Combining Voice and Face Content in the Primate Temporal Lobe

Catherine Perrodin[1] & Christopher I. Petkov[2*]

1. Institute of Behavioural Neuroscience, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom; e-mail: c.perrodin@ucl.ac.uk; phone: +44 (0) 207 679 21011

2. Institute of Neuroscience, Newcastle University Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, United Kingdom; e-mail: chris.petkov@ncl.ac.uk; phone: +44 (0) 191 208 3467

*Corresponding author

Keywords: communication, visual, auditory, comparative, superior-temporal sulcus, neurons, voice areas, face areas, neurophysiology, fMRI, spikes, oscillations

Abstract

The interactions of many social animals critically depend on identifying other individuals to approach or avoid. Recognizing specific individuals requires extracting and integrating cross-sensory indexical cues from richly informative communication signals, such as voice and face content. Knowledge on how the brain processes faces and voices as unisensory or multisensory signals has grown: neurobiological insights are now available not only from human neuroimaging data but also from comparative neuroimaging studies in nonhuman animals, which identify the correspondences that can be made between brain processes in humans and other species. These advances have also had the added benefit of establishing animal models in which neuronal processes and pathways can be interrogated at finer neurobiological scales than possible in humans. This chapter overviews the latest insights on neuronal representations of voice and face content, including information on sensory convergence sites and pathways that combine multisensory signals in the primate temporal lobe. The information synthesized here leads to a conceptual model whereby sensory integration of voice and face content depends on temporal lobe convergence sites, which are a midway processing stage and a conduit between audio-visual sensory processing streams and frontal cortex.

1.1 Neurobiological Processing of Voice and Face Content in Communication Signals

Far from being redundant, information from different sensory inputs is complementary and expedites behavioral recognition of an object or entity. Yet how the brain achieves sensory integration remains a challenging question to answer, in part because it has become evident that multisensory neural interactions are distributed, taking place between multiple sites throughout the brain. Although a handful of multisensory association areas are often emphasized in summaries and reviews for brevity (Schroeder and Foxe 2002; Stein and Stanford 2008), it is well accepted that multisensory influences abound from early cortical and subcortical sensory processing stages and beyond (Damasio 1989; Ghazanfar and Schroeder 2006).

Thus the task for neuroscientists has been steadily shifting away from a focus on particular sensory convergence sites towards an emphasis on identifying the neural multisensory influences and transformations that occur between sites along particular processing pathways (Yau et al. 2015; Bizley et al. 2016). Moreover, comparing the forms of multisensory convergence seen at different brain sites can pinpoint common principles of multisensory integration and identify how specializations in neural multisensory interactions may occur, such as duplication with differentiation. In this chapter evidence on neuronal representations and multisensory interactions along pathways involved in processing voice and face content will be considered. Finally, this chapter will conclude by identifying obvious epistemic gaps that inspired readers might be encouraged to empirically shrink in the future, helping to advance neurobiological knowledge.

Initial insights into how the brain processes identity-related information were obtained in the visual modality. Neurons responding stronger to faces than to other non-face objects were first identified in the monkey inferior temporal (IT) cortex (Bruce et al. 1981; Perrett et al. 1982). Subsequently, human neuroimaging studies identified face-category preferring regions in the fusiform gyrus, occipital cortex and adjacent visual areas (Sergent et al. 1992; Kanwisher et al. 1997). Shortly after, functionally homologous face-sensitive regions in the monkey inferior bank and fundus of the superior-temporal sulcus (STS) were identified (Logothetis et al. 1999; Tsao et al. 2006). More recently, as we next consider in more detail, complementary information from the auditory modality has become available (for a review: Perrodin et al. 2015b). Together these developments have opened pathways for understanding how multisensory (voice and face) content is combined in the brain (see Fig. 1, also see Plakke and Romanski, Chapt. 7).

1.2 Voice Sensitive Brain Regions in Humans, Monkeys and other Mammals

Human neuroimaging studies aiming to shed light on the processing of auditory communication signals tend to focus on the neurobiology of speech and language, which is a fundamental aspect of human communication (Hickok and Poeppel 2007; Binder et al. 2009). In parallel, work in carnivore, rodent, and primate models aims to unravel the neurobiological substrates for referential social communication (i.e., “what” was vocalized), a likely evolutionary precursor upon which human vocal communication evolved (Ghazanfar and Takahashi 2014; Seyfarth and Cheney 2014).

More recently, investigators focusing on understanding identity-related content ("who" vocalized) have identified voice-sensitive regions in the human brain using functional magnetic-resonance imaging (fMRI). The approach of comparing how the brain responds to voice versus non-voice content in communication signals is analogous to neurobiological studies in the visual domain comparing responses to face versus non-face objects (Belin et al. 2004). These and other studies have identified the presence of several voice-sensitive clusters in the human temporal lobe, including in the superior-temporal gyrus/sulcus (Belin et al. 2000; von Kriegstein et al. 2003).

However, it is known that human voice regions also strongly respond to speech (Fecteau et al. 2004), and that both speech and voice content can be decoded from largely overlapping areas in the superior portions of the human temporal lobe (Formisano et al. 2008). These observations left open the possibility that human voice and speech processes are so functionally intertwined that human brain specializations for voice processing may have occurred alongside those for speech, raising the question whether 'voice regions' would be evident in nonhuman animals.

This question of whether nonhuman animals have voice-sensitive regions as humans do was answered in the affirmative initially in rhesus macaques (*Macaca mulatta*), an Old World monkey species, with evidence in other primate species and mammals following shortly thereafter. The macaque monkey fMRI study identified temporal lobe voice-sensitive regions that are more strongly activated by voice than non-voice sounds (Petkov et al. 2008). Moreover, of the several fMRI identified voice-sensitive clusters in the monkey superior temporal lobe, the most anterior one was found to be particularly sensitive to "who" vocalized rather than "what" was vocalized, forging a more direct link to human fMRI studies on voice-identity sensitive processes in the anterior temporal lobe (Belin and Zatorre 2003; McLaren et al. 2009). More recently, Andics and colleagues (2014) imaged domesticated dogs with fMRI to reveal voice-preferring regions in the temporal lobe of these carnivores, broadening the evolutionary picture. Relatedly, an fMRI study in marmosets (*Callithrix jacchus*, a New World monkey species) identified temporal lobe regions that respond more strongly to conspecific vocalizations than other categories of sounds (Sadagopan et al. 2015), which in the future could be interrogated for voice content sensitivity.

In laboratory animals that are established neurobiological models, the fMRI identified voice-sensitive clusters can be targeted for neurophysiological study at a fundamental scale of neural processing, e.g. at the level of single neurons. Moreover, the identification of voice-sensitive regions in nonhuman animals also helps to forge links to analogous processes in the visual system.

1.3 Voice-Sensitive Neurons in the Ventral Auditory Processing Stream

The anterior voice-sensitive fMRI cluster in rhesus macaques is located in hierarchically higher neuroanatomically delineated cortical regions (Galaburda and Pandya 1983; Saleem and Logothetis 2007). These areas reside in the anterior portion of the superior-temporal plane (aSTP, the dorsal and anterior surface of the temporal lobe, see Fig. 1). The anterior temporal lobe voice-sensitive cluster falls somewhere between the 4th or 5th stage of processing in the auditory cortical hierarchy, rostral to the tonotopically organized core (1º), 'belt' (2º), and parabelt (3º) areas (Rauschecker 1998; Kaas and Hackett 2000). The anatomical localization of the voice area in the aSTP places it at an intermediate level in the ventral auditory 'object' processing stream (Rauschecker and Tian 2000; Romanski 2012). Other downstream cortical regions interconnected with the aSTP include the superior-temporal gyrus, sulcus, and frontal cortex; see Fig. 1 and (Perrodin et al. 2011).

Although the fMRI results on voice-identity sensitive processing and the corresponding anatomical findings identify the anterior aSTP region as higher-order cortex, the auditory feature selectivity displayed by neurons in this region was unknown prior to electrophysiological recordings from the fMRI identified clusters in a neurobiological model, here macaques. In the initial neuronal recording studies from the anterior voice sensitive cluster, neuronal spiking responses were modulated by differences in the vocal features of the auditory stimuli, such as call type, caller identity, and caller species (Perrodin et al. 2014). In particular, the results revealed a distinct subpopulation of voice-sensitive neurons, which accounted for much of the observed sensitivity to caller identity features ("who" vocalized).

Thus, the neuronal recordings showed that the responses of neurons in voice-sensitive clusters can simultaneously be sensitive to the category of voices over non-voice stimuli, and to auditory features of individual stimuli within the voice category. This dual sensitivity in the recordings is paralleled at a very different spatiotemporal scale by the human and monkey fMRI results, which in turn show that voice-sensitive clusters are sensitive to both the categorical distinction between voice versus non-voice stimuli *and* to specific conspecific voices within the category of voices (Belin and Zatorre 2003; Petkov et al. 2008).

