship to vowel formant frequencies. *J Speech Hear Res* 14, 565–577 (1971)

31. B Weinberg, S Bennett: A study of talker sex recognition of esophageal voices *J Speech Hear Res* 14, 391–395 (1971)

32. J Fellowes, R Remez, & P Rubin: Perceiving the sex and identity of a talker without natural vocal timbre. *Percept Psychophys* 59, 839–849 (1997)

33. D Childers, K Wu: Gender recognition from speech. Part II: Fine analysis. *J Acoust Soc Am* 90, 1841–1856 (1991)

34. D Smith, R Patterson: The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *J Acoust Soc Am* 118, 3177–3186 (2005)

35. M Gelfer, V Mikos: The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *J Voice* 19, 544–554 (2005)

36. D Honorof, D Whalen: Identification of speaker sex from one vowel across a range of fundamental frequencies. *J Acoust Soc Am* 128, 3095–3104 (2010)

37. D Smith, T Walters, R Patterson: Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *J Acoust Soc Am* 122, 3628–2639 (2007)

38. D Klatt, L Klatt: Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am* 87, 820–857 (1990)

39. P Assmann, T Neary, S Dembling. Effects of frequency shifts on perceived naturalness and gender information in speech. In: Proceedings of the 9th International Conference on Spoken Language Processing Pittsburgh, Pennsylvania (2006) (available at http://www.utdallas.edu/~assmann/icslp06.pdf)

40. A Simpson: Phonetic differences between male and female speech. Lang Linguist Compass 3, 621–640 (2009)

41. W Van Dommelen, B Moxness: Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Lang Speech* 38, 267–287 (1994)

42. S Collins: Men's voices and women's choices. *Anim Behav* 60, 773–780 (2000)

43. D Rendall, J Vokey, C Nemeth: Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size. *J Exp Psychol Hum Percept Perform* 33, 1208–1219 (2007)

44. K Pisanski, D Rendall: The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *J Acoust Soc Am* 129 2201–2212 (2011)

45. N Lass, W Brown: Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. *J Acoust Soc Am* 63, 1218–120 (1978)

46. H Künzel: How well does average fundamental frequency correlate with speaker height and weight? *Phonetica* 46, 117–125 (1989)

47. H Hollien, R Green, K Massey: Longitudinal research on adolescent voice change in males. *J Acoust Soc Am* 96, 2646–2653 (1994)

48. W Fitch, J Giedd: Morphology and development of the human vocal tract: A study using magnetic resonance imaging *J Acoust Soc Am* 106, 1511–1522 (1999)

49. J Laver. The Phonetic Description of Voice Quality. Cambridge, United Kingdom (1980)

50. J Laver. Phonetic evaluation of voice quality. In: Voice Quality Measurement. Eds: RD Kent, MJ Ball San Diego, California (2000)

51. M Hirano. Clinical Examination of Voice. New York, New York (1981)

52. J Kreiman, B Gerratt: Validity of rating scale measures of voice quality. *J Acoust Soc Am* 104, 1598-1608 (1998)

53. A Webb, P Carding, I Deary, K MacKenzie, N Steen, J Wilson: The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol* 261, 429–434 (2004)

54. I Titze, Y Horii, R Scherer: Some technical considerations in voice perturbation measurements. *J Speech Hear Res* 30, 252–260 (1987)

55. Y Maryn, N Roy, M De Bodt, P Van Cauwenberge, P Corthals: Acoustic measurement of overall voice quality: A meta-analysis. *J Acoust Soc Am* 126, 2619–2634 (2009)

56. B Fritzell, B Hammarberg, J Gauffin, I Karlsson, J Sundberg: Breathiness and insufficient vocal fold closure. *J Phon* 14, 549–553 (1986)

57. J Hillenbrand, R Cleveland, R Erickson: Acoustic correlates of breathy vocal quality. *J Speech Hear Res* 37, 769–778 (1994)

