Face-to-face communication is a temporally extended, multisensory process. In human speech, the auditory component is, for the most part, the result of vocal fold movements that release sound energy. This sound energy travels up through the oral and nasal cavities, where it is radiated out of the lips and nostrils. Changes in the internal and external shape of these cavities lead to different speech sounds as well as deformations of the external face around the oral aperture and other regions (Chandrasekaran, Trubanova, Stillitano, Caplier, & Ghazanfar, 2009; Yehia, Rubin, & Vatikiotis-Bateson, 1998). This orofacial motion acts as a cue to what is being heard and is referred to as visual speech. Visual speech is not just an epiphenomenon of the vocal production process but is a critical component of face-to-face communication. Visual speech provides considerable intelligibility benefits to the perception of auditory speech (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954), is difficult to ignore (integrating readily and automatically with auditory speech; McGurk & MacDonald, 1976), and interacts with auditory speech even at the earliest stages of human cognitive development (Lewkowicz & Ghazanfar, 2009). Thus, multisensory speech seems to be the primary mode of speech perception and is not a capacity that is simply piggybacked onto auditory speech perception (Rosenblum, 2005).

If the processing of multisensory signals forms the default mode of speech perception, then this should be reflected both in the evolution of vocal communication and the organization of neural processes related to communication. Naturally, any vertebrate organism (from fishes and frogs to birds and dogs) that can produce a vocalization will have a concomitant visual motion in the area of the mouth. In the primate lineage, the number of different vocalizations, and thus different patterns of facial motion, has increased during the course of evolution relative to other species. In light of this, it is not surprising that when nonhuman primates produce vocalizations, these utterances are accompanied by variety of visual cues encompassing a range of mouth, ear and head movements (Hauser, Evans, & Marler, 1993; Partan, 2002), as is the case for human speech production. Similarities between human and nonhuman primate vocal production imply that the perceptual and neural mechanisms underlying multisensory vocal perception and its evolution all could be illuminated by studying monkeys and other primates.

The purpose of this review is (1) to briefly describe the data that reveal that human speech is not uniquely multisensory, that in fact the default mode of communication is multisensory in nonhuman primates (hereafter, primates) as well, and (2) to suggest that this mode of communication is reflected in the organization of the neocortex. We summarize the underlying structure of communication signals in primates, highlighting the similarities and differences with human speech. We then show that nonhuman primates display very similar strategies and behavioral patterns as humans in their perception of vocal communication signals. Finally, we review what we know about the neurophysiological bases of multisensory vocal communication. To conclude, we explore a mechanistic account to explain the neurophysiological integration of visual and auditory communication channels and how such integration could guide behavior.

THE STRUCTURE OF PRIMATE COMMUNICATION SIGNALS

Humans and other primates share a remarkable number of similarities in their vocal signals ranging from production mechanisms and signal structure to the inextricable link between vision and audition. All primate vocal signals are produced through the coordinated movements of the lungs, larynx (vocal folds), and the supralaryngeal vocal tract (Fitch & Hauser, 1995; Ghazanfar & Rendall, 2008). In human speech, the signal—across all languages and contexts—is amplitude modulated, consisting of a rhythm that ranges from 2 to 7 Hz (Chandrasekaran et al., 2009; Drullman, 1995; Greenberg, Carvey, Hitchcock, & Chang, 2003), roughly matching the time scale for syllable production. Such temporal modulation in similar frequency ranges also
seems to be a common feature of several nonhuman primate vocalizations. For example, marmoset twitter calls are modulated in the range 5–9 Hz (Wang, Merzenich, Beitel, & Schreiner, 1995), squirrel monkey vocalizations in the range 6–10 Hz (Godfrey, Atencio, Bonham, Schreiner, & Cheung, 2005), and finally, cotton-top tamarins, macaques, and chimpanzees all seem to have vocalizations modulated in the range 3–10 Hz (Turesson, Chandrasekaran, & Ghazanfar, unpublished data).

