
Article:
Young, Andy orcid.org/0000-0002-1202-6297, Fruehholz, Sascha and Schweinberger, S.R. (2020) Face and voice perception : understanding commonalities and differences. Trends in Cognitive Sciences. ISSN 1364-6613

https://doi.org/10.1016/j.tics.2020.02.001

Authors' version of manuscript, as accepted for publication in *Trends in Cognitive Sciences*.

Date of acceptance: 3rd February 2020.

Face and voice perception: understanding commonalities and differences

Andrew W. Young (University of York, UK),
Sascha Frühholz (University of Zurich, Switzerland,
and University of Oslo, Norway)
and Stefan R. Schweinberger (University of Jena, Germany, and Swiss Center for Affective Sciences, University of Geneva, Switzerland)

Keywords: Face perception, voice perception, identity, emotion, speech

Contact details:

Andy Young: andy.young@york.ac.uk
Sascha Frühholz: sascha.fruehholz@uzh.ch
Stefan Schweinberger: stefan.schweinberger@uni-jena.de

Acknowledgements: Sascha Frühholz is supported by the Swiss National Science Foundation (SNSF PP00P1_157409/1 and PP00P1_183711/1). Stefan Schweinberger has been supported by grants from the Deutsche Forschungsgemeinschaft (DFG grant ref. SCHW 511/18-1 and SCHW 511/22-1) and by the Swiss Center for Affective Sciences at the University of Geneva, Switzerland, hosting a sabbatical leave in summer 2019. Although views expressed here remain the authors' responsibility, we are grateful for helpful comments from the Editor and four anonymous reviewers.

Abstract

Faces and voices are of high importance in interpersonal communication, and there are notable parallels between face and voice perception. However, these parallels do not sit entirely comfortably with the full range of available evidence. This review evaluates parallels between the functional and neural organisation of face and voice perception, whilst locating these in the context of ways in which faces and voices also differ. It takes the discussion to the next level by asking *why* these commonalities and differences exist. A novel synthesis is offered, grounded in the interaction between intrinsic characteristics of faces and voices and the demands of everyday life, showing how the pattern of findings reflects a system that can respond optimally to different everyday demands.

Highlights

- Similarities in functional organisation have led to the proposal of parallel, largely independent processing streams for voices and faces. Linked to this conception is the idea that the voice can be considered to be a kind of 'auditory face'.
- However, neuroimaging studies show a strong contribution of multimodal regions that respond both to voices and to faces. Closer examination of neuropsychological and behavioural studies supports this form of organisation.
- The contributions of differences between how relatively invariant information (such as a person's identity) and more rapidly changing information (such as their emotional state) must be represented need to be carefully considered.
- Understanding the everyday demands of different tasks involving voice and face perception offers a resolution in which these serve as strong drivers of the optimal functional and neural organisation.

Understanding face and voice perception

Human communication involves complex patterns of signals originating primarily from the face, voice, and body [1]. Whilst much of this communication takes the form of propositional speech, faces and voices can also convey common forms of information concerning a person's gender, age, identity, health and emotional state, and they create impressions of warmth, competence and other social traits [2,3]. Much modern research has therefore focussed on communication from the face and voice [4-9].

This review aims to strengthen theoretical approaches to key properties of face and voice perception. It offers a synthesis of existing evidence based on evaluating **functional perspectives** (**see Glossary**) and **neural perspectives** in light of the overarching background of what can be communicated through faces and voices, the different contingencies this creates, the demands of everyday life, and the ways in which these act as determinants of a communicative system that has to balance the needs of the sender and recipient.

Functional and neural perspectives

Communication from the face and voice can be considered from different perspectives that can be subdivided into those where the focus of interest is primarily functional (the organisation of cognitive processes and components underlying face and voice perception) and those where the focus of interest is primarily in terms of underlying neural mechanisms (the brain regions and neural pathways involved in perceiving faces and voices).

From a functional perspective, comparisons between face and voice perception have led to notable parallels and the useful and important theoretical suggestion that the voice can be considered as a kind of 'auditory face' [4,6,9,10] with a comparable functional organisation, as shown in Box 1.

BOX 1 ABOUT HERE

Box 1: *The voice as an auditory face in functional modelling*

Nonetheless, this view has not gone unchallenged, and in some ways it does not sit entirely happily with perspectives centred on underlying brain regions [8,16]. As noted in Box 2, areas involved in voice perception show a relatively lower degree of functional specificity than regions involved in face perception. Unsurprisingly, then, a recent meta-analysis has questioned the 'voice as an auditory face' interpretation, calling instead for a more modality-specific perspective [16]. At the same time, though, Box 2 highlights the fact that there are more brain regions with multi-modal responses to both faces and voices than are suggested by Box 1, implying clear limits to the modality-specific view. Theoretical progress depends on reconciling such differences.

BOX 2 ABOUT HERE

Box 2: *Brain regions involved in face and voice perception*

A longstanding debate involves the relation between functional and neural levels of explanation [30-32]. Whilst it is logically possible that these involve entirely different types of discourse that will not map on to each other, this theoretically possible scenario seems unlikely to turn out to be the case. Instead, it is reasonable to begin by expecting (and being reassured by) some degree of correspondence [2,30]. The fact that at present functional and neural models do not sit entirely comfortably together thus presents an interesting and unresolved theoretical puzzle.

This review therefore evaluates the ways in which the functional and neural organisation of face and voice perception offer parallels and the ways in which

they differ. Moreover, and importantly, it takes the discussion to the next level by asking *why* these commonalities and differences exist. It offers a novel perspective grounded in the interaction between intrinsic characteristics of faces and voices and the demands of everyday life.

The nature of visual and auditory transmission means there are a number of general differences between facial and vocal communication that form an essential background. The voice allows communication when the face is not visible, the voice can be silent even when the face is visible, and in most everyday contexts a person can hear their own voice but can't see their own face. Moreover, nonverbal communication often arises as a concomitant of propositional speech that may itself inform about a speaker´s thoughts and feelings or be influenced by the intention to elicit a specific response. Verbal and nonverbal content are thus often linked, and evidence from brain electrophysiology suggests that emotional word content affects early stages of processing [33]. However, this close coupling is not inevitable; whilst it is difficult to ignore emotional speech intonation, it is somewhat easier to ignore speech content when instructed to do so [34]. Because the present review focuses on commonalities and differences beween the face and the voice, addressing such interactions between verbal content and nonverbal cues in detail would be beyond its scope, but they have been discussed elsewhere [35].

In addition, some other more specific differences between faces and voices are immediately clear. For example, the important role of facial eye gaze in signalling someone´s focus of attention [2] has no direct counterpart in the voice, though it is true that other functions of gaze direction such as signalling conversational turn-taking [36] can equally well be communicated through the voice and that distinct neural mechanisms mediate the use of eye contact in spoken conversation [37].

As well as this background context involving the nature of visual or auditory signalling, a key factor involves the underlying time course of the communicative signals themselves [2,12,13]. In particular, some properties

signal things about an individual that are relatively stable across time (for example their identity, age, or gender) whilst other properties signal things that change from moment to moment (such as how they are feeling, or what they are saying). The consequences of this distinction are of critical theoretical importance.

By considering such influences, the review develops a new synthesis of evidence which is summarised in Figure 1 and explained in the following sections.