58. G de Krom: Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *J Speech Hear Res* 38, 794–811 (1995)

59. D Martin, J Fitch, V Wolfe: Pathologic voice type and the acoustic prediction of severity. *J Speech Hear Res* 38, 765–771 (1995)

60. G Peterson, H Barney: Control methods used in a study of vowels. *J Acoust Soc Am* 24, 175–184 (1952)

61. J Hillenbrand, L Getty, M Clark, K Wheeler: Acoustic characteristics of American English vowels. *J Acoust Soc Am* 97, 3099–3111 (1995)

62. R Hagiwara: Dialect variation and formant frequency: The American English vowels revisited. *J Acoust Soc Am* 102, 655–658 (1997)

63. C Clopper, D Pisoni, K de Jong: Acoustic characteristics of the vowel systems of six regional varieties of American English *J Acoust Soc Am* 118, 1661–1676 (2005)

64. B Clifford: Voice identification by human listeners: On earwitness reliability. *Law Hum Behav* 4, 373–394 (1980)

65. A Yarmey: Earwitness speaker identification. *Psychol Public Policy Law* 1, 792–816 (1995)

66. A Yarmey, R Lindsay, D Ross, J Read, M Toglia. The psychology of speaker identification and earwitness memory. In: The Handbook of Eyewitness Psychology: Volume II: Memory for People. Eds: R Lindsay, D Ross, J Read, M Toglia Mahwah, New Jersey (2007)

67. G Legge, C Grosmann, C M Pieper: Learning unfamiliar voices. *J Exp Psychol Learn Mem Cogn*10, 298–303 (1984)

68. J Kerstholt, N Jansen, A Van Amelsvoort, A Broeders: Earwitnesses: effects of speech duration, retention interval and acoustic environment. *Appl Cogn Psychol* 18, 327–336 (2004)

69. A Yarmey, E Matthys: Voice identification of an abductor. *Appl Cogn Psychol* 6, 367–377 (1992)

70. R Roebuck, J Wilding: Effects of vowel variety and sample length on identification of a speaker in a line-up. *Appl Cogn Psychol* 7, 475–481 (1993)

71. S Cook, J Wilding: Earwitness testimony: Never mind the variety, hear the length. *Appl Cogn Psychol* 11, 95–111 (1997)

72. B Clifford, H Rathborn, R Bull: The effects of delay on voice recognition accuracy. *Law Hum Behav* 5, 201–208 (1981)

73. G Papcun, J Kreiman, A Davis: Long-term memory for unfamiliar voices. *J Acoust Soc Am* 85, 913–925 (1989)

74. A Yarmey: Descriptions of distinctive and non-distinctive voices over time. *J Forensic Sci Soc* 31, 421–428 (1991)

75. A Goldstein, P Knight, K Bailis, J Conover: Recognition memory for accented and unaccented voices. *Bull Psychon Soc* 17, 217–220 (1981)

76. S Stevenage, G Clarke, A McNeill: The “other-accent” effect in voice recognition. *J Cogn Psychol* 24, 647–653 (2012)

77. C Thompson: A language effect in voice identification. *Appl Cogn Psychol* 1, 121–131 (1987)

78. J Goggin, C Thompson, G Strube, L Simental: The role of language familiarity in voice identification. *Mem Cogn* 19, 448–58 (1991)

79. A Philippon, J Cherryman, R Bull, A Vrij: Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Appl Cogn Psychol* 21, 539–550 (2007)

80. S Winters, S Levi, D Pisoni: Identification and discrimination of bilingual talkers across languages. *J Acoust Soc Am* 123, 4524–4538 (2008)

81. M Wester: Talker discrimination across languages. *Speech Commun* 54, 781–790 (2012)

82. J Wilding, S Cook: Sex differences and individual consistency in voice identification. *Percept Motor Skills* 91, 535–538 (2000)

83. A Reich, J Duke: Effects of selected vocal disguises upon speaker identification by listening. *J Acoust Soc Am* 66, 1023–1028 (1979)