In addition to commonalities in temporal structure, vocal production in primates results in the deformation of the face around the oral aperture and other parts of the face (Chandrasekaran et al., 2009; Hauser et al., 1993; Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Yehia et al., 1998; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). In humans, this suggests that visual cues in speech should be highly correlated with the spectral structure of the auditory component and modulated in a similar frequency range, which is precisely what is observed (Chandrasekaran et al., 2009). Nonhuman primate vocalizations also seem to share a similar link between acoustic output and facial dynamics. Different macaque monkey vocalizations are produced with unique lip configurations and mandibular positions, and the motion of such articulators influences the acoustics of the signal (Hauser et al., 1993; Hauser & Ybarra, 1994). Coo calls, like /u/ in speech, are produced with the lips protruded, whereas screams, like the /i/ in speech, are produced with the lips retracted (figure 22.1A).

In speech, the mouth opens and closes with a rhythm that matches the temporal structure of the acoustic output (Chandrasekaran et al., 2009). Oddly, however, for macaque vocalizations, the correlation of mouth opening to the temporal structure of the vocalizations is not apparent. In other words, although macaque vocalizations have an acoustic component that is modulated at a frequency that is between 2 and 7 Hz, their facial movements during vocal production do not have such a rhythm. Macaques produce a single ballistic mouth movement when producing vocalizations—they open their mouth during an utterance and then close it. Thus, the vocal tract origin of the temporal modulations embedded in the acoustics is not known. This suggests that although the acoustic rhythm in humans, apes, and monkeys is likely produced by the same mechanism, in humans the visual component was added at a later time. One idea is perhaps that the rhythmic visual component of audiovisual speech evolved in humans by the modification and co-opting of cyclical jaw movements in an ancestor (MacNeilage, 1998). For example, whereas rhythmic jaw movements are relatively rare during vocal production by monkeys, they are extremely common as facial communicative gestures. For example, lipsmacks, an affiliative signal produced during face-to-face contact, and teeth-grinds, a signal produced when a monkey is anxious, both involve cyclical movements of the mouth and are not accompanied by any vocalizations (Redican, 1975); notably, they have a temporal structure that matches audiovisual speech (2–7 Hz) (Ghazanfar, Chandrasekaran, & Morrill, 2010).

In addition to facial deformations, movement of other parts of the head and the body not directly associated with the articulation process also accompany speech. These are potentially informative. In humans, eyebrow raises, nods and other head movements, manual gestures, and blinks usually accompany speech (Munhall & Vatikiotis-Bateson, 2004). Similarly, in macaques, visual cues not part of the articulatory process are also present. For example, bark vocalizations are usually accompanied by retracted ears and a lowered head, whereas pant threats occur with ears forward and a raised head (Partan, 2002).

There is, naturally, a robust correlation between auditory and visual components of vocalizations. The relative timing of the two modalities is such that, in humans and other primates, facial motion typically precedes the onset of the auditory component (figure 22.1B,C). This delay between the onset of mouth movement and onset of sound, termed the time-to-voice (TTV; not to be confused with voice-onset time or VOT), can range anywhere from 100 to 300 msec for human speech (Chandrasekaran & Ghazanfar, 2009). For the limited data available in monkeys, a similar range has been observed with the lower limit around 60 msec and an upper limit of 300 msec (Chandrasekaran & Ghazanfar, 2009; Ghazanfar, Maier, Hoffman, & Logothetis, 2005). Although this feature of speech and primate vocalizations is well known, the physiological and behavioral consequences of onset of facial motion before the onset of the sound are only now beginning to be studied both in humans (Abry, Lallouache, & Cathiard, 1996; Munhall & Tohkura, 1998; van Wassenhove, Grant, & Poeppel, 2007) and in primates (Chandrasekaran & Ghazanfar, 2009; Ghazanfar et al., 2005).