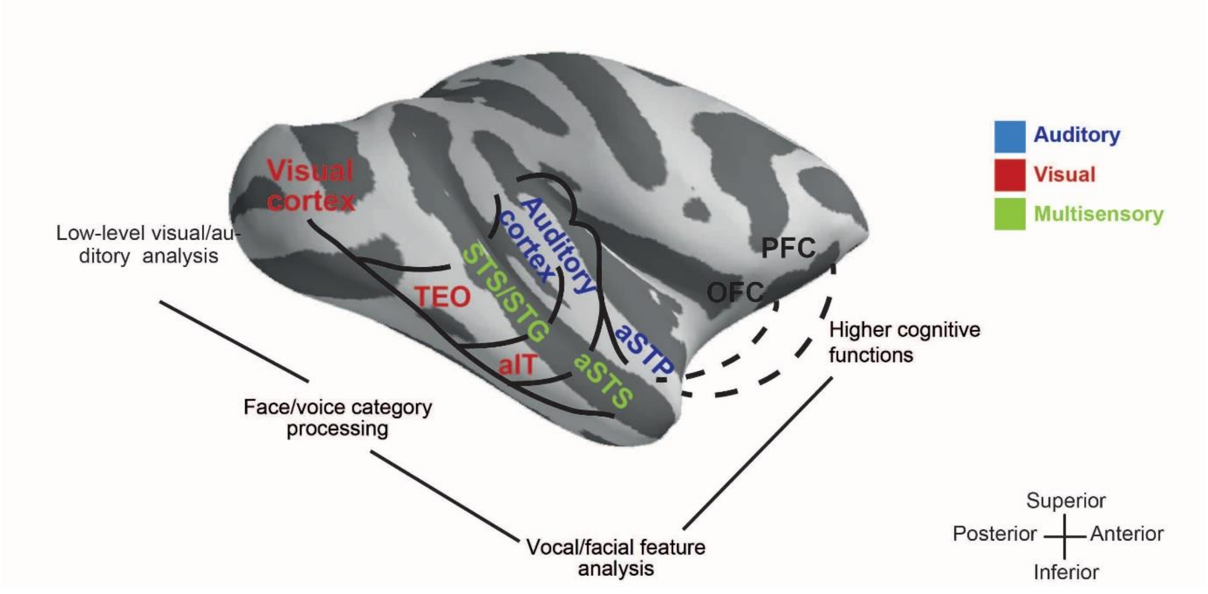

Fig. 1 Auditory voice and visual face processing pathways in the primate brain
This illustrates a primate model of ascending auditory and visual cortical streams rendered on a rhesus macaque brain. It features early sensory cortices, processing stages extracting face content in visual areas of the inferior temporal lobe (*TEO* and *aIT*), and auditory regions of the anterior superior-temporal plane/gyrus (*aSTP/STG*) extracting voice-related content. Multisensory interactions are possible

between voice and face processing regions including by way of association areas along the superior-temporal sulcus (*STS*) and frontal cortex (*PFC*: prefrontal cortex, *OFC*: orbitofrontal cortex). The cortical regions are interconnected via bidirectional pathways of inter-regional projections, including feed-forward and feed-back projections to the auditory and visual processing streams (dotted and solid black lines). *M:* medial, *p:* posterior, *a:* anterior. Reproduced from (Perrodin et al. 2015b).

The neuronal sensitivity to auditory vocal features in the voice area was compared to that in a different part of the anterior temporal lobe, the anterior upper bank of the superior temporal sulcus (aSTS, see Fig. 1), which has long been considered to be multisensory since it belongs to higher-order association cortex (Stein and Stanford 2008). More posterior regions of the STS were known to contain both auditory and visually responsive clusters of neurons (Beauchamp et al. 2004; Dahl et al. 2009); also see Beauchamp Chapt. 8. The neuronal recordings in the aSTS confirmed that a substantial proportion of neurons in this area are driven by sounds, but the results also showed that neurons in this area are not very sensitive to auditory vocal features, unlike the auditory-feature sensitive neurons in the aSTP voice-sensitive cluster (Perrodin et al. 2014).

By comparison to these observations from neural recordings in the aSTP and aSTS, neurons in the ventro-lateral prefrontal cortex, which are hierarchically further along the ventral processing stream, show a sensitivity to complex acoustical features of vocalizations, such as call-type (Gifford et al. 2005) and caller identity (Plakke et al. 2013). Why certain areas in the processing pathways to frontal cortex show less auditory feature specificity than others is a topic that will be visited later in this chapter, after considering more of the available information.

Converging evidence from the visual and auditory modalities in humans and monkeys points to anterior subregions of the temporal lobe being involved in the processing of identity-related features. In the visual domain, face regions in the anterior inferior-temporal lobe (aIT; see Fig. 1) are particularly sensitive to identity-related content in humans (Kriegeskorte et al. 2007; Tsao and Livingstone 2008) and monkeys (Freiwald and Tsao 2010; Morin et al. 2014). Likewise in the auditory modality more anterior temporal lobe areas are sensitive to identity-related content in communication sounds both in humans (e.g., Belin and Zatorre 2003; von Kriegstein et al. 2003) and monkeys (Petkov et al. 2008). A number of theoretical models also highlight the anterior temporal lobe as a region containing sites sensitive to voice or face identity-related content (Bruce and Young 1986; Campanella and Belin 2007). However, since anterior temporal lobe sites are nodes in a broader network processing voice and face content (Fecteau et al. 2005; Tsao et al. 2008), other more posterior voice or face sensitive sites in the temporal lobe are undoubtedly involved in ways that need to be better understood. For instance, more posterior superior temporal lobe regions can also be sensitive to identity-related information regardless of the sensory modality (Chan et al. 2011; Watson et al. 2014).

In summary, the results from neuronal recordings in the voice sensitive aSTP specify the auditory response characteristics of neurons in this region of the ventral processing stream and distinguish these characteristics in relation to those from neurons in the adjacent association cortex of the anterior STS. However, despite the surface resemblance, sensory processing of voices and faces in the auditory and visual modalities, respectively, does not seem to be identical, as we consider in the next section where we ask: do voice cells exist and, if so, are they direct analogs of visual face-sensitive neurons?

1.4 Do Voice-Sensitive Regions Contain 'Voice Cells,' and, If So, How Do Their Responses Compare to 'Face Cells' in the Visual System?

An initial question while interrogating neuronal responses in voice-sensitive cortex, given the evidence for face-sensitive cells in the visual system, is do 'voice cells' exist? Arguably, at the cellular level the auditory system tends to show relatively less tangible organizational properties than those seen in the visual and somatosensory systems for a host of fundamental sensory processing features (Griffiths et al. 2004; King and Nelken 2009). The better established view of auditory cortical processing is that many auditory functions are supported by neuronal processes that are distributed across populations of auditory neurons and do not require topographical maps or individual cells with high feature selectivity (Bizley et al. 2009; Mizrahi et al. 2014).

Thus another open question was whether fMRI-identified voice clusters contain highly voice-content sensitive neurons, or 'voice' cells. Voice cells could be defined as neurons that exhibit two-fold greater responses to voice vs. non-voice stimuli, in direct analogy to how face cells have been defined: This was the approach of Perrodin and colleagues (2011) in searching for 'voice cells' in the anterior voice-identity sensitive fMRI cluster in macaque monkeys. They first used an auditory voice localizer borrowed from the earlier monkey fMRI study (Petkov et al. 2008), which allowed them to identify neurons within the fMRI voice clusters that were preferentially sensitive to the voice category of stimuli. The voice localizer stimulus set included a collection of macaques voices from many individuals (many voices), and two comparison categories containing either animal vocalizations and voices from other species, or a set of natural sounds. All stimuli were subsampled from a larger stimulus set so that the selected sounds from each category were matched in multiple low-level acoustical features. Using these sounds as stimuli, the researchers observed a modest proportion (~25% of the sample) of neurons within the aSTP that could be defined as voice cells.

Yet, comparisons of the proportions and response characteristics of the 'voice cells' in relation to what is known about face cells suggest that voice cells may not be direct analogs to face cells. For instance, visual studies of face clusters identified a high density of face cells in monkey face sensitive fMRI regions (Tsao et al. 2006; Aparicio et al. 2016). This very high clustering (>90%) of face cells in these fMRI face clusters is in stark contrast to the much more modest (~25%) clustering of voice cells (Perrodin et al. 2011). Further, the voice-sensitive cells in the anterior temporal lobe fMRI cluster are remarkably stimulus-selective, responding to only a small proportion or just a few of the voices within the category of stimuli (Perrodin et al. 2011). This high neuronal selectivity of voice cells seems to diverge from the functional encoding properties that have been reported for face cells, whereby face cells typically respond more broadly to the majority of faces in the stimulus set (Hasselmo et al. 1989; Tsao et al. 2006).