FIGURE 1 ABOUT HERE

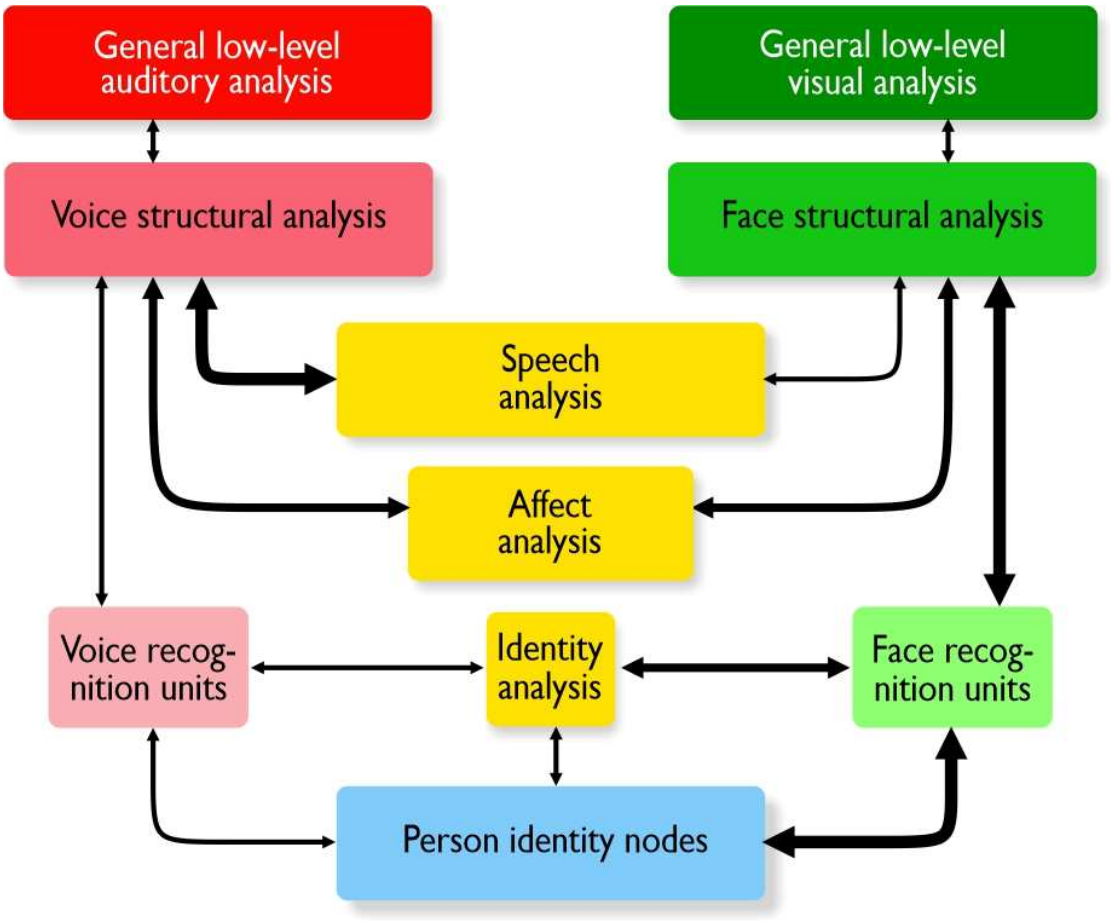

Figure 1: *Revised functional model of face and voice perception. The model shows components that involve relatively unimodal responses to voices in*

reddish highlighting, and components that involve relatively unimodal responses to faces in green highlighting. Relatively low-level analyses are indicated by more intense colour. Components shown in yellow highlighting involve multimodal perceptual integration for speech, affect, and identity, with the size of boxes and the weight of arrows indicating the relative importance of perceptual integration for speech, affect, and identity. A component that involves post-perceptual representations at the level of episodic or semantic processing is highlighted in blue. Line thicknesses are used to indicate the relative weighting of different functional connections at higher processing levels.

Recognition of identity

A useful place to begin is by considering the recognition of identity, which is often thought to offer a paradigmatic example of the need to determine a relatively stable personal characteristic [12,23,38-40]; identity does not change during a social encounter. The principal purpose of recognising a person's identity is, of course, to allow someone to bring to mind pertinent previously experienced facts and episodes that enable them to interact with the recognised person as a unique individual with a known background history [11,39,41,42]. This requirement has been hugely increased by modern life through the sheer number of familiar individuals most of us now know and recognise; a number which dwarfs the capacity needed in prehistoric times [43,44].

Evidence of modality-specific face and voice recognition

Whilst person identity can be determined from facial, vocal or body cues [9,45-47], it is clear that there exist parallel forms of **neuropsychological deficits** of face and voice recognition following brain injury [48]. In some cases these involve severely impaired recognition of the identities of familiar faces (**prosopagnosia**) or severely impaired recognition of familiar voices (**phonagnosia**). Such deficits can be strikingly selective; in prosopagnosia, the severe face recognition deficit is not accompanied by a correspondingly

severe problem in recognising familiar voices, and in phonagnosia the severe voice recognition deficit is not accompanied by a correspondingly severe problem in recognising familiar faces. These patterns strongly suggest a degree of modality-specificity in face and voice recognition mechanisms [3,45,47,48]. Consistent with these observations, neuropsychological and functional brain imaging studies implicate substantially different underlying brain regions for initial stages of face and voice recognition [3,24,49-51].

This modality-dependent organisation may itself be driven by natural environments, in which a person's face is often seen before their voice is heard. Indeed, familiar face recognition is remarkably efficient; not only is recognition so readily achieved from even a severely degraded image of a familiar face that any contribution from the voice is not needed in most contexts [39.40], but identification is also more efficient for faces than voices [52]. Occasionally, though, the voice is heard before the face is seen; a circumstance that again puts a premium on a modality-specific mechanism that does not demand multimodal input [13].

Cross-talk between face and voice recognition

Despite the substantial degree of modality-specificity of face and voice recognition noted above, commonalities in the neural coding of faces and voices have also been found. These include suggestions of some degree of cross-modal face-voice integration at an early stage of identity processing [53,54] and phenomena that were first demonstrated for faces, but have later been observed for voices too, such as **contrastive adaptation aftereffects** [55,56]. Such findings are consistent with the suggestion of a unified coding strategy for faces and voices [9,57]. Nonetheless, whilst there may be overarching common properties, the neural coding principles underlying the analysis and representation of faces and voices must also differ in important respects. Face perception can use a mix of spatially distributed (eyes, nose, mouth) as well as temporal information (facial movements and the effects of saccades), whereas voice perception inevitably depends heavily on the integration of temporal information (acoustic information over time).

Functional demands of face and voice recognition

Importantly, there are other notable parallels between the functional demands of face and voice recognition. These primarily involve the need to recognise familiar identities across substantial natural variation [3,39,40,58-61]. For face recognition, images or views of a familiar individual differ hugely across changes in lighting, viewpoint, expression, and even the time of day or a person's state of health [39,59,62]. Recognition of a familiar face across these enormous visual changes presents a substantial challenge. At first sight this challenge is exacerbated by the fact that some of this variability in appearance is identity-specific, in the sense that the way one person's face can differ across different views will not be the same as the ways in which another face differs [63]. However, it turns out that this identity-specific variability can facilitate recognition as long as the recognition mechanism can learn to encompass its implications [58,64,65]. In this sense, recognising faces involves being able to group very different images together (i.e. to recognise that despite the differences they represent the same identity) rather than (as is more often assumed) merely being able to tell similar images apart [59,62,66].