PERCEPTUAL RESPONSES OF PRIMATES TO COMMUNICATION SIGNALS

Given that both humans and other extant primates use both facial and vocal expressions as communication signals, it is perhaps not surprising that the behavior of primates in response to such communication signals mirrors the behavior of humans. First, both humans and primates use temporal cues for the recognition of
Figure 22.1  Structure of communication signals. (A) Exemplars of the facial expressions produced concomitantly with vocalizations. Rhesus monkey coo and scream taken at the midpoint of the expressions with their corresponding spectrograms. (B) Visual and auditory dynamics during the production of the word “PROBLEM” by a single speaker. Upper panel denotes the inter-lip distance as a function of time. The lower panel shows waveform of the sound. Dots denote the full- and half-opening points of the mouth. Note that the mouth dynamics seem to start nearly 275 msec before the onset of the sound. (C) Visual and auditory dynamics during a coo vocalization. Figure conventions as in B. Note again how mouth opening precedes onset of the sound even for nonhuman primate vocalizations.
conspecific (i.e., from the same species) signals. Humans are able to recognize speech with very few spectral frequency channels as long as the temporal modulations of the envelope are intact (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), and the critical frequency band of the temporal envelope for normal speech comprehension is between 4 and 16 Hz (Drullman, Festen, & Ploomp, 1994; Van Der Horst, Leeuw, & Wouter, 1999). Furthermore, chimeric speech sounds, in which the amplitude (temporal) envelope of one sentence is combined with the fine structure of another sentence, reveal that human listeners primarily use the envelope in speech perception (Smith, Delgutte, & Oxenham, 2002).

Few studies have investigated the role of different acoustic features in primate vocal recognition, but the available data strongly suggest that the temporal structure of vocalizations is critical. For example, at the population level, wild rhesus monkeys tend to orient their heads to the right when they hear conspecific sounds (Hauser & Andersson, 1994; Teufel, Ghazanfar, & Fischer, 2010; see Teufel et al., 2010, for a critique of this approach). When the amplitude envelopes of grunts and shrill barks are stretched beyond the species-typical range, they no longer orient to the right for these calls (Hauser, Aghnetta, & Perez, 1998). Similarly, when tested with time-reversed versions of harmonic arches or shrill barks—a manipulation that changes the direction of the envelope but preserves the overall spectral content—rhesus monkeys again orient toward the left as though they are not hearing the calls as conspecific signals (Ghazanfar, Smith-Rohrberg, & Hauser, 2001). These studies suggest that the amplitude envelope is critical for call recognition in rhesus monkeys. Along the same lines (but using a different methodology), tamarin monkeys will respond vocally to conspecific “long calls” that have the species-typical amplitude envelope, but with the spectral structure replaced with white noise (Ghazanfar, Smith-Rohrberg, Pollen, & Hauser, 2002). They, however, respond significantly less to calls without the species-typical envelope.

The link between facial motion and vocalizations presents an obvious opportunity to exploit the concordance of both channels. Thus, it is not surprising that many primates other than humans recognize the correspondence between the visual and auditory components of vocal signals. Rhesus and Japanese macaques (Macaca mulatta and Macaca fuscata), capuchins (Cebus apella), and chimpanzees (Pan troglodytes) (the only non-human primates tested thus far) all recognize auditory-visual correspondences between their various vocalizations (Adachi, Kuwahata, Fujita, Tomonaga, & Matsuzawa, 2006; Evans, Howell, & Westergaard, 2005; Ghazanfar & Logothetis, 2003; Izumi & Kojima, 2004; Parr, 2004). For example, rhesus monkeys readily match the facial expressions of “coo” and “threat” calls with their associated vocal components (Ghazanfar & Logothetis, 2003). Perhaps more pertinent, rhesus monkeys can also segregate competing voices in a chorus of coos, much as humans might with speech in a cocktail party scenario, and match them to the correct number of individuals seen cooing on a video screen (Jordan, Brannon, Logothetis, & Ghazanfar, 2005). Finally, macaque monkeys use formants (i.e., vocal tract resonances) as acoustic cues to assess age-related body size differences among conspecifics (Ghazanfar et al., 2007). They do so by linking across modalities the body size information embedded in the formant spacing of vocalizations (Fitch, 1997) with the visual size of animals who are likely to produce such vocalizations (Ghazanfar et al., 2007). Taken together these data suggest that humans are not at all unique in their ability to receive communication information across modalities.

Although both humans and other primates readily link facial expressions with appropriate, congruent vocal expressions, the cues and behavioral strategies they use to make such matches are not known. One method for investigating behavioral strategies is the measurement of eye movement patterns. For example, when human participants are given no task or instruction regarding what acoustic cues to attend to, they will consistently look at the eye region more than the mouth when viewing videos of human speakers (Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Macaque monkeys exhibit a very similar strategy. The eye movement patterns of monkeys viewing conspecifics producing vocalizations reveal that monkeys spend most of their time inspecting the eye region relative to the mouth (Ghazanfar, Nielsen, & Logothetis, 2006) (figure 22.2A). When they did fixate on the mouth, it was highly correlated with the onset of mouth movements (figure 22.2B). Indeed, such a strategy is probably adaptive, as suggested by evidence from human paradigms. For example, human subjects asked to identify words fixated more on the mouth region at the onset of facial motion (Lansing & McConkie, 2003).