The high stimulus-selectivity of voice cells is on par with the level of selectivity measured in neurons responding to conspecific vocalizations in the ventro-lateral prefrontal cortex (Gifford et al. 2005; Romanski et al. 2005); both temporal and frontal sites show higher stimulus selectivity than that reported for neurons in and around primary auditory cortex (Tian et al. 2001; Recanzone 2008), an auditory region in the insula (Remedios et al. 2009), or parts of the superior-temporal gyrus (Russ et al. 2008). Thus, in relation to reports on face cell selectivity, so far the available evidence raises the intriguing possibility that voice cells are not direct auditory analogs of face cells, which may reflect specialization under different evolutionary pressures in the auditory versus visual domains (Miller and Cohen 2010; Perrodin et al. 2011).

2 How Multisensory is the Anterior Voice-Sensitive Temporal Cortex?

The authors involved in the initial monkey neuroimaging and electrophysiological studies on voice regions and voice cells were rather bold in their initial claims identifying the anterior temporal fMRI-identified cluster as a voice sensitive area. In the initial monkey neuronal recording studies multisensory stimulation conditions were not used to rule out or rule in that the region is multisensory rather than auditory. Moreover, human fMRI studies had already shown evidence for both functional crosstalk and direct structural connections between voice- and face-sensitive regions (von Kriegstein et al. 2005; Blank et al. 2011), which suggests that the neuroanatomical pathways for the exchange of visual face and auditory voice information are in place. Other potential sources of visual input into the auditory STP include cortico-cortical projections from visual areas (Bizley et al. 2007; Blank et al. 2011), feedback projections from higher association areas such as inferior frontal cortex (Romanski et al. 1999a; Romanski et al. 1999b), and the upper-bank of the STS (Pandya et al. 1969; Cappe and Barone 2005). Multisensory projections with subcortical origins could also directly or indirectly influence cross-modal interactions, such as those from the suprageniculate nucleus of the thalamus or the superior colliculus.

A number of electrophysiological studies have directly evaluated the multisensory influences of face input on voice processing in nonhuman primates at a number of cortical sites, including posterior auditory regions closer to primary and adjacent auditory cortex (Ghazanfar et al. 2005; Kayser et al. 2008), and higher-order association regions such as the STS (Chandrasekaran and Ghazanfar 2009; Dahl et al. 2009) or ventro-lateral prefrontal cortex (Sugihara et al. 2006; Romanski 2007); also see Plakke and Romanski, Chapt. 7. However, whether the aSTP could be classified as auditory or association/multisensory cortex had been ambiguous based on its neuroanatomy (Galaburda and Pandya 1983; Kaas and Hackett 1998), begging the question whether multisensory interactions in the voice-sensitive aSTP are comparable to those in early auditory cortex, or alternatively, whether they are more like those seen in multisensory association areas.

2.1 How Multisensory are Neurons in the Anterior Voice-Identity Sensitive fMRI Cluster?

To directly study whether and how auditory responses to voices are affected by simultaneously presented visual facial information, neuronal recordings were performed from the anterior aSTP voice cluster during auditory, visual, or audiovisual presentation of dynamic face and voice stimuli. As might be expected of a predominantly auditory area, neurons in the voice-sensitive cortex primarily respond to auditory stimuli while silent visual stimuli are mostly ineffective in eliciting neuronal firing (see Fig. 2B and Perrodin et al. 2014). Yet, comparing the amplitudes of spiking responses to unimodal (auditory alone) vs bimodal (audiovisual) stimulation conditions revealed clear nonlinear influences (sub-additive or super-additive) on the responses of auditory neurons (Fig. 2A, B). This provided evidence for robust visual modulation of auditory neuronal responses at the anterior voice-sensitive fMRI cluster in the aSTP (Perrodin et al. 2014). From here, a comparison can be made between these multisensory influences in the anterior voice area and those seen in earlier auditory areas. Interestingly, similar proportions and types of multisensory influences have been reported for neurons in the posterior core/belt auditory areas (Ghazanfar et al. 2005; Kayser et al. 2008), suggesting qualitatively comparable multisensory interactions throughout auditory cortex, from earlier auditory cortical processing stages to the anterior voice cluster.

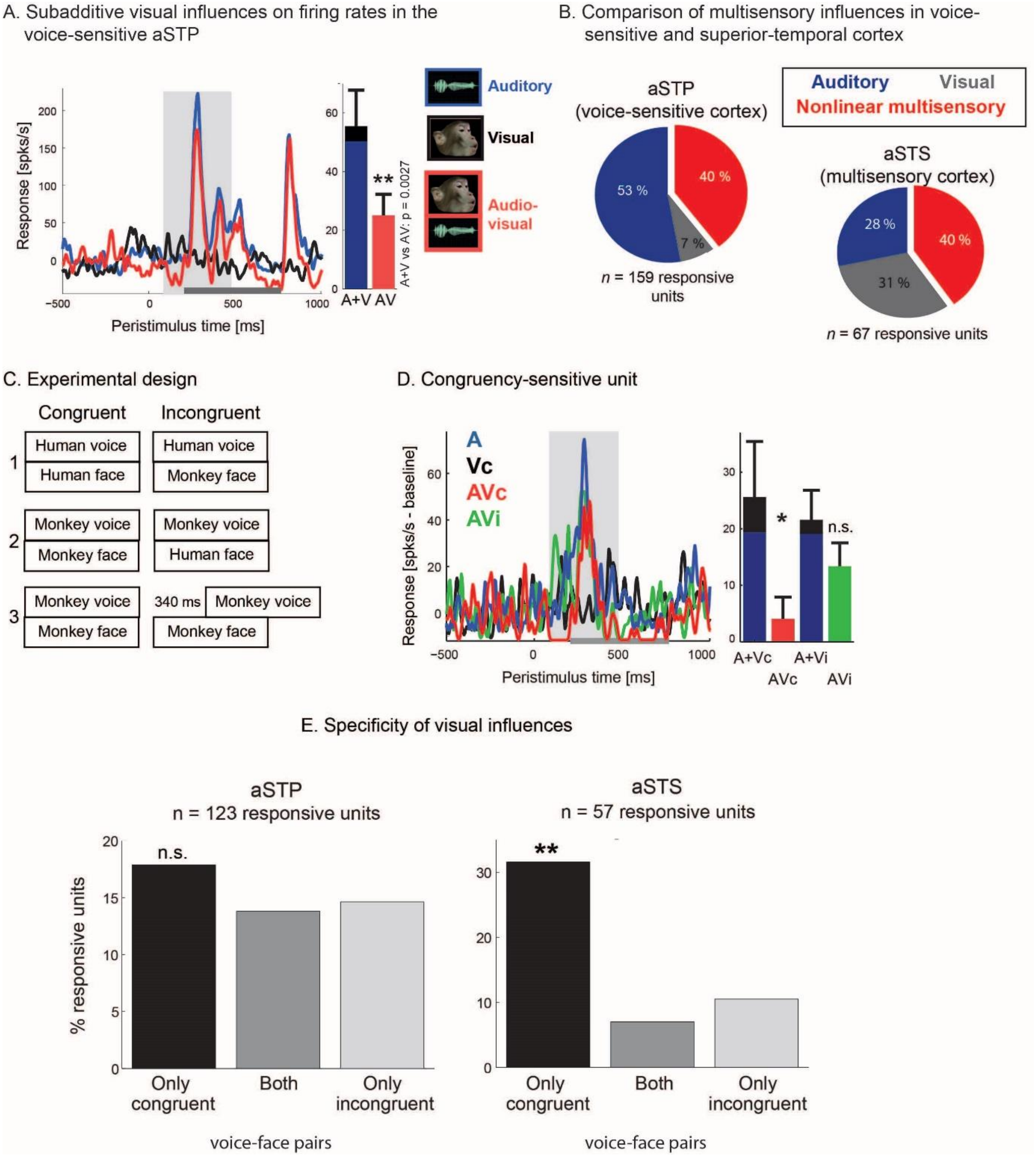

Fig. 2 Neuronal multisensory influences and effect of voice-face congruency in voice-sensitive and superior-temporal cortex