84. H Saslove, A DYarmey: Long-term auditory memory: Speaker identification. *J Appl Psychol* 65, 111–116 (1980)

85. H Hollien, W Majewski, E Doherty: Perceptual identification of voices under normal, stress and disguise speaking conditions. *J Phon*, 10, 139–148 (1982)

86. T Orchard, A Yarmey: The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Appl Cogn Psychol* 9, 249–260 (1995)

87. J Kreiman, G Papcun: Comparing discrimination and recognition of unfamiliar voices. *Speech Commun* 10, 265–275 (1991)

88. M Posner, S Keele: On the genesis of abstract ideas. *J Exp Psychol* 77, 353–363 (1968)

89. S Reed: Pattern recognition and categorization. *Cogn Psychol* 3, 382–407 (1972)

90. E, Rosch: Natural categories. *Cogn Psychol* 4, 328–350 (1972)

91. D Homa, S Sterling, L Trepel Limitations of exemplar-based generalization and the abstraction of categorical information. *J Exp Psychol Hum Percept Perform* 7, 418–439 (1981)

92. J.Mullennix, A Ross, C Smith, K Kuykendall, J Conard, S Barb, Typicality effects on memory for voice: Implications for earwitness testimony. *Appl Cogn Psychol* 25, 29–34 (2009)

93. I Pollack, J M Pickett, W Sumby: On the identification of speakers by voice. *J Acoust Soc Am* 26, 403–406 (1954)

94. A Compton: Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally. *J Acoust Soc Am* 35, 1748–1752 (1963)

95. P Bricker, S Pruzansky: Effects of stimulus content and duration on talker identification. *J Acoust Soc Am* 40, 1441–1449 (1966)

96. S Schweinberger, A Herholz, W Sommer: Recognising famous voices: Influence of stimulus duration and different types of retrieval cues. J Speech Lang *Hear Res* 40, 453–463 (1997)

97. C LaRivere, Contributions of fundamental frequency and formant frequencies to speaker identification. Phonetica 31, 185–197 (1975)

98. E Abberton, A Fourcin: Intonation and speaker identification. *Lang Speech* 21, 305–318 (1978)

99. T Carrell. Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification. Doctoral dissertation, Indiana University, Bloomington, Indiana (1984)

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

101. D Van Lancker, J Kreiman, T Wickens: Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *J Phon* 13, 39–52 (1985)

102. A Schmidt-Nielsen, K Stern: Identification of known voices as a function of familiarity and narrow band coding. *J Acoust Soc Am* 77, 658– 663 (1985)

103. W Van Dommelen: The contribution of speech rhythm and pitch to speaker recognition. *Lang Speech* 30, 325–338 (1987)

104. W Van Dommelen: Acoustic parameters in human speaker recognition. *Speech* 33, 259–272 (1990)

105. Y Lavner, I Gath, J Rosenhouse: The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Comm* 30, 9– 26 (2000)

106. Y Lavner, J Rosenhouse, I Gath: The prototype model in speaker identification by human listeners. *Int J Speech Technol* 4, 63–74 (2001)

107. S Sheffert, D Pisoni, J Fellowes, R Remez: Learning to recognise talkers from natural, sinewave, and reversed speech samples. *J Exp Psychol Hum Percept Perform* 28, 1447–1469 (2002)

108. J Allen, J Miller: Listener sensitivity to individual talker differences in voice-onset-time. *J Acoust Soc Am* 115, 3171–3183 (2004)

109. R Theodore, J Miller, D DeSteno: Individual talker differences in voice-onset-time: contextual influences. *J Acoust Soc Am* 125, 3974–3982 (2009)

110. R Theodore, J Miller: Characteristics of listener sensitivity to talker-specific phonetic detail. *J Acoust Soc Am* 128, 2090–2099 (2010)

111. R Remez, J Fellowes, P Rubin: Talker identification based on phonetic information. *J Exp Psychol Hum Percept Perform* 23, 651–666 (1997)