In the same way, the sound of a familiar voice will vary depending on a comparably wide range of factors involving local acoustics, the prevailing context (e.g. a job interview vs. an informal conversation), the person's emotional state, their health, whether they are talking, joking, asking questions or making nonverbal sounds, and so on [3,67,68]. Recent work shows that these differences in the sound of the same voice work in much the same way as differences in the appearance of the same face to drive a corresponding form of organisation that can achieve recognition of familiar individuals across widely differing examples of the same voice [60,61,68]. For this reason, recognition of familiar faces or familiar voices by most neurologically normal individuals far outstrips recognition of unfamiliar faces and voices [39,40,60], and impairments in recognition of familiar faces or voices can occur even in the context of relatively preserved recognition or

matching of their unfamiliar counterparts [24,45,47]. In effect, a person can learn about the idiosyncratic variability of each familiar face and each familiar voice in order to recognise them. In contrast, recognising unfamiliar faces and unfamiliar voices presents a quite different set of challenges because their idiosyncratic identity-specific variability is by definition unknown, leading most of us to make substantial errors in recognising unfamiliar people [40,60,65].

This combination of functional demands, then, leads both to a substantial degree of modality-specificity of face and voice recognition and to a substantial degree of parallel functional organisation. That said, and as already noted, it is also clear that the frequent co-occurrence of facial and vocal communicative signals in the natural environment (i.e. the fact that people are often seen and heard at the same time) does lead to a degree of cross-modal integration that is evident in some circumstances [23-25,53,54,69-72].

It is also evident that modality-specific face and voice recognition mechanisms must access modality-independent semantic and episodic information about the recognised individuals. This is highlighted in Box 1 and Box 2. It would be inefficient and probably ineffective to store separately from seeing someone and from talking to them the informative things that have been learnt about an individual; their likes and dislikes, their past history, the shared experiences, and so on. The same set of memories need to be accessible in any social interaction, and with sufficient flexibility to allow pertinent facts to be quickly brought to mind. You might be talking to your friend about their holidays one minute and then their new job the next, but of course their identity does not change across the shift in context.

This point is clearly seen in case studies of neuropsychological deficits involving structural damage to anterior temporal lobes, which point strongly to the existence of modality-independent forms of loss of memory for people in which severe deficits affect the retrieval of identity-specific information about a familiar individual from their face, voice, and even their name

[26,27,48,49,73,74]. The same underlying functional architecture is evident in widely-used functional models [11,75-78].

Recognition of emotion

Having considered the factors that shape functional mechanisms underlying recognition of identity, it is instructive to contrast these with recognition of emotion. Critically, whereas a person's identity remains consistent throughout a social encounter, their emotions can change from moment to moment [2,12,13]. In consequence a strikingly different type of functional organisation arises for emotion recognition, where it has become evident that comprehension of facial and vocal cues is closely integrated [13,79,80] in a way that is not captured adequately by Box 1.

Cross-modal integration characterises facial and vocal emotion recognition

Most past research on emotion recognition was predicated on an assumption that modality-specific mechanisms underlie the recognition of facial expressions and vocal expressions, in much the same way that largely modality-specific mechanisms underlie the recognition of facial and vocal identity [8,81]. Based on this working assumption, the neuropsychological research literature has been dominated by studies that look exclusively at problems in facial expression recognition [8]. These have shown double dissociations between the recognition of facial identity and facial expressions that can be interpreted as consistent with the view that different mechanisms are involved in analysing identity and expression [81,82].

However, when the perspective is broadened to look also at recognition of vocal expressions, it turns out that patients with neuropsychological deficits following brain injury that affect emotion recognition invariably have problems that affect both facial and vocal expressions [13,79,82,83]. Even though the deficits in such cases may affect the recognition of some emotions more than others - for example compromising recognition of fear and anger after amygdala damage [83-85] and disgust after insula damage [83,85] - they

have a comparable impact on the recognition of these emotions from faces, voices and bodies. This is the case whether acoustic cues to emotion are conveyed through nonverbal sounds (laughing, crying, screaming, etc.), through tone of voice (prosody), or even through music [83-87]. Indeed, such impairments clearly affect the experience of the corresponding emotions themselves [85,86,88,89]. Moreover, facial and vocal emotion recognition impairments also co-occur in other disorders including Parkinson´s disease and autism [90-92].

Functional brain imaging studies also show the importance of a multimodal contribution to emotion recognition by demonstrating that audio-visual signals of emotion are integrated at a relatively early stage of processing and that posterior superior temporal cortex (**pSTC)**, including posterior parts of the superior temporal sulcus (**STS**), plays a critical role in this integration [13,79,93]. For example, **MEG** reveals integrative responses to faces and voices in pSTC within the first 200 ms of stimulus onset [94,95].

Contrasting patterns of functional organisation for recognising identity and emotion from face and voice

The functional organisation of emotion recognition described above is strikingly different from the functional organisation involved in recognising familiar identities. This is especially clearly seen from neuropsychological studies. Prosopagnosia and phonagnosia tend to be modality-specific, affecting only recognition of identity from the face or voice, but in each case virtually all familiar faces or familiar voices are affected [3,45,47,48]. In contrast, emotion recognition deficits are cross-modal, affecting recognition from the face and the voice, but they can compromise the recognition of some emotions more than others [13,79,82-89]. Brain imaging and MEG studies add the important information that this audio-visual integration arises at a relatively early stage of processing signals of emotion [13,79,93-95].

Functional demands of facial and vocal emotion recognition

From the standpoint of everyday functional demands, the organisation of emotion recognition is primarily driven by the need to resolve transient signals about mental and emotional states that require rapid readjustment of the perceiver's interpretation and intentions [13,96]; if your friend's mood shifts suddenly from apparent happiness to anger, it becomes a priority to understand why. In this context, pooling all sources of available information can maximise the speed and accuracy of responses. An additional strong driver is that many signals of emotion are themselves inherently somewhat ambiguous [97-99], but these ambiguities arise in different ways that will often make the signals complementary [100]. A multimodal mechanism that can integrate facial and vocal cues with contextual constraints [101] thus represents an optimal solution to these environmental and behavioural demands. Indeed, there is substantial overlap between the underlying structure of emotions recognised from faces and voices, with comparable patterns of confusion between different emotions despite the fundamental differences in how these are signalled [102].

Emotion and other changeable social signals

Although implications of the interpretation of propositional speech have mostly been set aside for this review, there is an instructive parallel to be made between recognition of emotion and early stages of speech perception.

Cross-modal integration between faces and voices in speech perception

The multimodal organisation of emotion recognition should perhaps not have come as such a surprise, as it has been known for decades that multimodality is especially clearly seen in the case of early stages of speech perception. Whilst it is natural to think of speech perception in terms of decoding the acoustic signal, and of course we know that purely acoustic analyses can support speech perception when we listen to a radio or talk to someone on the telephone, there is none the less substantial evidence that seeing

someone's facial movements can make an important contribution to understanding what they are saying. A compelling example of this is the **McGurk illusion** [103,104], which shows that in hearing what someone says we make use of the correspondence between movements of their lips (and tongue) and the speech sounds.

An important clue to why this happens comes from classic work on speech perception in noise demonstrating a substantial improvement when the speaker's face was visible [105]. In considering the cause of this effect, it was noted that cues such as place of articulation are particularly hard to hear but easy to see on a talker's lips. Conversely, other features can be hard to see, but easy to hear [105]. Hence the voice and face can to some degree offer complementary information to support early stages of speech perception; this offers a clear parallel with the complementarity between facial and vocal signals of emotion already noted [100,101]. Considered more generally, it seems that because speech signals involve changes across different time-scales that have to be decoded as they unfold [14,15], integrating complementary information from face and voice offers an optimal way of dealing with these temporal constraints. In fact, whether the brain opts for integration or segregation is dependent on the degree of audio-visual synchrony [106,107] and studies of infants suggest that sensitivity to these audio-visual correspondences begins early in life [108,109].