A. Example spiking response of a unit in the anterior voice-sensitive fMRI cluster on the superior-temporal plane showing nonlinear (subadditive) visual modulation of auditory activity: firing rates in response to combined audio-visual stimulation (*AV*, voice and face) are significantly lower than the sum of the responses to the unimodal stimuli (*A*: auditory and *V*: visual; AV vs (A+V), z-test, **: p<0.01). The horizontal gray line indicates the duration of the auditory stimulus, and the light grey box represents the 400ms peak-centered response window. Bar plots indicate the response amplitudes in the 400ms response window (shown is mean ± SEM). **B**. Neuronal multisensory influences are prominent in voice-sensitive cortex (anterior supratemporal plane; *aSTP*) but are qualitatively different from those in the anterior superior-temporal sulcus (*aSTS*). For example, aSTS neurons more often display bimodal responses (Perrodin et al. 2014). Panel B reproduced from (Perrodin et al. 2015b). **C**. Illustration of the stimulus set containing three congruency violations in primate voice/face pairs. **D**. Example response of a unit with congruency-specific visual influences: a congruent, but not an incongruent, visual stimulus significantly modulated the auditory response. The plot shows spiking activity in response to the auditory stimulus alone (*A*), the congruent visual stimulus alone (*Vc*), the congruent audio-visual (*AVc*), and the incongruent audio-visual (*AVi*) pairs. The horizontal gray line indicates the duration of the auditory stimulus, and the light grey box represents the 400ms response window in which the response amplitude was computed. Bar plots indicate the response amplitudes in the 400ms response window (mean ± sem). The symbols refer to significantly nonlinear audio-visual interactions, defined by comparing the audio-visual response with all possible summations of auditory and visual responses (AVc vs (A+Vv) and AVi vs (A+Vi), z-test, * $p<0.05$; n.s., not significant). **E**. Summary of the congruency specificity of visually modulated units in the anterior voice-sensitive cortex (*aSTP*; *left*) and the anterior superior-temporal sulcus (*right*). Bar plots indicate the percentage of auditory responsive units that showed significant non-additive audio-visual interactions in response to the congruent pair only (black bars), the incongruent pair only (light grey bar), or both the congruent and the incongruent stimuli (dark grey bar). Stars indicate p-values (**: $p<0.01$; n.s., not significant) resulting from a Chi-square test comparing the numbers of visually modulated units for each of the three categories to a uniform distribution. Panels *A*, *C-F* reproduced from (Perrodin et al. 2014) with permission from the Society for Neuroscience. Panel B reproduced from (Perrodin et al. 2015b)

Beyond the proportions of modulated neurons, and the types of multisensory interactions, cross-modal influences are also known to differ in their specificity to behaviorally relevant multisensory combinations used for stimulation (Werner and Noppeney 2010). The neuronal sensitivity of visual influences to speaker congruency was investigated using a set of congruent and incongruent audiovisual stimulus conditions, in which a voice was paired with a mismatched face (Fig. 2C). The neuronal responses to these conditions showed that multisensory influences on aSTP units were relatively insensitive to speaker congruency, and were not strongly affected by the mismatching audio-visual stimulus conditions, such as when a monkey voice was paired with a human face (see Fig. 2E and Perrodin et al. 2014). The relative lack of specificity of visual influences in the aSTP is consistent with the notion that the anterior voice cluster shows more general cross-modal influences, belonging to a relatively early stage of audiovisual processing, which includes primary auditory cortex and surrounding auditory areas (Schroeder et al. 2003; Ghazanfar and Schroeder 2006).

The impressions given by these observations is that there are clear visual influences on many auditory neurons in the anterior voice-sensitive cluster in the aSTP. These multisensory influences are qualitatively more like those reported in early auditory cortical fields, potentially differing from those seen in neurons from multisensory association cortex in the aSTS (see Sect 2.3). Thereby, the anterior voice identity sensitive cluster in monkeys is primarily sensitive to auditory features, with the multisensory influences seen in this region being of a more general modulatory form.

2.2 Natural Asynchronies in Audio-Visual Communication Signals and their Impact on Neuronal Excitability

As we have considered, neurons in the anterior voice-sensitive cluster in the aSTP, although predominantly auditory, do show certain types of cross-modal influences from faces on their auditory spiking responses to voices. During audiovisual communication, a caller is perceived to produce a vocalization while the facial expression changes. However, although these multisensory signals are often perceived to emanate in concord, in natural communication signals there is a considerable level of temporal asynchrony between the onset of informative content in one modality relative to the other. For instance, visual orofacial movements can precede the vocalization by tens to hundreds of milliseconds (see Fig. 3A-C and Ghazanfar et al. 2005; Chandrasekaran et al. 2009). While a coherent multisensory percept can be maintained across a wide range of spatial and temporal discrepancies (McGrath and Summerfield 1985; Slutsky and Recanzone 2001), these subtle to moderate temporal misalignments have the potential to drastically impact on neuronal excitability and population dynamics. Yet because neurons in voice sensitive cortex lack the temporal response fidelity of neurons in and around primary auditory cortical or subcortical regions (Creutzfeldt et al. 1980; Bendor and Wang 2007), it was uncertain whether such stimulus asynchronies affect audiovisual influences on neurons in the aSTP voice-sensitive cortex.

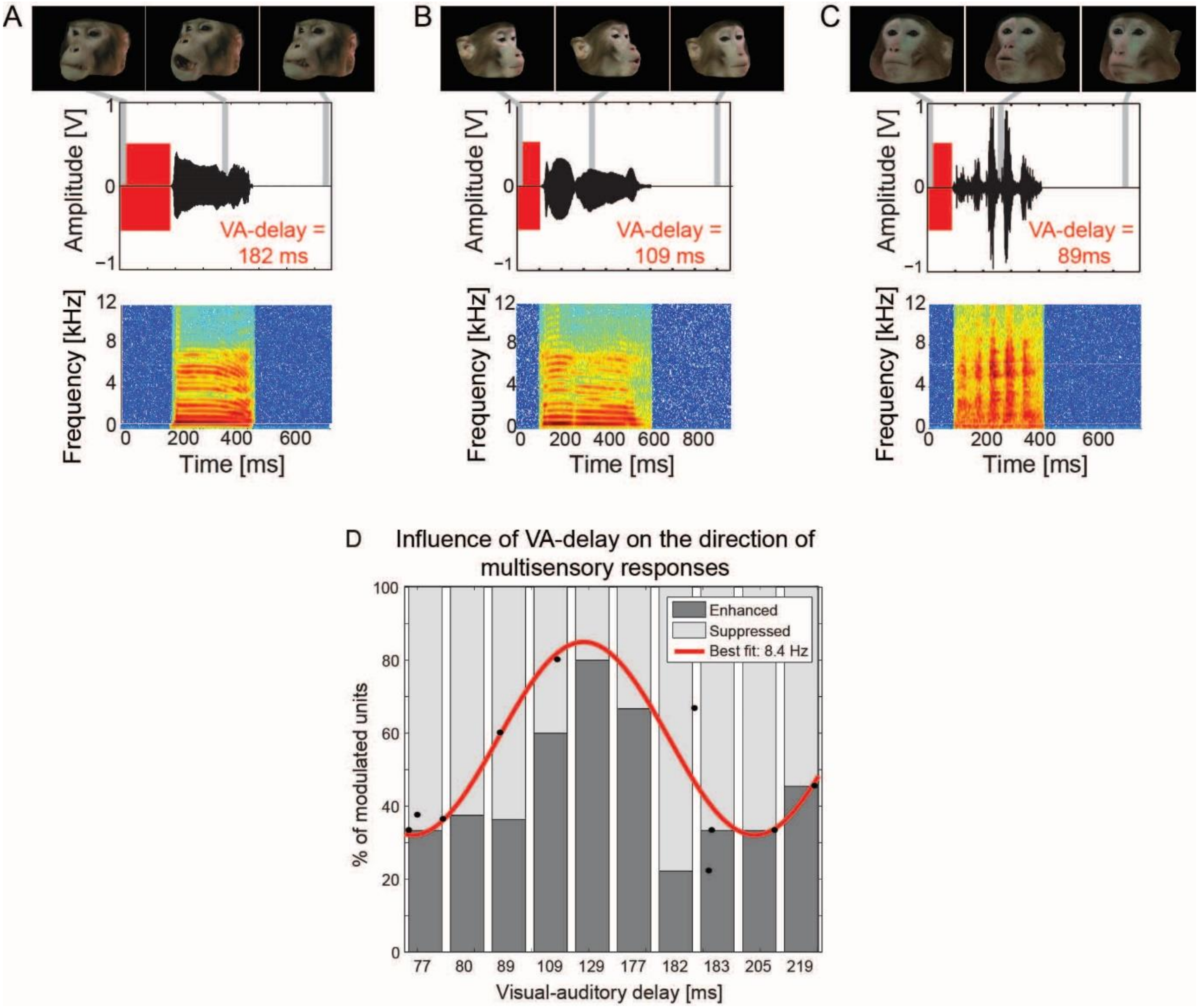

Fig. 3 Audiovisual primate vocalizations, visual-auditory onset delays and the direction (sign) of multisensory interactions

A-C. Examples of audiovisual rhesus macaque ‘coo’ (*A*,*B*) and ‘grunt’ (*C*) vocalizations used for stimulation and their respective natural visual to auditory onset asynchronies/delays (time interval between the onset of mouth movement and the onset of the vocalization; red bars). The video starts at the onset of mouth movement, with the first frame showing a neutral facial expression, followed by mouth movements associated with the vocalization. Gray lines indicate the temporal position of the representative video frames (*top row*). The amplitude waveforms (*middle row*) and the spectrograms (*bottom row*) of the corresponding auditory vocalization are displayed below. **D**. Proportions of enhanced and suppressed multisensory units by stimulus, arranged as a function of increasing visual to auditory onset delays (*VA-delay*, n = 81 units). Note that the bars are spaced at equidistant intervals for display purposes. Black dots indicate the proportion of enhanced units for each VA-delay value, while respecting the real relative positions of VA-delay values. The red line represents the sinusoid with the best-fitting frequency (8.4 Hz, adjusted $R^2 = 0.58$). Reproduced from (Perrodin et al. 2015a).