For the work with monkeys described above, the focus has been on spontaneous behavioral responses to communication signals; that is, no training or reward was involved in the behavior. This makes it difficult to compare their behavior with the numerous human studies that used psychophysical measures and parameterized tests of audiovisual speech processing. More studies requiring monkeys to discriminate or detect audiovisual communication signals would significantly

410 ASIF A. GHAZANFAR AND CHANDRAMOULI CHANDRASEKARAN
expands our understanding of the similarities and differences between communication signal processing in humans and nonhuman primates.

THE AUDITORY CORTEX IN AUDIOVISUAL COMMUNICATION

Because nonhuman primates seem to exhibit some of the very same behaviors that humans have when confronted with audiovisual communication signals, they provide a powerful model to investigate the mechanisms of communication. Traditionally, association areas around the principal sulcus in the frontal lobe, the intraparietal sulcus in the parietal lobe, and the superior temporal sulcus of the temporal lobe have been thought to be responsible for linking signals across multiple modalities (Ettlinger & Wilson, 1990). Although these areas are important for the processing of communication signals, they are by no means the only structures that integrate auditory and visual cues to support communication. Indeed, recent efforts suggest that perhaps even classical “unisensory” areas such as auditory cortex integrate cues from faces and voices among other sensory signals. Here we focus on studies investigating how areas such as the core and belt regions of auditory cortex integrate faces and voices and how such integration may be mediated by interactions with the superior temporal sulcus.

Two types of neural signals in auditory cortex have been investigated: spiking activity and local field potentials. Spiking activity represents the output responses (in the form of action potentials) of one or more neurons in a given area. Local field potentials (LFPs), on the other hand, largely reflect input processes that occur on a slower time scale, including synaptic potentials. Local field potentials can be further broken down into distinct rhythms. In fact, the neocortex in general produces stereotypical rhythms or oscillations (Buzsaki & Draguhn, 2004). These neocortical oscillations are regular amplitude modulations, representing fluctuations between low- and high-excitability states. These modulations range in frequency from very slow (e.g., 0.05 Hz) to very fast (e.g., 500 Hz). The different frequencies are thought to represent the size of the underlying neural network that produces them, with higher-frequency oscillations mediated by smaller networks of neurons than lower-frequency oscillations. The set of oscillations that we know the most about are labeled with Greek letters: delta (1–3 Hz), theta (3–8 Hz), alpha (8–14 Hz), beta (14–30 Hz), and gamma (>30 Hz). These different oscillations are associated with different brain states, and, under certain conditions, multiple rhythms can occur at the same time in a particular neocortical area or even across different areas (Chandrasekaran & Ghazanfar, 2009; Ghazanfar, Chandrasekaran, & Logothetis, 2008).

Consistent with evoked potential studies in humans (Besle, Fort, Delpuech, & Giard, 2004; van Wassenhove, Grant, & Poeppel, 2005), recordings from both primary (A1) and lateral belt (middle lateral belt area, ML) auditory cortex in the monkey reveal that responses to the voice are influenced by the presence of a dynamic face (figure 22.3). Monkey subjects viewing unimodal and bimodal versions of two different species-typical vocalizations (“coos” and “grunts”) show both enhanced and suppressed LFP responses in the bimodal condition relative to the unimodal auditory condition (Ghazanfar et al., 2005). These modulations of neural signal strength are considered “integrative” in the sense that the change in response magnitude is significantly significant.
greater than the strongest unimodal response and, in most cases, greater than the sum of the two unimodal responses. In monkeys, the combination of faces and voices led to integrative responses in the vast majority of auditory cortical sites—both in primary auditory cortex and the lateral belt auditory cortex. The data demonstrated that LFP signals in the auditory cortex are capable of multisensory integration of facial and vocal signals in monkeys (Ghazanfar et al., 2005), and such a hypothesis has subsequently been shown also to be true at the level of spiking activity from single neurons in the lateral belt cortex as well (Ghazanfar, Chandrasekaran, & Logothetis, 2008).