112. M Vongphoe, F Zeng: Speaker recognition with temporal cues in acoustic and electric hearing. *J Acoust Soc Am* 118, 1055–1061 (2005)

113. T Perrachione, J Pierrehumbert, P Wong: Differential neural contributions to native- and foreign-language talker identification. *J Exp Psychol Hum Percept Perform* 35, 1950–1960 (2009)

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

115. T Perrachione, S Del Tufo, J Gabrieli: Human voice recognition depends on language ability. *Science* 333, 595 (2011)

116. M Latinus, P Belin: Anti-voice adaptation suggests prototype-based coding of voice identity. *Front Psychol* 2, 175 (2010)

117. M Webster, D MacLeod: Visual adaptation and face perception. *Philos Trans R Soc Lond B Biol Sci* 366, 1702–1725 (2011)

118. S Schweinberger, C Casper, N Hauthal, J Kaufmann, H Kawahara, N Kloth, D Robertson, A Simpson, R Zäske: Auditory adaptation in voice perception. *Curr Biol* 18, 684–648 (2008)

119. R Zäske, S Schweinberger, H Kawahara: Voice aftereffects of adaptation to speaker identity. *Hear Res* 268, 38–45 (2010)

120. D Van Lancker, G Canter: Impairment of voice and face recognition in patients with hemispheric damage. *Brain Cogn* 1, 185–195 (1982)

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

122. A Ellis, A Young, E Critchley: Loss of memory for people following temporal lobe damage. *Brain* 112, 1469–1483 (1989)

123. J Hanley, A Young, N Pearson: Defective recognition of familiar people. *Cogn Neuropsychol* 6 179–210 (1989)

124. G Gainotti, A Barbier, C Marra: Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. *Brain* 126, 792–803 (2003)

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

126. J Hailstone, G Ridgway, J Bartlett, J Goll, A Buckley, S Crutch, J Warren, Voice processing in dementia: a neuropsychological and neuroanatomical analysis. *Brain* 134, 2535–2547 (2011)

127. G Gainotti: Different patterns of famous people recognition disorders in patients with right and left anterior temporal lesions: A systematic review. *Neuropsychologia* 45, 1591–607 (2007)

128. G Gainotti, C Marra: Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders., *Front Hum Neurosci* 5, 55 (2011)

129. D Tranel: The left temporal pole is important for retrieving words for unique concrete entities. *Aphasiology* 23, 7–8 (2009)

130. H Damasio, D Tranel, T Grabowski, R Adolphs, A Damasio: Neural systems behind word and concept retrieval. *Cognition* 92, 179–229 (2004)

131. D Drane, G Ojemann, E Aylward, J Ojemann, L Johnson, D Silbergeld, J Miller, D Tranel: Category-specific naming and recognition deficits in temporal lobe epilepsy surgical patients. *Neuropsychologia* 46, 1242–1455 (2008)

132. D Van Lancker, J Kreiman: Voice discrimination and recognition are separate abilities. *Neuropsychologia* 25, 829–34 (1987)

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

134. D Van Lancker, J Kreiman, J Cummings: Voice perception deficits: Neuroanatomical correlates of phonagnosia. *J Clin Exp Neuropsychol* 11, 665–674 (1989)

135. C Lang, O Kneidl, M Hielscher-Fastabend, J Heckmann: Voice recognition in aphasic and non-aphasic stroke patients. *J Neurol* 256, 1303–1306 (2009)

136. F Neuner, S Schweinberger: Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain Cogn* 44, 342–366 (2000)

137. L Garrido, F Eisner, C McGettigan, L Stewart, D Sauter, J R Hanley, S R Schweinberger, J D Warren, B Duchaine: Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia* 47, 123–131 (2009)

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

139. K Specht, J Reul: Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: an auditory rapid event-related fMRI-task. *NeuroImage* 20, 1944–1954 (2003)