Relevant studies have shown that the temporal dynamics of sensory streams, such as those typical for processing human speech or other natural stimuli, can shape and synchronize cortical oscillations through entrainment (Ghitza 2011; Giraud and Poeppel 2012). More generally, neuronal oscillations are thought to reflect the state-dependent excitability of local networks to subsequent incoming sensory inputs (Schroeder et al. 2008; Thut et al. 2012). These oscillatory responses are considered to reflect neuronal population mechanisms for routing information to downstream stages and prioritizing the processing at sensory nodes in the brain network (Bastos et al. 2015). It is also known that cortical oscillations are influenced by rhythmic multisensory input (Thorne and Debener 2014; van Atteveldt et al. 2014).

The impact of cross-modal stimulus asynchronies on neuronal responses and cortical oscillations in voice-sensitive cortex of the primate aSTP was assessed using a set of dynamic face and voice combinations spanning a broad range of naturally occurring audio-visual asynchronies (Fig. 3A-C). The results of this study revealed that the prevalence of two key forms of audiovisual interactions in neuronal spiking responses (multisensory enhancement or suppression) varied according to the degree of asynchrony between the onsets of informative communication content in either sensory input stream (see Fig. 3D and Perrodin et al. 2015a). Time-frequency analyses of the local-field potential signal in the aSTP showed that this cross-modal asynchrony selectively affects low-frequency neuronal oscillations (Perrodin et al. 2015a). By aligning and transiently synchronizing the phase of ongoing low-frequency cortical oscillations, the visual input cyclically influences the excitability of auditory neuronal responses in the aSTP. Thus, whether the majority of neurons show enhancement or suppression in their multisensory responses depends to a large extent on the visual-to-auditory stimulus onset delay present in natural communication signals. These effects on neuronal excitability span several hundreds of milliseconds, or the full range of asynchronies observed in audiovisual communication signals (Chandrasekaran et al. 2009; Perrodin et al. 2015a).

The functional role of cortical oscillations and how they modulate sensory perception is the topic of ongoing research. By comparison, comparable cross-modal phase-resetting in local-field potentials is also seen in early auditory and visual cortical areas (Lakatos et al. 2007; Mercier et al. 2013). In primary auditory cortex, another study shifting somatosensory nerve stimulation combined with pure tone stimuli found a comparable alternating pattern of multisensory enhancement and suppression of multi-unit activity for different relative stimulus onset asynchronies (Lakatos et al. 2007). Other studies on visual influences in primary auditory cortex have reported comparable neural response dependancies on temporal stimulus alignment (Ghazanfar et al. 2005; Bizley et al. 2007). Thus in combination with the naturally occuring timing differences in multisensory streams, cross-modal resetting of ongoing oscillations allows the leading visual input to shape or 'window' subsequent auditory responses.

One hypothesis proposes that these temporal relationships in natural communicative situations segment sensory input into an appropriate temporal granularity (Giraud and Poeppel 2012; Gross et al. 2013). For instance, neurons in the monkey STS show specific patterns of slow oscillatory activity and spike timing that reflect visual category-specific information in faces versus other objects (Turesson et al. 2012). Anterior voice area neurons seem to be comparably involved in oscillatory responses, whereby the neuronal spiking responses display different types of multisensory interactions (enhancement vs suppression) depending on the phase alignment of low frequency oscillatory responses. Taken together, these findings suggest an interplay between neuronal firing and the surrounding oscillatory context that needs to be better explored in terms of the causal interactions underlying auditory and audio-visual transformations of neural responses between brain areas. The potential behavioral relevance of these oscillatory phenomena for stimulus identification and detection will also need to be described. These issues are currently being investigated in humans (Henry and Obleser 2012; Keil et al. 2014) for a host of perceptual processes (Strauss et al. 2015; Ten Oever and Sack 2015); also see Keil and Senkowski, Chapt. 10. These efforts could benefit from insights obtained at the neuronal level in animal models, especially from subjects participating in active tasks (Fetsch et al. 2012; Osmanski and Wang 2015) to better understand the perceptual correlates (Chen et al. 2016).

2.3 How do Visual Interactions at Voice Clusters Compare to those in Multisensory Areas of the Temporal Lobe?

Extracellular recordings of neuronal activity in the anterior upper-bank of the STS in response to the same voice and face stimulus set described in Sect. 2.1 revealed a comparable proportion of nonlinear audiovisual interactions as in aSTP neurons (Fig. 2B). However, in agreement with previous electrophysiological studies (Benevento et al. 1977; Dahl et al. 2009), evidence for a greater level of cross-modal convergence was prevalent in the STS, where neurons showed a balance of both auditory and visual responses alongside modulatory multisensory influences (Fig. 2B). These observations are consistent with studies highlighting the STS as a cortical association area, and a prominent target for both auditory and visual afferents in the temporal lobe (Seltzer and Pandya 1994; Beauchamp et al. 2004).

The presentation of incongruent audiovisual stimuli revealed that, in contrast to the generic visual influences in voice-sensitive neurons, those modulating the auditory responses of STS neurons occurred with greater specificity: multisensory interactions were sensitive to the congruency of the presented voice-face pairing, and nonlinear multisensory responses (both super- and sub-additive) occurred more frequently in response to matching compared to mismatching audiovisual stimuli (e.g., were more likely to be disrupted by incongruent stimulation, see Fig. 2D, E). Dahl et al. (2010) similarly reported congruency-sensitive auditory influences on visual responses in the monkey lower-bank STS. These observations are consistent with the evidence for integrative multisensory processes in the human and monkey STS (Beauchamp et al. 2004; Dahl et al. 2009), potentially at the cost of decreased specificity for representing unisensory features (see Sect. 1.3 and Werner and Noppeney 2010; Perrodin et al. 2014). More generally, this increased audiovisual feature-specificity in STS neurons, a classically defined multisensory region, is in agreement with current models of audiovisual processing and the important role of the STS in multisensory integration (Beauchamp et al. 2004; Stein and Stanford 2008).

Thereby, neurons in the anterior voice-sensitive cluster in the aSTP show a double dissociation in functional properties with respect to neurons in the aSTS: aSTP neurons, maybe because they are primarily engaged in sensory analysis in the unisensory (auditory) modality, show little specificity in their cross-sensory influences. In contrast, neurons in the STS show more specific multisensory influences but display a lack of fidelity in their unisensory representations (see also Sect. 1.3). Together, these observations suggest that a high level of specificity is not retained in both the unisensory and the multisensory domain. As such, the results are consistent with the notion of reversed gradients of functional specificity in the unisensory vs multisensory pathways, whereby unisensory stimulus response fidelity decreases along the sensory processing hierarchy as multisensory feature sensitivity and specificity increases. These comparisons of results across different brain areas suggest an intermediate functional role of voice-sensitive neurons in the auditory and audiovisual processing hierarchies relative to early auditory fields and the multisensory STS, which is of relevance for building better neurobiologically informed models of multisensory integration (for reviews see, e.g.: Ghazanfar and Schroeder 2006; Stein and Stanford 2008).

3 Multisensory Pathways to Primate Prefrontal Cortex

Sect. 2.1, 2.2, and 2.3 above reviewed some of the evidence for visual influences on the neuronal processing of voices at voice-sensitive and association regions in the temporal lobe. However these findings do not address whether and how voice regions in the primate temporal lobe are interconnected. Previous neuroanatomical studies have identified pathways for visual and auditory input to the frontal lobe, including dense projections from the second stage of auditory cortical processing, the auditory belt (Romanski et al. 1999b; Plakke and Romanski 2014); also see Plakke and Romanski, Chapt. 7. Projections to frontal cortex from association cortex in the anterior superior-temporal gyrus are considerable (Petrides and Pandya 1988; Seltzer and Pandya 1989). Yet, the strength and functional impact of the connections between the aSTP and frontal cortex was unclear.