Functional brain imaging studies offer an important contribution here by identifying brain regions that are involved in lipreading. Multimodal responses in audio-visual integration studies using talking faces are consistently found in left pSTC, including left posterior STS and possibly the adjacent left superior temporal gyrus [22,110.111]. The potential importance of left pSTC to audio-visual integration has been confirmed by demonstrating that **TMS** to this region disrupts the McGurk effect [112].

Parallels between the functional demands of emotion recognition and speech perception

Compared to emotion recognition, the temporal demands of speech perception are even higher - requiring disambiguation of cues that may only last for milliseconds [14,107] - and the nature of the complementarity between auditory cues and cues that can be read from movements of the lips and tongue is correspondingly strongly established. But the underlying drivers of needing to be able to interpret and respond to rapidly changing and partially ambiguous signals are much the same for emotion and for speech.

Pushing this point further, it is also evident that integration of facial and vocal sources of information characterises other everyday tasks, such as determining the authenticity of someone's expressed emotion [113], in which there are ambiguous signals that may need a fairly rapid response. An important example involves forming a first impression of an unfamiliar individual, where there is again some overlap between the information that can be gained from faces and voices [114,115] and we seem to integrate this so readily that it is difficult to attend selectively to what is being gleaned from the face or voice itself [7,116].

Concluding remarks and future directions

This review has shown that that there are indeed parallels between face and voice perception that make the 'voice as an auditory face' metaphor a useful and informative place to begin. However, it is time also to take into account strong differences between face and voice perception on the cognitive and neural levels that are best understood as consequences of behavioural and environmental demands. Box 3 offers an overview of some of these differing functional demands across the recognition of identity and emotion and their implications.

BOX 3 ABOUT HERE

Box 3: *Functional demands of identity recognition and emotion recognition*

Whilst understanding the functional demands created by our everyday lives helps resolve fundamental issues of neural organisation involving the relatively unimodal or multimodal contributions of different brain regions, it is clear that there are also unresolved issues. For example, differences in the neural network for emotional processing across modalities obtained in functional brain imaging studies, in which the amygdala is involved in emotion processing both from faces and voices [84,85,121-124] but its response is more consistently noted for emotion perceived from faces than for voices [20,80,112]. Moreover, posterior STS should not be overinterpreted as the only region involved in audiovisual integration of speech; it clearly forms part of a larger network that is apparent in studies that have used different criteria [125-128]. This network includes other regions along the superior temporal sulcus and superior temporal gyrus that include classical auditory areas.

An interesting and at present largely unexplored issue here concerns the role of familiarity in interpreting highly changeable social signals. Whilst familiarity is central to recognition of identity across different types of perceptual change [39,40,60-65,68] it seems less central to interpreting changeable signals. People do not make substantially more errors in understanding the speech or emotions of unfamiliar individuals than they make in understanding the speech or emotions of familiar individuals, though there are clearly some benefits to familiarity [111,129]. For example, familiarisation with individual talkers' voices promotes better speech recognition in noise [130], and facial familiarity or identity [131,132] may exert a small effect on emotion perception. One reason why the role of familiarity for interpreting highly changeable signals is small (relative to its prominent role for recognition) seems to be that whilst there are identity-specific differences that can to some extent limit the universality of characteristics that underlie social signals, these are relatively

small compared to the identity-specificity of perceptual signals of personal identity [133]. This has the useful consequence of facilitating the many interactions with strangers that characterise much of modern life (or simply watching television).

A potentially important and at present under-researched source of insight may also come from studies of the factor structure underlying individual differences in ability to recognise identity and emotion from faces and voices, which are beginning to offer complementary support to the type of model shown in Figure 1 [134,135].

Considering together the complementary influences of the intrinsic differences between faces and voices and the impact of the different demands of everyday life thus gives a richer understanding of properties that underlie the functional and neural organisation of how faces and voices are used in interpersonal perception. Approaching the issues in this way shows that the pattern of findings reflects a system that is well-tuned to respond optimally to different types of everyday demand, and that this point is key to understanding parallels, differences and convergences between face and voice perception. These insights offer a new perspective to drive the agenda for further advances (see **Outstanding Questions**).

Outstanding Questions

- Some brain regions show strongly multimodal responses in which information from vocal and facial signals is integrated at relatively early stages of processing, but how is this achieved?
- Integration is achieved despite many obvious differences between the basic perceptual mechanisms demanded by voices and faces, so are there common higher-order coding principles that facilitate this integration? If so, what are they?
- Integration is most useful and most strongly evident for rapidly changing signals such as emotion, rather than for fixed characteristics such as identity, so why do some findings point to cross-talk between recognition of voice identity and face identity?
- Does cross-talk between face and voice identity have functional significance, or is it better considered to be a by-product of testing the limits of a system whose organisation is primarily unimodal but has high interactivity across its components?
- Is the extent of unimodal or multimodal organisation of identity recognition influenced to some degree by the associative contingencies of the natural environment (i.e. the fact that a person's face and voice will so often be experienced together in spatiotemporal correspondence)?
- Emotions and speech need to be interpreted for both familiar and unfamiliar individuals, so to what extent is integration of audiovisual information in speech and emotion perception influenced by familiarity with the speaker?

Glossary

Contrastive adaptation aftereffects: Transient changes in perception induced by exposure to stimuli with particular characteristics. A classic example is the motion aftereffect, in which viewing a unidirectionally moving stimulus induces illusionary perception of motion in the opposite direction in a static scene. Within the last two decades, contrastive adaptation aftereffects have been demonstrated for the perception of complex stimuli, including emotion or identity perception in faces and voices.

Functional perspectives: Used here to indicate approaches that attempt to delineate the organisation of cognitive processes and components underlying face and voice perception; often this is done with a 'box and arrow' type model. 'Functional' in this sense refers to how a particular function is organised - sometimes called 'cognitive architecture' in the research literature - rather than to function in the sense of utility to the organism.

McGurk illusion: An audio-visual illusion in which a video showing the face of a person saying one phoneme (for example, "ga") is combined with a different phoneme (for example, "ba") on the soundtrack. Remarkably, the heard phoneme can then correspond neither to the auditory nor the visual part of the video, but reflects a fusion of the two (heard as "da" in the example used here).

MEG: Magnetoencephalography. A measure of changes in the magnetic field around the skull resulting from neural activity that has excellent temporal resolution and is capable of localising activity to sources such as posterior **STS** that are close to the sensors.

Neural perspectives: Used here to indicate approaches where the focus of interest is primarily in terms of the brain regions and neural pathways involved in perceiving faces and voices.

Neuropsychological deficits: Used here to refer to consequences of brain injury.

Prosopagnosia: A severe problem in recognising familiar faces that cannot be explained by more general visual or intellectual difficulties. This can

be due to acquired brain injury (**neuropsychological deficits**) or congenital causes. Usually, even the most familiar faces are not recognised in neuropsychological cases. Studies of such cases have been very influential.

Phonagnosia: A severe problem in recognising familiar voices that cannot be explained by more general auditory or intellectual difficulties. This can be due to acquired brain injury (**neuropsychological deficits**) or congenital causes. Although less widely reported and investigated than **prosopagnosia**, such cases are also of substantial importance.

pSTC: The posterior part of superior temporal cortex, including posterior superior temporal sulcus and adjacent left superior temporal gyrus. Often noted to be involved in integration of visual and auditory signals in speech and in emotion perception.

STS: Superior temporal sulcus. A major sulcus in the temporal lobe.

TMS: Transcranial magnetic stimulation. A strong local magnetic field can disrupt neural activity in the affected region.