Is such modulation specific to a dynamic face, or would any arousing dynamic stimulus do? The specificity of integration was tested by replacing the dynamic faces with dynamic disks that mimicked the aperture and displacement of the mouth. In human psychophysical experiments, such artificial dynamic stimuli can still lead to enhanced speech detection, but not to the same degree as a real face (Bernstein, Auer, & Takayanagi, 2004; Schwartz, Berthommier, & Savariaux, 2004). When LFP or spiking activity was investigated following presentations with dynamic disks, far less integration was seen when compared to when real monkey faces were presented (Ghazanfar et al., 2008; Ghazanfar et al., 2005) (figure 22.4A,B). This was true primarily for the lateral belt auditory cortex and was observed to a lesser extent in the primary auditory cortex.

THE UPPER BANK OF THE SUPERIOR TEMPORAL SULCUS IS A PUTATIVE SOURCE OF VISUAL MODULATION IN AUDITORY CORTEX

Face-specific modulation of auditory cortex raises a question: What is the source of face-specific visual input into auditory cortex? Although there are several visually sensitive regions that project to auditory cortex (Cappe & Barone, 2005; Ghazanfar & Schroeder, 2006), one region ideally suited to modulate activity in auditory
Figure 22.4 Integrating vision and audition in auditory cortex and the upper bank of STS. (A) Single-neuron examples of multisensory integration of Face + Voice stimuli compared with Disk + Voice stimuli in the lateral belt area. The left panel shows an enhanced response when voices are coupled with faces but no similar modulation when coupled with disks. The right panel shows similar effects for a suppressed response. X-axes show time aligned to onset of the face (solid line). Gray lines indicate the onset and offset of the voice signal. Y-axes depict the firing rate of the neuron in spikes per second. Shaded regions denote the SEM. (B) Neurons in the upper bank of the STS respond to dynamic faces. The left panel shows a neuron in the upper bank of the STS that responded to a teeth-grind. The right panel shows another neuron that responded well to a yawn. X-axes show time aligned to onset of the face (solid line) in milliseconds. Y-axes depict the firing rate of the neuron in spikes per second. Shaded regions denote the SEM. (C) Single-neuron examples of multisensory integration of communication signals in the upper bank of the STS. The left panel shows an enhanced response when voices are coupled with faces. The right panel shows similar effects for another coo vocalization. Conventions as in A.
cortex is the upper bank of the superior temporal sulcus (STS) (Baylis, Rolls, & Leonard, 1987; Bruce, Desimone, & Gross, 1981; Hikosaka, Iwai, Saito, & Tanaka, 1988) (figure 22.3). The STS is an excellent candidate for visual input into auditory cortex because neurons in this region are predominantly visual (Bruce et al., 1981), highly sensitive to biological motion such as dynamic faces (Ghazanfar et al., 2010), and able to integrate faces and voices (Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Chandrasekaran & Ghazanfar, 2009) (figure 22.4B,C). Reciprocal anatomical connections are known to be present between parts of the superior temporal sulcus and the belt region of auditory cortex (Barnes & Pandya, 1992; Seltzer & Pandya, 1994).

The functional relationships between STS and the lateral belt region of auditory cortex during audiovisual vocalization processing was tested in the following manner: (1) recording LFPs from the lateral belt region of auditory cortex and STS concurrently; (2) breaking the LFP signal into the different frequency bands reflecting neural oscillations; and (3) measuring the correlations between these oscillations as a function of stimulus condition using a "cross-spectrum" analysis. The focus of the analyses were oscillations in the high-frequency gamma band (>30 Hz), and they revealed the activity between the auditory cortex and the STS to be more strongly correlated during the presentation of faces and voices together relative to the unimodal conditions (Ghazanfar et al., 2008) (figure 22.5A). Because the cross-spectrum analysis conflates coordinated changes in strength with those of timing, a separate analysis, called “phase coherence” was used to examine only changes in timing. This analysis revealed that correlation strength changes between the two structures were also driven by tight temporal coordination of their respective gamma oscillations (figure 22.5B). In sum, faces and voices generate stronger functional interactions between the auditory cortex and the STS.