Insights into the effective connectivity between some of these regions were recently provided using combined microstimulation and fMRI in monkeys. This approach allows charting the directional connectivity of a specific pathway, in this case between temporal and prefrontal brain regions. Namely, electrically stimulating a specific cortical brain region and using fMRI to assess which regions are activated in response can reveal synaptic targets of the stimulated site, a presumption supported by the fact that target regions activated by stimulation are often consistent with those identified using neuronal anterograde tractography (e.g., Matsui et al. 2011; Petkov et al. 2015). Surprisingly, microstimulating voice-identity sensitive cortex did not strongly activate prefrontal cortex, unlike stimulation of downstream multisensory areas in the STS and upstream auditory cortical areas in the lateral belt (Petkov et al. 2015): The voice-sensitive cortex in the primate aSTP seemed to interact primarily with a local multisensory network in the temporal lobe, including the upper bank of the aSTS and regions around the temporal pole (Fig. 4A). By contrast, stimulating the aSTS resulted in significantly stronger frontal fMRI activation, particularly in orbital frontal cortex (Fig. 4B). These observations complement those on inter-regional connectivity (Frey et al. 2004; Plakke and Romanski 2014) and information on neuronal properties in some of these regions (Kikuchi et al. 2010; Perrodin et al. 2014), which altogether suggest that multisensory voice/face processes are initially integrated in a network of anterior temporal lobe regions, only parts of which have direct access to frontal cortex.

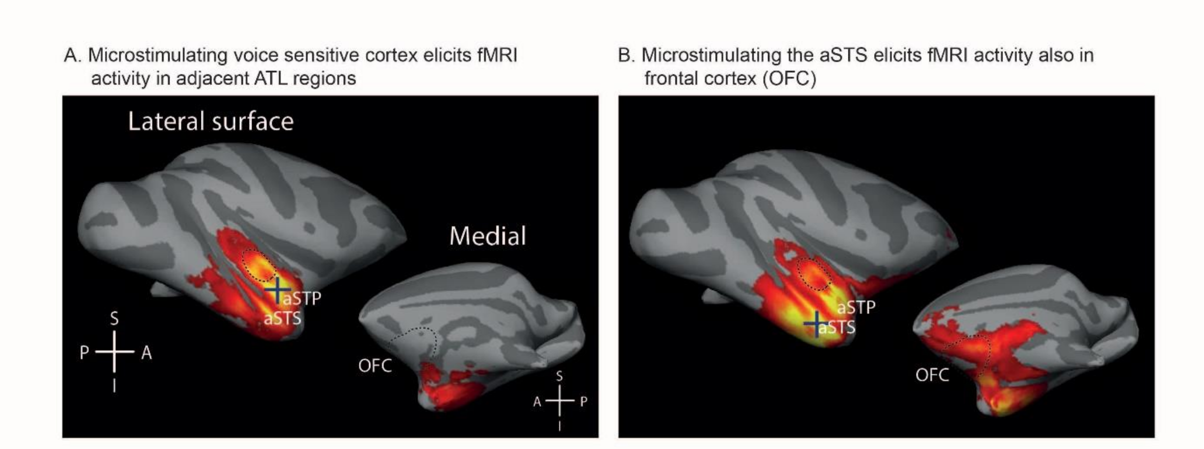
A. Microstimulating voice sensitive cortex elicits fMRI activity in adjacent ATL regions
Lateral surface
Medial
aSTP
aSTS
OFC
S
P
A
I
S
A
P
I
B. Microstimulating the aSTS elicits fMRI activity also in frontal cortex (OFC)
aSTP
aSTS
OFC

Fig. 4 Effective functional connectivity between voice-sensitive and frontal cortices
A. A study of effective functional connectivity using combined microstimulation and fMRI shows that stimulating voice-sensitive cortex (*blue cross*) on the anterior supratemporal plane (*aSTP*) tends to elicit fMRI activity in neighboring regions of the anterior temporal lobe (Petkov et al., 2015). **B**. By contrast, stimulating the anterior superior-temporal sulcus (*aSTS*) also elicits fMRI activity in frontal cortex, in particular the orbitofrontal cortex (*OFC*). *A: anterior, P: posterior, S: superior, I: inferior*. Reproduced from (Perrodin et al. 2015b).

4. Voice and Face Processing Pathways: Comparative Perspective

Much of this review has thus far focused on studies in human and nonhuman primates. However, it is important to at least briefly consider the benefits of pursuing a broader evolutionary perspective for advancing the understanding of the neurobiology of communication signal processing and integration (Fig. 5). Some animal models will allow teasing apart mechanisms and processes that remain difficult or not possible to achieve in primate models.

Although a number of non-primate species rely less on voices and faces for social interactions than other forms of communication, relevant ethologically suitable paradigms can be found. The ferret *(Mustela putorius*) is an animal in which both the auditory and multisensory cortical representations and pathways are actively being studied. For instance, studies in ferrets are relied on to reveal the neuronal coding principles supporting the representation of multiple simultaneous auditory features, such as pitch and the timbre of resonant sources (Bizley et al. 2009; Walker et al. 2011). These auditory features, while more generally found in many natural sounds, nevertheless are related to the processing of voice content, given that prominent indexical cues for identifying an individual by voice are provided by formants, with the vocal folds as the source and vocal tract as the filter (Fitch 2000; Smith and Patterson 2005). Multisensory interactions between auditory and visual stimuli have also been well studied, both anatomically and functionally, in ferrets (Bizley et al. 2007).

Many rodents, including mice (*Mus musculus*), rats (*Rattus norvegicus*), and gerbils (*Meriones unguiculatus*), strongly rely on olfactory/pheromonal and auditory information for social interactions, and these animals can readily identify each other by odor (Brennan 2004). Information about odor identity is represented in the olfactory piriform cortex (Kadohisa and Wilson 2006; Gire et al. 2013), and can synergistically interact with vocalization sounds to influence maternal behavior in mice (Okabe et al. 2013). There appear to be multisensory interactions between the rodent olfactory and auditory processing systems associated with improved maternal behavior (Budinger et al. 2006; Cohen et al. 2011). A broader comparative approach will clarify evolutionary relationships and better define the function of behaviorally relevant uni- and multi-sensory pathways.