References

1. Schweinberger, S.R. and Burton, A.M. (Eds.) (2011) Person perception 25 years after Bruce and Young (1986). *Br. J. Psychol. 102*, 695-974.
2. Bruce, V. and Young, A. (2012) *Face perception*. Hove, East Sussex: Psychology Press.
3. Schweinberger, S.R. *et al.* (2014) Speaker perception. *WIREs Cogn. Sci. 5*, 15-25.
4. Belin, P. *et al.* (2004) Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci. 8*, 129-135.
5. Belin, P. *et al.* (2011) Understanding voice perception. *Br. J. Psychol. 102*, 711-725.
6. Belin, P. (2017) Similarities in face and voice cerebral processing. *Vis. Cogn. 25*, 658-665.
7. Mileva, M. *et al.* (2018) Audiovisual integration in social evaluation. *J. Exp. Psychol. Hum. Percept. Perform. 44*, 128-138.
8. Schirmer, A. and Adolphs, R. (2017) Emotion perception from face, voice, and touch: comparisons and convergence. *Trends Cogn. Sci. 21*, 216-218.
9. Yovel, G. and Belin, P. (2013) A unified coding strategy for processing faces and voices. *Trends Cogn. Sci. 17*, 263-271.
10. Campanella, S. and Belin, P. (2007) Integrating face and voice in person perception. *Trends Cogn. Sci. 11*, 535-543.
11. Bruce, V. and Young, A. (1986) Understanding face recognition. *Br. J. Psychol. 77,* 305-327.
12. Haxby, J.V. *et al.* (2000) The distributed human neural system for face perception. *Trends Cogn. Sci. 4*, 223-233.
13. Young, A.W. (2018) Faces, people and the brain: the 45th Sir Frederic Bartlett Lecture. *Q. J. Exp. Psychol. 71*, 569-594.
14. Eimas, P.D. and Corbit, J.D. (1973) Selective adaptation of linguistic feature detectors. *Cognitive Psychol. 4*, 99-109.

15. McGettigan, C. and Scott, S.K. (2012) Cortical asymmetries in speech perception: what's wrong, what's right and what's left? *Trends Cogn. Sci. 16*, 269-276.
16. Schirmer, A. (2018) Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. *Soc. Cogn. Affect. Neurosci. 13*, 1-13.
17. Kanwisher, N. (2017) The quest for the FFA and where it led. *J. Neurosci. 37*, 1056-1061.
18. Pernet, C.R. *et al.* (2015) The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage 119*, 164-174.
19. Frühholz, S. and Belin, P. (2019) The science of voice perception. In S. Frühholz & P. Belin (Eds). *The Oxford handbook of voice perception*. Oxford, UK: Oxford University Press, 3-14.
20. Frühholz, S. *et al.* (2014) The role of the medial temporal limbic system in processing emotions in voice and music. *Progr. Neurobiol. 123*, 1-17.
21. Janak, P.H. and Tye, K.M. (2015) From circuits to behaviour in the amygdala. *Nature 517*, 284-292.
22. Calvert, G.A. (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex 11*, 1110-1123.
23. Perrodin, C. *et al.* (2015) Who is that? Brain networks and mechanisms for identifying individuals. *Trends Cogn Sci. 19*, 783-796.
24. Maguiness, C. *et al.* (2018) Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia 116*, 179-193.
25. Tsantani, M. *et al.* (2019) Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. *NeuroImage 201*, 116004.
26. Gainotti, G. (2014) The neuropsychology of familiar person recognition from face and voice. *Psychol. Belg. 54*, 298-309.

27. Cosseddu, M. *et al.* (2018) Multimodal face and voice recognition disorders in a case with unilateral right anterior temporal lobe atrophy. *Neuropsychology 32*, 920-930.
28. Mormann, F. *et al.* (2015) Neurons in the human amygdala encode face identity but not gaze direction. *Nat. Neurosci. 18*, 1568-1570.
29. Wang, S. *et al.* (2017) The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nat. Comm. 8*, 14821.
30. Henson, R.N.A. (2005) What can functional neuroimaging tell the experimental psychologist? *Q. J. Exp. Psychol. 58A*, 193-233.
31. Coltheart, M. (2006) What has functional neuroimaging told us about the mind (so far)? *Cortex 42*, 323-331.
32. Page, M.P.A. (2006) What can't functional neuroimaging tell the experimental psychologist? *Cortex 42*, 428-433.
33. Scott, G.G. *et al.* (2009) Early emotion word processing: evidence from event-related potentials. *Biol. Psychol. 80,* 95-104.
34. Filippi, P. *et al.* (2017) More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing. *Cognition Emotion 31*, 879-891.
35. Schirmer, A, and Kotz, S.J. (2006) Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends Cogn. Sci. 10,* 24-30.
36. Kleinke, C.L. (1986) Gaze and eye contact: a research review. *Psychol. Bull. 100*, 78–100.
37. Jiang, J. *et al.* (2017) Neural mechanisms of eye contact when listening to another person talking. *Soc. Cogn. Affect. Neurosci. 12*, 319-328.
38. Bernstein, M. and Yovel, G. (2015) Two neural pathways of face processing: a critical evaluation of current models. *Neurosci. Biobehav. R. 55*, 536-546.
39. Young, A.W. and Burton, A.M. (2017) Recognizing faces. *Curr. Dir. Psychol. Sci. 26*, 212-217.
40. Young, A.W. and Burton, A.M. (2018) Are we face experts? *Trends Cogn. Sci. 22,* 100-110.

41. Gobbini, M.I. and Haxby, J.V. (2007) Neural systems for recognition of familiar faces. *Neuropsychologia 45*, 32-41.
42. Wiese, H. *et al.* (2019) A robust neural index of high face familiarity. *Psychol. Sci. 30*, 261-272.
43. Layton, R. *et al.* (2012) Antiquity and social functions of multilevel social organization among human hunter- gatherers. *Int. J. Primatol. 33*, 1215-1245.
44. Jenkins, R. *et al.* (2019) How many faces do people know? *P. Roy. Soc. B-Biol. Sci. 285,* 20181319.
45. Barton, J.J.S. and Corrow, S.L. (2016) Recognizing and identifying people: a neuropsychological review. *Cortex 75*,132-150.
46. Yovel, G. and O'Toole, A.J. (2016) Recognizing people in motion. *Trends Cogn. Sci. 20*, 383-395.
47. Biederman, I. *et al.* (2018) The cognitive neuroscience of person identification. *Neuropsychologia 116*, 205-214.
48. Neuner, F. and Schweinberger, S.R. (2000) Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain Cognition 44*, 342-366.
49. Blank, H. *et al.* (2014) Person recognition and the brain: merging evidence from patients and healthy individuals. *Neurosci. Biobehav. R. 47*, 717-734.
50. Liu, R. R. *et al.* (2016) Voice recognition in face-blind patients. *Cereb. Cortex 26*, 1473-1487.
51. Roswandowitz, C. *et al.* (2018) Obligatory and facultative brain regions for voice-identity recognition. *Brain 141*, 234-247.
52. Hanley, J.R. and Damjanovic, L. (2009) It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory 17*, 830-839.
53. Schweinberger, S.R. *et al.* (2011) Hearing facial identities: brain correlates of face-voice integration in person identification. *Cortex 47,* 1026-1037.
54. Schall, S. *et al.* (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage 77,* 237-245.