140. S Fecteau, J L Armony, Y Joanette, P Belin: Is voice processing species-specific in human auditory cortex? An fMRI study. *NeuroImage* 23, 840–848 (2004)

141. K Specht, B Osnes, K Hugdahl: Detection of differential speech-specific processes in the temporal lobe using fMRI and a dynamic "sound morphing" technique. *Hum Brain Mapp* 30, 3436–3444 (2009)

142. S Shultz, A Vouloumanos, K Pelphrey: The superior temporal sulcus differentiates communicative and noncommunicative auditory signals. *J Cogn Neurosci* 24, 1224–1232 (2012)

143. P Belin, R J Zatorre, P Ahad: Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 13, 17–26 (2002)

144. A Giraud, C Kell, C Thierfelder, P Sterzer, M O Russ, C Preibisch, A Kleinschmidt: Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cereb Cortex* 14, 247–255 (2004)

145. P Rämä, A Poremba, J Sala, L Yee, M Malloy, M Mishkin, S Courtney: Dissociable functional cortical topographies for working memory maintenance of voice identity and location. *Cereb Cortex* 14, 768–80 (2004)

146. P Rämä, S M Courtney, Functional topography of working memory for face or voice identity. *NeuroImage* 24, 224–234 (2005)

147. K Relander, P Rämä: Separate neural processes for retrieval of voice identity and word content in working memory. *Brain Res* 1252, 143–151 (2009)

148. K von Kriegstein, E Eger, A Kleinschmidt, A Giraud: Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Research* 17, 48–55 (2003)

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

150. M Meyer, S Zysset, D von Cramon, K Alter: Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. *Brain Res Cogn Brain Res* 24, 291–306 (2005)

151. J Obleser, H Boecker, A Drzezga, B Haslinger, A Hennenlotter, M Roettinger, C Eulitz, J P Rauschecker: Vowel sound extraction in anterior superior temporal cortex. *Hum Brain Mapp* 27, 562–571 (2006)

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

153. N J Shah, J C Marshall, O Zafiris, A Schwab, K Zilles, H J Markowitsch, G R Fink: The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. *Brain* 124, 804–815 (2001)

154. H Tiitinen, P Sivonen, P Alku, J Virtanen, R Näätänen: Electromagnetic recordings reveal latency differences in speech and tone processing in humans. *Brain Res Cogn Brain Res* 8, 355–363 (1999)

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

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

157. A Gunji, S Koyama, R Ishii, D Levy, H Okamoto, R Kakigi, C Pantev: Magnetoencephalographic study of the cortical activity elicited by human voice. *Neurosci Lett* 348, 13–16 (2003)

158. T Mizuochi, M Yumoto, S Karino, K Itoh, K Yamakawa, K Kaga: Perceptual categorization of sound spectral envelopes reflected in auditory-evoked N1m. *Neuroreport* 16, 555–558 (2005)

159. T Mizuochi, M Yumoto, S Karino, K Itoh, T Yamasoba: Latency variation of auditory N1m responses to vocal and nonvocal sounds. *Neuroreport* 18, 1945–1949 (2007)

160. I Charest, C Pernet, G Rousselet, I Quiñones, M Latinus, S Fillion-Bilodeau, J Chartrand, P Belin: Electrophysiological evidence for an early processing of human voices. *BMC Neurosci* 10, 127 (2009)

161. M De Lucia, S Clarke, M M Murray: A temporal hierarchy for conspecific vocalization discrimination in humans. *J Neurosci* 30, 11210–11221 (2010)

162. P Bestelmeyer, P Belin, M Grosbras: Right temporal TMS impairs voice detection. *Curr Biol* 21, 838–839 (2011)

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

164. A Blasi, E Mercure, S Lloyd-Fox, A Thomson, M Brammer, D Sauter, Q Deeley, G Barker, V Renvall, S Deoni, D Gasston, S Williams, M Johnson, A Simmons, D Murphy: Early specialization for voice and emotion processing in the infant brain. *Curr Biol* 21, 1220–1224 (2011)