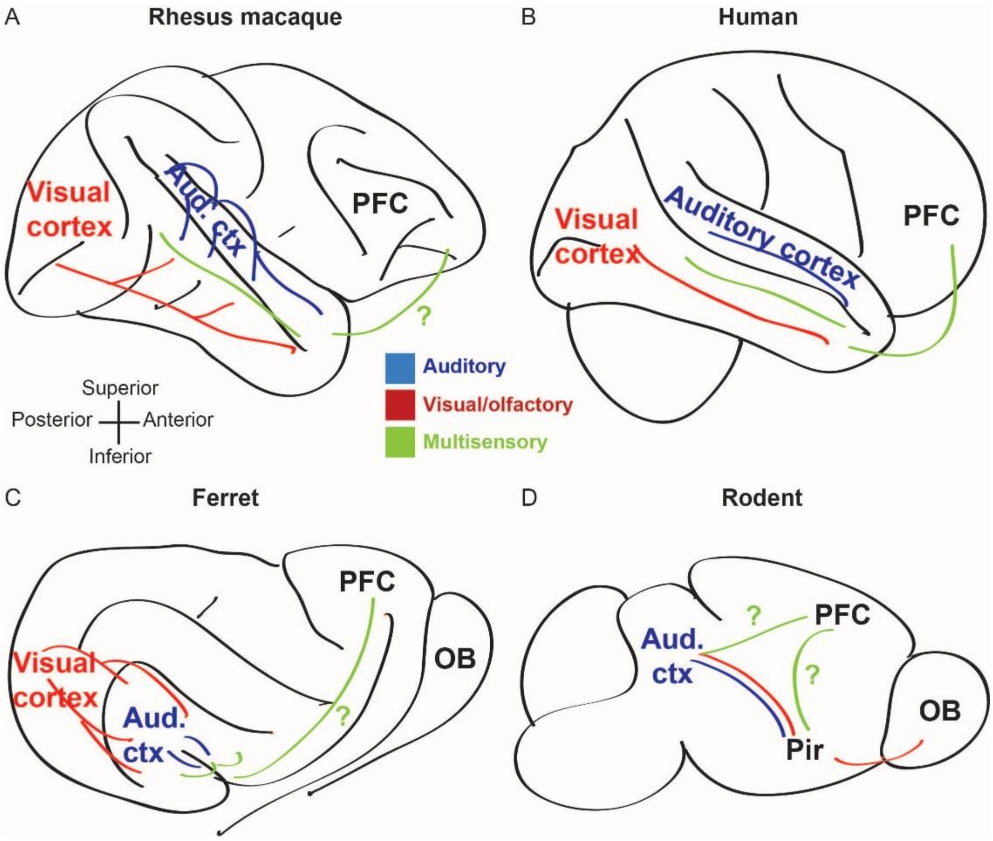

Fig. 5 Sensory processing pathways supporting social communication in the mammalian brain A comparative view of ascending auditory and visual or olfactory cortical streams illustrated on the right hemisphere of brains across several mammalian species. **A**. Rhesus macaque monkey, **B**. Human, **C**. Ferret, **D**. Rodent. Unisensory neuronal representations of communication signals (visual: face or facial expressions in primates, auditory: voices or vocalizations, olfactory: odor or pheromonal cues) become progressively more selective in relation to primary sensory cortices (auditory projections: blue lines, visual and/or olfactory: red. *OB*: olfactory bulb, *Pir*: piriform (olfactory) cortex). Bidirectional anatomical and functional crosstalk occurs at multiple levels throughout the sensory streams, for instance visual projections into auditory cortices (Bizley et al. 2007), or auditory projections from primary auditory cortex into olfactory cortex (Budinger et al. 2006). There are also feedforward and feedback projections between cortical areas, including multisensory influences from high-order association areas in the frontal lobes (PFC: prefrontal cortex) onto sensory processing streams (Hackett et al. 1998; Romanski et al. 1999a). Directions for future study include better understanding the nature and dynamics of the bidirectional functional links between higher-level unisensory and frontal cortices, and how these mediate the transformation/abstractions of multisensory neuronal representations.

5. Summary, Conclusions and Look Ahead

This chapter has reviewed the current state of neuroscientific knowledge on the neural representation of voice and face content in communication signals, focusing in particular on some of the processing sites in the anterior temporal lobe in primates. Guided by neuroimaging results in humans and rhesus macaques and the resulting functional correspondences across the species, invasive electrophysiological recordings in the nonhuman primates revealed evidence for voice cells and characterized their basic functional properties, including how these relate to information on face cell characteristics in the visual system. Neuronal processing in the aSTP voice cluster was found to be sensitive to voice identity, and very acoustically stimulus selective in relation to upstream auditory areas. Additionally, a double dissociation in the auditory feature sensitivity versus the specificity of multisensory interactions was identified between, on the one hand, neurons in the anterior voice-sensitive cluster on the supratemporal plane and, on the other, adjacent regions in temporal association cortex. Insights into the directed functional connectivity have also been obtained, providing information on inter-regional connectivity to complement that on neuronal response characteristics. Together, these initial forays into the neurobiological substrates of voice processing in the temporal lobe raise a number of new questions: What are the perceptual and behavioral correlates of the observed neuronal, and oscillatory, responses and multisensory interactions? What are the transformations and causal interactions that occur between brain regions involved in voice and face processing, as well as multisensory integration for identifying individuals and other entities? Pursuing answers to these questions will be essential for solidifying the next set of advances in understanding how the brain integrates sensory information to guide behavior.

Acknowledgments

We thank Christoph Kayser and Nikos Logothetis for useful discussion, support, and encouragement. This work was supported by the Max Planck Society, Wellcome Trust (CIP, Investigator Award WT092606AIA; CP, Sir Henry Wellcome Fellowship 110238/Z/15/Z), the European Research Council (CIP, MECHIDENT), and the Biotechnology and Biological Sciences Research Council (BB/J009849/1 to CIP joint with Q. Vuong).

Compliance with Ethics Requirements

Catherine Perrodin declares that she has no conflict of interest.
Christopher I. Petkov declares that he has no conflict of interest.

References

Andics, A., Gácsi, M., Faragó, T., Kis, A., & Miklósi, Á. (2014). Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology, 24*(5), 574-578.

Aparicio, P. L., Issa, E. B., & DiCarlo, J. J. (2016). Neurophysiological organization of the middle face patch in macaque inferior temporal cortex. *Journal of Neuroscience, 36*(50), 12729-12745.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron, 85*(2), 390-401.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7*(11), 1190-1192.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105-2109.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129-135.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309-312.

Bendor, D., & Wang, X. (2007). Differential neural coding of acoustic flutter within primate auditory cortex. *Nature Neuroscience, 10*(6), 763-771.

Benevento, L. A., Fallon, J., Davis, B. J., & Rezak, M. (1977). Auditory--visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology, 57*(3), 849-872.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex, 19*(12), 2767-2796.

Bizley, J. K., Jones, G. P., & Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? *Current Opinion in Neurobiology, 40*, 31-37.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex, 17*(9), 2172-2189.

Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J., & Schnupp, J. W. H. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *The Journal of Neuroscience, 29*(7), 2064-2075.

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice-and face-recognition areas. *The Journal of Neuroscience, 31*(36), 12906-12915.

Brennan, P. A. (2004). The nose knows who's who: chemosensory individuality and mate recognition in mice. *Hormones and Behavior, 46*(3), 231-240.

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46*(2), 369-384.

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

Budinger, E., Heil, P., Hess, A., & Scheich, H. (2006). Multisensory processing via early cortical stages: Connections of the primary auditory cortical field with other sensory systems. *Neuroscience, 143*(4), 1065-1083.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535-543.

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience, 22*(11), 2886-2902.

Chan, A. M., Baker, J. M., Eskandar, E., Schomer, D., Ulbert, I., Marinkovic, K., Cash, S. S., & Halgren, E. (2011). First-pass selectivity for semantic categories in human anteroventral temporal lobe. *The Journal of Neuroscience, 31*(49), 18119-18129.

Chandrasekaran, C., & Ghazanfar, A. A. (2009). Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *Journal of Neurophysiology, 101*(2), 773-788.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436.

Chen, A., Gu, Y., Liu, S., DeAngelis, G. C., & Angelaki, D. E. (2016). Evidence for a Causal Contribution of Macaque Vestibular, But Not Intraparietal, Cortex to Heading Perception. *The Journal of Neuroscience, 36*(13), 3789-3798.

Cohen, L., Rothschild, G., & Mizrahi, A. (2011). Multisensory integration of natural odors and sounds in the auditory cortex. *Neuron, 72*(2), 357-369.

Creutzfeldt, O., Hellweg, F. C., & Schreiner, C. (1980). Thalamocortical transformation of responses to complex auditory stimuli. *Experimental Brain Research, 39*(1), 87-104.

Dahl, C. D., Logothetis, N. K., & Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *The Journal of Neuroscience, 29*(38), 11924-11932.

Dahl, C. D., Logothetis, N. K., & Kayser, C. (2010). Modulation of visual responses in the superior temporal sulcus by audio-visual congruency. *Frontiers in Integrative Neuroscience, 4*, 10.

Damasio, A. R. (1989). The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. *Neural Computation, 1*(1), 123-132.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage, 23*(3), 840-848.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2005). Sensitivity to voice in human prefrontal cortex. *Journal of Neurophysiology, 94*(3), 2251-2254.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15*(1), 146-154.

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences, 4*(7), 258-267.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*(5903), 970-973.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science, 330*(6005), 845-851.

Frey, S., Kostopoulos, P., & Petrides, M. (2004). Orbitofrontal contribution to auditory encoding. *Neuroimage, 22*(3), 1384-1389.

Galaburda, A. M., & Pandya, D. N. (1983). The intrinsic architectonic and connectional organization of the superior temporal region of the rhesus monkey. *Journal of Comparative Neurology, 221*(2), 169-184.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10*(6), 278-285.

Ghazanfar, A. A., & Takahashi, D. Y. (2014). The evolution of speech: vision, rhythm, cooperation. *Trends in Cognitive Sciences, 18*(10), 543-553.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience, 25*(20), 5004-5012.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology, 2*, 130.

Gifford, G. W., 3rd, MacLean, K. A., Hauser, M. D., & Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *Journal of Cognitive Neuroscience, 17*(9), 1471-1482.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511-517.

Gire, D. H., Whitesell, J. D., Doucette, W., & Restrepo, D. (2013). Information for decision-making and stimulus identification is multiplexed in sensory cortex. *Nature Neuroscience, 16*(8), 991-993.

Griffiths, T. D., Warren, J. D., Scott, S. K., Nelken, I., & King, A. J. (2004). Cortical processing of complex sound: a way forward? *Trends in Neurosciences, 27*(4), 181-185.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology, 11*(12), e1001752.

Hackett, T. A., Stepniewska, I., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology, 394*(4), 475-495.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research, 32*(3), 203-218.