55. Webster, M. A. *et al.* (2004) Adaptation to natural facial categories. *Nature 428*, 557-561.
56. Schweinberger, S.R. *et al.* (2008) Auditory adaptation in voice perception. *Curr. Biol. 18*, 684-688.
57. Watson, R. *et al.* (2014) Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. *J. Neurosci. 34*, 6813-6821.
58. Bruce, V. (1994) Stability from variation: the case of face recognition. *Q. J. Exp. Psychol. 47A*, 5-28.
59. Burton, A.M. (2013) Why has research in face recognition progressed so slowly? The importance of variability. *Q. J. Exp. Psychol. 66*, 1467-1485.
60. Lavan, N. *et al.* (2019) Flexible voices: identity perception from variable vocal signals. *Psychon. B. Rev. 26*, 90-102.
61. Lavan, N. *et al.* (2019) Listeners form average-based representations of individual voice identities. *Nat. Commun. 10,* 2404.
62. Jenkins, R. *et al.* (2011) Variability in photos of the same face. *Cognition, 121*, 313–323.
63. Burton, A.M. *et al.* (2016) Identity from variation: representations of faces derived from multiple instances. *Cognitive Sci., 40,* 202–223.
64. Kramer, R.S.S. *et al.* (2017) Robust social categorization emerges from learning the identities of very few faces. *Psychol. Rev. 124*, 115-129.
65. Kramer, R.S.S. *et al.* (2018) Understanding face familiarity. *Cognition, 172*, 46-58.
66. Andrews, S. *et al.* (2015) Telling faces together: learning new faces through exposure to multiple instances. *Q. J. Exp. Psychol. 68*, 2041-2050.
67. Lavan, N. *et al.* (2018) Impoverished encoding of speaker identity in spontaneous laughter. *Evol. Hum. Behav. 39,* 139-145.
68. Lee, Y. *et al.* (2019) Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am. 146*, 1568-1579.

69. von Kriegstein, K. *et al.* (2006) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cereb. Cortex 16*, 1314-1322.
70. von Kriegstein, K. *et al.* (2008) Simulation of talking faces in the human brain improves auditory speech recognition. *P. Natl. Acad. Sci. USA 105*, 6747-6752.
71. Maguiness, C. and von Kriegstein, K. (2017) Cross-modal processing of voices and faces in developmental prosopagnosia and developmental phonagnosia. *Vis. Cogn. 25*, 644-657.
72. Schweinberger, S.R. and Robertson, D.M.C. (2017) Audiovisual integration in familiar person recognition. *Vis. Cogn. 25*, 589-610.
73. Ellis, A.W. *et al.* (1989) Loss of memory for people following temporal lobe damage. *Brain, 112*, 1469-1483.
74. Hanley, J.R. *et al.* (1989) Defective recognition of familiar people. *Cogn. Neuropsychol. 6*, 179-210.
75. Burton, A.M. *et al.* (1990) Understanding face recognition with an interactive activation model. *Br. J. Psychol. 81*, 361-380.
76. Burton, A.M. *et al.* (1999) From pixels to people: a model of familiar face recognition. *Cognitive Sci. 23*, 1-31.
77. Schweinberger, S.R. and Neumann, M.F. (2016) Repetition effects in human ERPs to faces. *Cortex 80* 141-153.
78. Young, A.W. and Burton, A.M. (1999) Simulating face recognition: implications for modelling cognition. *Cogn. Neuropsychol. 16*, 1-48.
79. Calder, A.J. and Young, A.W. (2005) Understanding the recognition of facial identity and facial expression. *Nat. Rev. Neurosci. 6,* 641-651.
80. Dricu, M. and Frühholz, S. (2016) Perceiving emotional expressions in others: activation likelihood estimation meta-analyses of explicit evaluation, passive perception and incidental perception of emotions. *Neurosci. Biobehav. R. 71*, 810-828.
81. Young, A.W. *et al.* (1993) Face perception after brain injury: selective impairments affecting identity and expression. *Brain 116*, 941-959.
82. Keane, J. *et al.* (2002) Face and emotion processing in frontal variant frontotemporal dementia. *Neuropsychologia 40*, 655-665.

83. Calder, A.J. *et al.* (2001) Neuropsychology of fear and loathing. *Nat. Rev. Neurosci. 2*, 352-363.
84. Scott, S.K. *et al.* (1997) Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature 385*, 254-257.
85. Sprengelmeyer, R. *et al.* (1999) Knowing no fear. *P. Roy. Soc. B-Biol. Sci. 266*, 2451-2456.
86. Calder, A.J. *et al.* (2000) Impaired recognition and experience of disgust following brain injury. *Nat. Neurosci. 3*, 1077-1078.
87. Gosselin, N. *et al.* (2007) Amygdala damage impairs emotion recognition from music. *Neuropsychologia 45*, 236-244.
88. Broks, P. *et al.* (1998) Face processing impairments after encephalitis: amygdala damage and recognition of fear. *Neuropsychologia 36*, 59-70.
89. Feinstein, J.S. *et al.* (2011) The human amygdala and the induction and experience of fear. *Curr. Biol. 21*, 34-38.
90. Gray, H.A. and Tickle-Degnen, L. (2010) A meta-analysis of performance on emotion recognition tasks in Parkinson's disease. *Neuropsychology 24*, 176-191.
91. Philip, R.C.M. *et al.* (2010) Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychol. Med. 40*, 1919-1929.
92. Frühholz, S. and Staib, M. (2017) Neurocircuitry of impaired affective sound processing: a clinical disorders perspective. *Neurosci. Biobehav. R. 83*, 516-524.
93. Gao, C. *et al.* (2019) The brain basis of audiovisual affective processing: evidence from a coordinate-based activation likelihood estimation meta-analysis. *Cortex,* in press.
94. Hagan, C.C. *et al.* (2009) MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *P. Natl. Acad. Sci. USA 106*, 20010-20015.
95. Hagan, C.C. *et al.* (2013) Involvement of right STS in audio-visual integration for affective speech demonstrated using MEG. *PLoS One 8*, e70648.

96. Oatley, K. and Johnson-Laird, P.N. (2014) Cognitive approaches to emotions. *Trends Cogn. Sci. 18*, 131-140.
97. Russell, J.A. and Fehr, B. (1987) Relativity in the perception of emotion in facial expressions. *J. Exp. Psychol. Gen. 116*, 223-237.
98. Israelashvili, J. *et al.* (2019) When emotions run high: a critical role for context in the unfolding of dynamic, real-life facial affect. *Emotion 19*, 558-562.
99. Barrett, L.F. *et al.* (2019) Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychol. Sci. Publ. Int. 20*, 1-68.
100. Sauter, D.A. (2017) The nonverbal communication of positive emotions: an emotion family approach. *Emot. Rev. 9*, 222-234.
101. Sander, D. *et al.* (2018) An appraisal-driven componential approach to the emotional brain. *Emot. Rev. 10*, 219-231.
102. Kuhn, L. et al. (2017) Similar representations of emotions across faces and voices. *Emotion 17*, 912-937.
103. McGurk, H. and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature 264* 746-748.
104. Magnotti, J.F. *et al.* (2018) A causal inference explanation for enhancement of multisensory integration by co-articulation. *Sci. Rep. 8*, 18032.
105. Miller, G.A. and Niceley, P. (1955) An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am. 27*, 338-352.
106. Gau, R. and Noppeney, U. (2016) How prior expectations shape multisensory perception. *NeuroImage 124*, 876-886.
107. Lee, H. and Noppeney, U. (2014) Temporal prediction errors in visual and auditory cortices. *Curr. Biol. 24*, R309-R310.
108. Kuhl, P.K. and Meltzoff, A.N. (1982) The bimodal perception of speech in infancy. *Science 218*, 1138-1141.
109. Patterson, M.L. and Werker, J.F. (2003) Two-month-old infants match phonetic information in lips and voice. *Developmental Sci. 6*, 191-196.