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

166. C Perrodin, C Kayser, N Logothetis, C Petkov: Voice cells in the primate temporal lobe. *Curr Biol* 21, 1408–1415 (2011)

167. P Belin, S Fecteau, C Bédard: Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 8, 129–135 (2004)

168. S Campanella, P Belin, Integrating face and voice in person perception. *Trends Cogn Sci* 11, 535–543 (2007)

169. P Belin, P Bestelmeyer, M Latinus, R Watson: Understanding voice perception. *Br J Psychol* 102, 711–725 (2011)

170. K Walker, J Bizley, A King, J Schnupp: Cortical encoding of pitch: recent results and open questions. *Hear Res* 271, 74–87 (2010)

171. R Patterson, S Uppenkamp, I Johnsrude, T Griffiths: The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36, 767–776 (2002)

172. K Krumbholz, R Patterson, A Seither-Preisler, C Lammertmann, B Lütkenhöner: Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. *Cereb Cortex* 13, 765–772 (2003)

173. H Penagos, J Melcher, A Oxenham: A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J Neurosci* 24, 6810–6815 (2004)

174. K von Kriegstein, J Warren, D Ives, R Patterson, T Griffiths: Processing the acoustic effect of size in speech sounds. *NeuroImage* 32, 368–375 (2006)

175. K von Kriegstein, D Smith, R Patterson, S Kiebel, T Griffiths: How the human brain recognises speech in the context of changing speakers. *J Neurosci* 30, 629–638 (2010)

176. K von Kriegstein, D Smith, R Patterson, D Ives, T Griffiths: Neural representation of auditory sise in the human voice and in sounds from other resonant sources. *Curr Biol* 17, 1123–1128 (2007)

177. P Belin, R Zatorre: Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14, 2105–2109 (2003)

178. B Chandrasekaran, A Chan, P Wong: Neural processing of what and who information in speech *J Cogn Neurosci* 23, 2690–2700 (2011)

179. E Formisano, F De Martino, M Bonte, R Goebel: "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322, 970–973 (2008)

180. J Warren, S Scott, C Price, T Griffiths: Human brain mechanisms for the early analysis of voices. *NeuroImage* 31, 1389–1397 (2006)

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

182. M Latinus, F Crabbe, P Belin: Learning-induced changes in the cerebral processing of voice identity. *Cereb Cortex* 21, 2820–2828 (2011)

183. N Kanwisher, G Yovel: The fusiform face area: a cortical region specialised for the perception of faces. *Philos Trans R Soc Lond B Biol Sci* 361, 2109–2128 (2006)

184. K O'Craven, P Downing, N Kanwisher: fMRI evidence for objects as the units of attentional selection. *Nature* 401, 584–587 (1999)

185. G Loffler, G Yourganov, F Wilkinson, H Wilson: fMRI evidence for the neural representation of faces. *Nature Neurosci* 8, 1386–1390 (2005)

186. S Lattner, M Meyer, A Friederici, Voice perception: Sex, pitch, and the right hemisphere. *Human Brain Mapp* 24, 11–20 (2004)

187. K Nakamura, R Kawashima, M Sugiura, T Kato, A Nakamura, K Hatano, S Nagumo, K Kubota, H Fukuda, K Ito, S Kojima: Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia* 39, 1047–1054 (2000)

188. K von Kriegstein, A Kleinschmidt, P Sterzer, A Giraud: Interaction of face and voice areas during speaker recognition. *J Cogn Neurosci* 17, 367–376 (2005)

189. K von Kriegstein, A Giraud: Implicit multisensory associations influence voice recognition. *PLoS Biol* 4, e326 (2006)

190. K von Kriegstein, A Kleinschmidt, A Giraud: Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cereb Cortex* 16, 1314–1322 (2006)