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences of the United States of America, 109*(49), 20095-20100.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Kaas, J. H., & Hackett, T. A. (1998). Subdivisions of auditory cortex and levels of processing in primates. *Audiology and Neuro-Otology, 3*(2-3), 73-85.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America, 97*(22), 11793-11799.

Kadohisa, M., & Wilson, D. A. (2006). Separate encoding of identity and similarity of complex familiar odors in piriform cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(41), 15206-15211.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302-4311.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18*(7), 1560-1574.

Keil, J., Muller, N., Hartmann, T., & Weisz, N. (2014). Prestimulus beta power and phase synchrony influence the sound-induced flash illusion. *Cerebral Cortex, 24*(5), 1278-1288.

Kikuchi, Y., Horwitz, B., & Mishkin, M. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *The Journal of Neuroscience, 30*(39), 13021-13030.

King, A. J., & Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? *Nature Neuroscience, 12*(6), 698-701.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 104*(51), 20600-20605.

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53*(2), 279-292.

Logothetis, N. K., Guggenberger, H., Peled, S., & Pauls, J. (1999). Functional imaging of the monkey brain. *Nature Neuroscience, 2*(6), 555-562.

Matsui, T., Tamura, K., Koyano, K. W., Takeuchi, D., Adachi, Y., Osada, T., & Miyashita, Y. (2011). Direct comparison of spontaneous functional connectivity and effective connectivity measured by intracortical microstimulation: an fMRI study in macaque monkeys. *Cerebral Cortex, 21*(10), 2348-2356.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77*(2), 678-685.

McLaren, D. G., Kosmatka, K. J., Oakes, T. R., Kroenke, C. D., Kohama, S. G., Matochik, J. A., Ingram, D. K., & Johnson, S. C. (2009). A population-average MRI-based atlas collection of the rhesus macaque. *Neuroimage, 45*(1), 52-59.

Mercier, M. R., Foxe, J. J., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., & Molholm, S. (2013). Auditory-driven phase reset in visual cortex: human electrocorticography reveals mechanisms of early multisensory integration. *Neuroimage, 79*, 19-29.

Miller, C. T., & Cohen, Y. E. (2010). Vocalizations as Auditory Objects: Behavior and Neurophysiology. In A. Ghazanfar & M.L. Platt (Eds.), *Primate Neuroethology* (pp. 237–255). Oxford University Press.

Mizrahi, A., Shalev, A., & Nelken, I. (2014). Single neuron and population coding of natural sounds in auditory cortex. *Current Opinion in Neurobiology, 24*, 103-110.

Morin, E. L., Hadj-Bouziane, F., Stokes, M., Ungerleider, L. G., & Bell, A. H. (2014). Hierarchical encoding of social cues in primate inferior temporal cortex. *Cerebral Cortex, 25*(9), 3036-3045.

Okabe, S., Nagasawa, M., Kihara, T., Kato, M., Harada, T., Koshida, N., Mogi, K., & Kikusui, T. (2013). Pup odor and ultrasonic vocalizations synergistically stimulate maternal attention in mice. *Behavioural Neuroscience, 127*(3), 432-438.

Osmanski, M. S., & Wang, X. (2015). Behavioral Dependence of Auditory Cortical Responses. *Brain topography, 28*(3), 365-378.

Pandya, D. N., Hallett, M., & Kmukherjee, S. K. (1969). Intra- and interhemispheric connections of the neocortical auditory system in the rhesus monkey. *Brain Research, 14*(1), 49-65.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*(3), 329-342.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2011). Voice cells in the primate temporal lobe. *Current Biology, 21*(16), 1408-1415.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2014). Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *The Journal of Neuroscience, 34*(7), 2524-2537.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2015a). Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. *Proceedings of the National Academy of Sciences of the United States of America, 112*(1), 273-278.

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K., & Petkov, C. I. (2015b). Who is That? Brain Networks and Mechanisms for Identifying Individuals. *Trends in Cognitive Sciences, 19*(12), 783-796.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice region in the monkey brain. *Nature Neuroscience, 11*(3), 367-374.

Petkov, C. I., Kikuchi, Y., Milne, A. E., Mishkin, M., Rauschecker, J. P., & Logothetis, N. K. (2015). Different forms of effective connectivity in primate frontotemporal pathways. *Nature Communications, 6*, 10.1038/ncomms7000.

Petrides, M., & Pandya, D. N. (1988). Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *Journal of Comparative Neurology, 273*(1), 52-66.

Plakke, B., & Romanski, L. M. (2014). Auditory connections and functions of prefrontal cortex. *Front in Neuroscience, 8*, 199.

Plakke, B., Diltz, M. D., & Romanski, L. M. (2013). Coding of vocalizations by single neurons in ventrolateral prefrontal cortex. *Hearing Research, 305*, 135-143.

Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiology and Neurotology, 3*(2-3), 86-103.

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America, 97*, 11800-11806.

Recanzone, G. H. (2008). Representation of con-specific vocalizations in the core and belt areas of the auditory cortex in the alert macaque monkey. *The Journal of Neuroscience, 28*(49), 13184-13193.

Remedios, R., Logothetis, N. K., & Kayser, C. (2009). An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. *The Journal of Neuroscience, 29*(4), 1034-1045.

Romanski, L. M. (2007). Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cerebral Cortex, 17 Suppl 1*, i61-69.

Romanski, L. M. (2012). Integration of faces and vocalizations in ventral prefrontal cortex: implications for the evolution of audiovisual speech. *Proceedings of the National Academy of Sciences of the United States of America, 109 Suppl 1*, 10717-10724.

Romanski, L. M., Bates, J. F., & Goldman-Rakic, P. S. (1999a). Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology, 403*(2), 141-157.

Romanski, L. M., Averbeck, B. B., & Diltz, M. (2005). Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *Journal of Neurophysioly, 93*(2), 734-747.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999b). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience, 2*(12), 1131-1136.

Russ, B. E., Ackelson, A. L., Baker, A. E., & Cohen, Y. E. (2008). Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *Journal of Neurophysiology, 99*(1), 87-95.

Sadagopan, S., Temiz-Karayol, N. Z., & Voss, H. U. (2015). High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Scientific Reports, 5*.

Saleem, K. S., & Logothetis, N. K. (2007). *A Combined MRI and Histology: Atlas of the Rhesus Monkey Brain in Stereotaxic Coordinates*. London: Academic Press.

Schroeder, C. E., & Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research, 14*(1), 187-198.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*(3), 106-113.

Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., & Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology, 50*(1-2), 5-17.

Seltzer, B., & Pandya, D. N. (1989). Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. *Journal of Comparative Neurology, 281*(1), 97-113.

Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *Journal of Comparative Neurology, 343*(3), 445-463.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain, 115 Pt 1*, 15-36.

Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology, 28*, 5-9.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport, 12*(1), 7-10.

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America, 118*(5), 3177-3186.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neurosci, 9*(4), 255-266.

Strauss, A., Henry, M. J., Scharinger, M., & Obleser, J. (2015). Alpha phase determines successful lexical decision in noise. *The Journal of Neuroscience, 35*(7), 3256-3262.

Sugihara, T., Diltz, M. D., Averbeck, B. B., & Romanski, L. M. (2006). Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *The Journal of Neuroscience, 26*(43), 11138-11147.

Ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences of the United States of America, 112*(52), 15833-15837.

Thorne, J. D., & Debener, S. (2014). Look now and hear what's coming: on the functional role of cross-modal phase reset. *Hearing Research, 307*, 144-152.

Thut, G., Miniussi, C., & Gross, J. (2012). The functional importance of rhythmic activity in the brain. *Current Biology, 22*(16), R658-663.

Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science, 292*(5515), 290-293.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*, 411-437.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science, 311*(5761), 670-674.

Tsao, D. Y., Schweers, N., Moeller, S., & Freiwald, W. A. (2008). Patches of face-selective cortex in the macaque frontal lobe. *Nature Neuroscience, 11*(8), 877-879.

Turesson, H. K., Logothetis, N. K., & Hoffman, K. L. (2012). Category-selective phase coding in the superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America, 109*(47), 19438-19443.

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory Integration: Flexible Use of General Operations. *Neuron, 81*(6), 1240-1253.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48-55.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*(3), 367-376.

Walker, K. M., Bizley, J. K., King, A. J., & Schnupp, J. W. (2011). Multiplexed and robust representations of sound features in auditory cortex. *The Journal of Neuroscience, 31*(41), 14565-14576.

Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex, 50,* 125-136.

Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *The Journal of Neuroscience, 30*(7), 2662-2675.

Yau, J. M., DeAngelis, G. C., & Angelaki, D. E. (2015). Dissecting neural circuits for multisensory integration and crossmodal processing. *Philosophical Transactions of the Royal Society B, 370*(1677), 20140203.