110. Calvert, G.A. *et al.* (2001) Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage 14* 427-438.
111. Riedel, P. *et al.* (2015) Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex 68*, 86-99.
112. Beauchamp, M.S. *et al.* (2010) fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J. Neurosci. 30*, 2414-2417.
113. Lavan, N. *et al.* (2015) I thought that I heard you laughing: contextual facial expressions modulate the perception of authentic laughter and crying. *Cognition Emotion 29*, 935-944.
114. Oosterhof, N. N. and Todorov, A. (2008) The functional basis of face evaluation. *P. Natl. Acad. Sci. USA 105*, 11087-11092.
115. McAleer, P. *et al.* (2014) How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS One 9*, e90779.
116. Rezlescu, C. *et al.* (2015) Dominant voices and attractive faces: the contribution of visual and auditory information to integrated person impressions. *J. Nonverbal Behav. 39*, 355-370.
117. Du, S. *et al.* (2014) Compound facial expressions of emotion. *P. Natl. Acad. Sci. USA 111*, E1454–E1462.
118. Ekman, P. (1972) Universals and cultural differences in facial expressions of emotion. In J.K. Cole (Ed.), *Nebraska symposium on motivation, 1971 (pp. 207-283)*. Lincoln, Nebraska: University of Nebraska Press.
119. Ekman, P. (1992) An argument for basic emotions. *Cognition Emotion 6*, 169-200.
120. Vytal, K. and Hamann, S (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *J. Cognitive Neurosci. 22*, 2864-2885.
121. Frühholz, S. *et al.* (2015) Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. *P. Natl. Acad. Sci. USA 112*, 1583-1588.

122. Frühholz, S. *et al.* (2016) The sound of emotions: towards a unifying neural network perspective of affective sound processing. *Neurosci. Biobehav. R. 68***,** 96-110.
123. Fecteau, S. *et al.* (2007) Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage 36,* 480-487.
124. Bestelmeyer, P. *et al.* (2014) Adaptation to vocal expressions reveals multistep perception of auditory emotion. *J. Neurosci. 34,* 8098-8105.
125. Hall, D.A. *et al.* (2005) Reading fluent speech from talking faces: typical brain networks and individual differences. *J. Cognitive Neurosci. 17*, 939-953.
126. Macaluso, E. *et al.* (2004) Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage 21*, 725-732.
127. Wright, T.M. *et al.* (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex 13*, 1034-1043.
128. McGettigan, C. *et al.* (2017) *You talkin' to me?* Communicative talker gaze activates left-lateralized superior temporal cortex during perception of degraded speech. *Neuropsychologia 100,* 51-63.
129. Walker, S. *et al.* (1995) Facial identity and facial speech processing: familiar faces and voices in the McGurk effect. *Percept. Psychophys. 57*, 1124-1133.
130. Nygaard, L.C. *et al.* (1994) Speech perception as a talker-contingent process. *Psychol. Sci. 5*, 42-46.
131. Kahana-Kalman, R. and Walker-Andrews, A.S. (2001) The role of person familiarity in young infants' perception of emotional expressions. *Child Dev. 72*, 352-369.
132. Martens, U. *et al.* (2010) Parallel processing in face perception. *J. Exp. Psychol. Hum. Percept. Perform. 36*, 103-121.
133. Mileva, M. *et al.* (2019) Understanding facial impressions between and within identities. *Cognition 190,* 184-198.
134. Palermo, R. *et al.* (2013) New tests to measure individual differences in matching and labelling facial expressions of emotion, and their

association with ability to recognise vocal emotions and facial identity. *PLoS One 8*, e68126.

135. Connolly, H. *et al.* (2020) Emotion recognition ability: evidence for a supramodal factor and its links to social cognition. *Cognition 197*, 104166.

Box 1: The voice as an auditory face

Functional models offer a simplified and potentially falsifiable overview of the organisation of cognitive processes that underlie a particular task.

FIGURE I ABOUT HERE

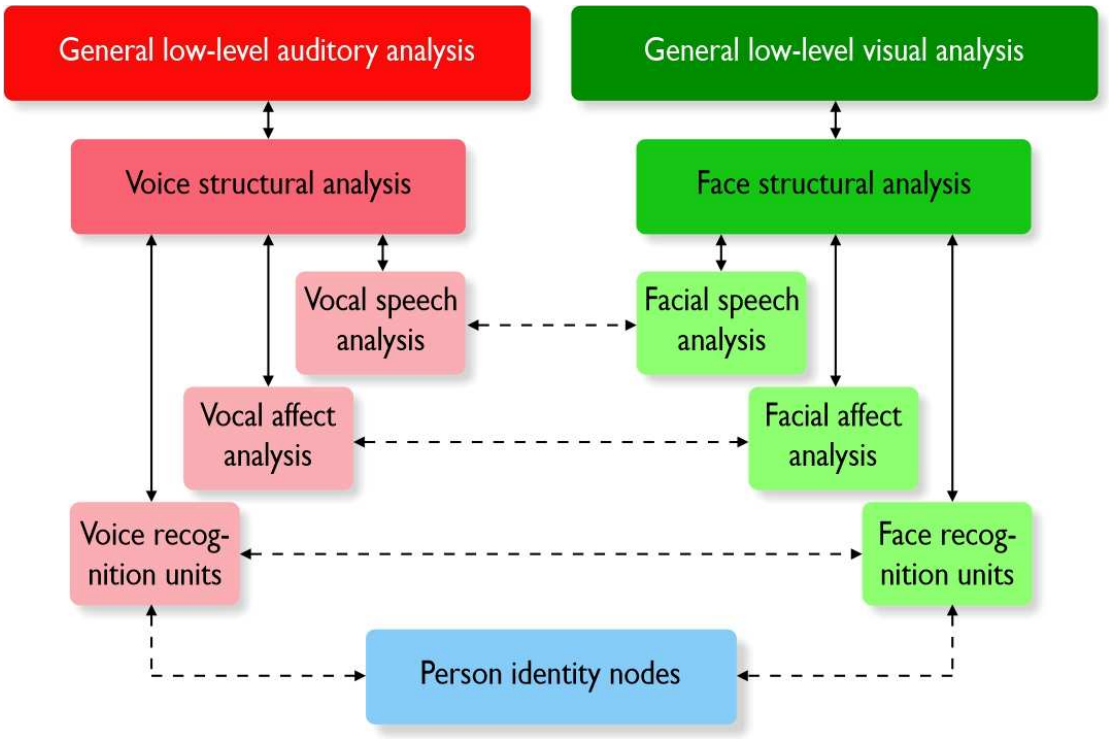

***Figure I**: An influential functional model of face and voice perception. Adapted from Belin et al. [4] and reproduced with publisher's permission (RightsLink 4681990357267).*

Figure I shows an adapted version of an influential model of face and voice perception [4] based on considering the functional organisation of voice perception as offering a close parallel to a widely-used functional model of face perception [11]. The model proposes that voice perception (highlighted in reddish tints) and face perception (green tints) each involve distinct modality-specific pathways for recognising familiar people (voice recognition units and

face recognition units), recognising emotion (vocal affect analysis and facial affect analysis), and for speech perception (vocal speech analysis and facial speech analysis). Relatively low-level analyses are indicated by more intense colour. Access to post-perceptual episodic and semantic representations of identity-specific information is highlighted in blue.

In this model, any audiovisual perceptual integration of speech, affect, or identity is implemented via direct links between modality-specific representations, rather than via potentially multimodal perceptual analysis components (see later, Box 3). Modality-specific perceptual recognition of a familiar identity from voice or face then converges on multimodal episodic and semantic representations of identity-specific information (via person identity nodes).

The suggested parallel organisation of voice and face perception underscores the idea that the voice can be considered to be a kind of 'auditory face'. This type of model does not itself maintain that comparable analyses of voices and faces (for example, those involved in recognising emotion) are achieved in precisely the same way, but it is consistent with proposals for common coding mechanisms [9].