191. K von Kriegstein, O Dogan, M Grüter, A Giraud, C Kell, T Grüter, A Kleinschmidt, S Kiebel: Simulation of talking faces in the human brain improves auditory speech recognition. *Proc Natl Acad Sci U S A* 105, 6747–6752 (2008)

192. P Birkett, M Hunter, R Parks, T Farrow, H Lowe, I Wilkinson, P Woodruff: Voice familiarity engages auditory cortex. *Neuroreport* 18, 1375–1378 (2007)

193. A Bethmann, H Scheich, A Brechmann: The temporal lobes differentiate between the voices of famous and unknown people: An event-related fMRI study on speaker recognition. *PLoS One* 7(10), e47626 (2012)

194. R Campbell: The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci* 363, 1001–1010 (2007)

195. L Rosenblum, D Yakel, N Baseer, A Panchal, B Nodarse, R Niehus: Visual speech information for face recognition. *Percept Psychophys* 64, 220–229 (2002)

196. M Kamachi, H Hill, K Lander, E Vatikiotis-Bateson: "Putting the face to the voice": matching identity across modality. *Curr Biol* 13, 1709–1714 (2003)

197. L Lachs, D Pisoni: Crossmodal Source Identification in Speech Perception. *Ecol Psychol* 16, 159–187 (2004)

198. L Lachs, D Pisoni: Specification of cross-modal source information in isolated kinematic displays of speech. *J Acoust Soc Am* 116, 507–518 (2004)

199. L Rosenblum, N Smith, S Nichols, S Hale, J Lee: Hearing a face: Cross-modal speaker matching using isolated visible speech. Atten *Percept Psychophys* 68, 84–93 (2006)

200. K Lander, H Hill, M Kamachi, E Vatikiotis-Bateson: It's not what you say but the way you say it: matching faces and voices. *J Exp Psychol Hum Percept Perform* 33, 905–914 (2007)

201. S Schweinberger, D Robertson, J Kaufmann: Hearing facial identities. *Q J Exp Psychol* 60, 1446–1456 (2007)

202. D Robertson, S Schweinberger: The role of audiovisual asynchrony in person recognition. *Q J Exp Psychol* 63, 23–30 (2010)

203. S Schweinberger, N Kloth, D Robertson: Hearing facial identities: brain correlates of face--voice integration in person identification. *Cortex* 47, 1026–1037 (2011)

204. H Blank, A Anwander, K von Kriegstein: Direct structural connections between voice- and face-recognition areas. *J Neurosci* 31, 12906–12915 (2011)

205. S Schall, S.J Kiebel, B Maess, K von Kriegstein: Early auditory sensory processing of voices is facilitated by visual mechanisms. *Neuroimage* 77, 237–347 (2013)

206. V Bruce, A Young: Understanding face recognition. *Br J Psychol* 77, 305–327 (1986)

207. A Burton, V Bruce, R Johnston: Understanding face recognition with an interactive activation model. *Br J Psychol* 81, 361–380 (1990)

208. S Brédart, T Valentine, A Caldor, L Gassi: An interactive activation model of face naming. *Q J Exp Psychol* A 48, 466–486 (1995)

209. T Valentine, T Brennan, S Brédart: The cognitive psychology of proper names. London, United Kingdom (1996)

210. A Burton, V Bruce, P Hancock: From Pixels to People: A Model of Familiar Face Recognition. *Cogn Sci* 21, 1–31 (1999)

211. A Young, V Bruce: Understanding person perception. *Br J Psychol* 102, 959–974 (2011)

212. H Ellis, D Jones, N Mosdell: Intra-and inter-modal repetition priming of familiar faces and voices. *Br J Psychol* 88, 143–156 (1997)

Key Words: Speaker Recognition, Voice Perception, Psychophysics, Neuroimaging, Review

Send correspondence to: Samuel R Mathias, Center for Computational Neuroscience and Neurotechnology, Boston University, 677 Beacon Street, Boston, MA 02215, Tel: 617-353-5760, Fax: 617-353-77550, E-mail: smathias@bu.edu