Such models reflect increasing interest in how voices, faces, and other sources of information (such as bodies) interact in interpersonal perception [1]. A key point in understanding how this may happen is to note that the different analyses themselves have different temporal demands. Because a person's identity is stable across a social encounter [12], recognition of identity (through voice or face recognition) has relatively low temporal demands, although efficient identity recognition is undoubtedly beneficial at the onset of an encounter. In contrast, emotions can change from moment to moment and these changes have important social implications, meaning that vocal and facial affect must be constantly monitored and have relatively high temporal demands [13]. Speech has the highest temporal demands of all, with differences between consonants involving only tens of milliseconds [14] and other meaningful changes across a range of time-scales [15].

Box 2: Brain regions involved in face and voice perception

FIGURE II ABOUT HERE

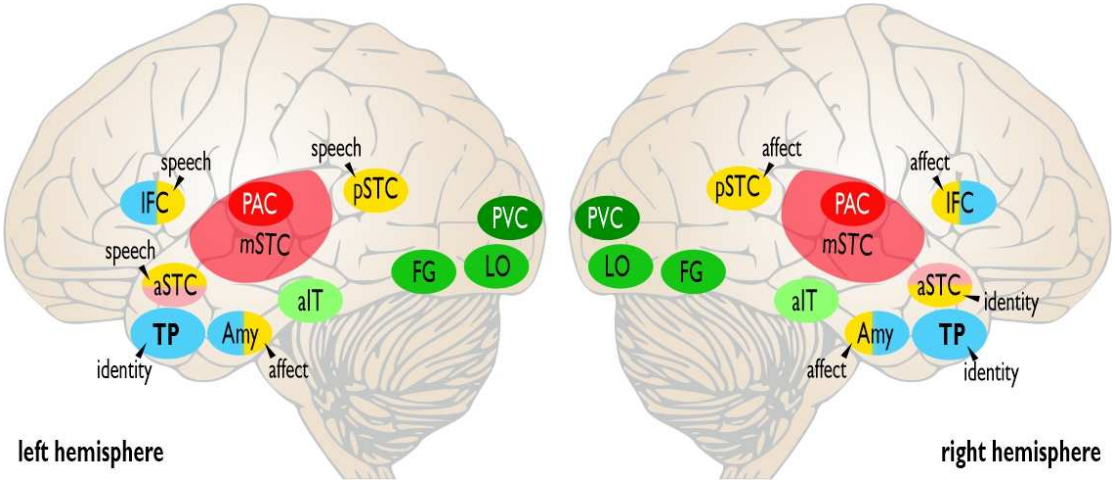

Figure II: *Brain regions responsive to faces, voices, or both. [Abbreviations: PAC primary auditory cortex; a/m/pSTC anterior/mid/posterior superior temporal cortex; PVC primary visual cortex; FG fusiform gyrus; LO lateral occipital cortex; aIT anterior inferior temporal lobe; Amy amygdala; TP temporal pole; IFC inferior frontal cortex; MFC medial frontal cortex].*

Figure II shows locations of unimodal and multimodal brain regions in the left and right hemispheres that are involved in decoding speech, affect, and identity information from facial and vocal signals. These are colour-coded with reddish tints for unimodal voice regions, green for unimodal face regions, yellow for regions with potential audiovisual responses, and blue for regions where information is likely to be represented at a post-perceptual level.

These regions for voice and face processing are in some cases defined through higher responses to faces than to other visual stimuli [17] and stronger activation to voices than to other auditory stimuli [18,19]. In other cases, regions such as the amygdala are defined anatomically [20,21].

Functional brain imaging studies demonstrate that, after basic sensory processing in primary sensory cortices (PAC, PVC), distinct cortical brain regions show strong and predominant unimodal responses to voices (the temporal voice area, located predominantly in an extended region of mSTC [18,19], indicated by red highlighting in Figure II) and unimodal responses to faces (the fusiform face area, located in FG, and occipital face area located in LO [17]; green highlighting) that provide a basic structural analysis of vocal and facial signals, respectively. While face-responsive regions usually appear as localised subregions of the cortical visual system [17], voice-responsive regions typically show spatial extension across subregions of the auditory cortex (PAC, mSTC) [18,19]. Overall, areas involved in voice perception show a relatively lower degree of regional functional specificity than regions involved in face perception [17-19].

Other regions show responses to both faces and voices (yellow highlighting), thus forming candidates for integrating vocal and facial signals involving speech, affect or identity [3,8,22-25]. Regions likely to also have relatively post-perceptual responses are highlighted in blue [19,23,26-27].

This neural network for the unimodal and multimodal processing of voice and face signals shows more regions with multimodal (yellow) responses than might have been expected from Box 1. Whilst functional (Box I) and neural levels of description (Box 2) need not necessarily be in full correspondence, this relative preponderance of multimodally responsive brain regions needs to be explained and if possible reconciled with the functional approach.

The major source of current evidence underlying Figure II involves fMRI data with limited temporal resolution. It can therefore be anticipated that new evidence from methods with high temporal resolution (EEG, MEG, or intracranial recordings) will help further to integrate functional and neural levels of description [28,29].

Box 3: Functional demands of identity recognition and emotion recognition

Table 1 summarises contrasting demands of identity recognition and emotion recognition from voices and faces. These clearly differ markedly along dimensions that involve the core task requirements, their behavioural implications, task complexity, the roles of within-person variability, and temporal demands. Taken together, these demands create compelling drivers of functional differences between the ways voices and faces are used in the recognition of identity and emotion. In effect, the everyday demands act as key determinants of the optimal functional organisation.

TABLE 1 ABOUT HERE

The points made concerning emotion recognition apply equally (and perhaps more strongly) to initial stages of speech perception, but Table 1 uses identity and emotion to point up key contrasts.

These differing demands also need to be considered alongside differences between the information that is most readily accessed from voices and faces. Often, social signals can be ambiguous [97-99], but the nature of these ambiguities can differ between voices and faces in ways that make them complementary [100]. When this is the case, pooling information from voices and faces through a fundamentally multimodal mechanism will optimise the speed and accuracy of responses.

The points summarised in Table 1 and discussed extensively in the main text of this review have been incorporated into the revised model of face and voice perception presented in Figure 1. In particular, Figure 1 emphasises the relative importance of multimodal perceptual integration for speech, affect, and identity.

The model's architecture necessarily represents a degree of simplification in detail. For example, the main text makes clear how person familiarity is central to identity processing, whereas familiarity provides a much smaller but systematic benefit to emotion and speech processing that is not represented graphically in Figure 1. This small benefit of familiarity for emotion and speech can be thought of as resulting from relatively efficient structural analysis of familiar perceptual patterns, though further research is needed to address this point.

Table 1: Different demands of identity recognition and emotion recognition.

	IDENTITY RECOGNITION	EMOTION RECOGNITION
Core requirement	Recognise faces or voices across different emotions [2,11,13]	Recognise emotions across different identities [2,11,13]
Behavioural implications	Access identity-specific episodic and semantic information to allow appropriate interaction [11,39-42]	Modulate ongoing priorities: an obvious change in mood signals that something needs immediate attention [96]
Task complexity	High: most people can recognise thousands of familiar individuals from their faces [44] and a substantial number of voices [24]	Moderate: limited number of basic emotions [99,117-120], but these are often dependent on context for correct interpretation [97-99] and some expressions involve blends of different emotions [117]
Role of within-person variability across different instances	Largely meaningless for identity [2,11,13]	Highly meaningful [2,11,13]
Temporal demands	Relatively low (except at onset): once established, identity does not change during a social encounter [2,11-13]	Relatively high: constant monitoring needed because moods and feelings can change in any direction from moment to moment [2,11-13]