These data suggest that the influence of vision via the STS does not drive auditory cortical neurons but rather modulates their excitability, leading to enhancement or suppression. Furthermore, the influence of the STS on auditory cortex was not merely restricted to gamma oscillations. Spiking activity also seems to be modulated, but not “driven,” by ongoing activity arising from the STS. That is, the spiking activity in auditory cortex seemed to be influenced by the activity in the STS that preceded it. Two lines of evidence suggest this “modulation” scenario. First, visual influences on single neurons were most robust when in the form of dynamic faces and were apparent only when neurons had a significant response to a vocalization (i.e., there were no overt responses to faces alone). Second, these integrative responses were often “face-specific” and had a wide distribution of latencies, which suggested that the face signal was an ongoing signal that influenced auditory responses (Ghazanfar et al., 2008).

Local field potential signals from both the auditory cortex and the STS have multiple bands of oscillatory activity generated in response to stimuli that might mediate different functions (Chandrasekaran & Ghazanfar, 2009; Lakatos et al., 2005). In the STS, these different bands of oscillatory activity seem to integrate faces and voices differently (Chandrasekaran & Ghazanfar, 2009), and such integration was dependent on the

![Figure 22.5](image_url)  
**Figure 22.5** STS is a source of visual input into auditory cortex. (A) Cross-spectral power between the LFPs in auditory cortex and the upper bank of the STS from 0 to 300 msec. X-axes depict the neural frequency band in Hertz. Y-axes depict the normalized cross-spectral power. (B) Population phase concentration between STS and auditory cortex from 0 to 300 msec after voice onset. X-axes depict frequency in Hertz. Y-axes depict the average normalized phase concentration. Shaded regions denote the SEM across all electrode pairs and calls. All values are normalized by the baseline mean for different frequency bands.
time-to-voice interval (figure 22.6). Because different oscillation frequencies imply different cortical network sizes, these differences in integration may reflect different underlying multisensory computations using networks with different spatial scales (Senkowski, Schneider, Foxe, & Engel, 2008). Below 20 Hz, and in response to naturalistic audiovisual stimuli, there are directed interactions from auditory cortex to STS, whereas above 20 Hz there are directed interactions from STS to auditory cortex (Kayser & Logothetis, 2009). Because mouth movements are in the range of 2–7 Hz (at least in humans; Chandrasekaran et al., 2009), and eye

![Figure 22.6](image-url)

**Figure 22.6** Different neural frequency bands integrate faces and voices differently. (A) Top, baseline corrected alpha-band activity (8–14 Hz) for the three conditions, face + voice, face alone, and voice alone, for a coo call with a 331-msec time-to-voice interval. Bottom panel, baseline-corrected gamma-band (60–95 Hz) activity for the same coo call. X-axes depict time in milliseconds; Y-axes depicts baseline-corrected power. Black line denotes onset of the face. Solid gray lines the onset and offset of the voice. (B) Left, percentage integration of the peak face + voice alpha response relative to the voice-alone response as a function of the time-to-voice interval. X-axes depict time to voice in milliseconds; Y-axes depict integration in percent. Labels co and gt denote coos and grunts, respectively. Right, percentage integration of the peak face + voice gamma response relative to the voice-alone response as a function of the time to voice. Conventions as in left panel.
movements such as saccades and microsaccades are in the range of 3–4 Hz (Otero-Millan, Troncoso, Macknik, Serrano-Pedraza, & Martinez-Conde, 2008; Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010) it is also possible that these lower-frequency interactions between the STS and auditory cortex also represent distinct multisensory processing channels.

In light of our focus on communication signals, two things should be noted. The first is that functional interactions between STS and auditory cortex are not likely to occur solely during the presentation of faces with voices. Other congruent, behaviorally salient audiovisual events such as looming signals (Cappe, Thut, Romei, & Murray, 2009; Gordon & Rosenblum, 2005; Maier, Neuhoff, Logothetis, & Ghazanfar, 2004) or other temporally coincident signals can elicit similar functional interactions (Maier, Chandrasekaran, & Ghazanfar, 2008; Noesselt et al., 2007). The second is that there are other areas that, consistent with their connectivity and response properties (e.g., sensitivity to faces and voices), could also (and very likely) have a visual influence on auditory cortex. These include the ventrolateral prefrontal cortex (Romanski, Averbeck, & Diliz, 2005; Sugihara, Diliz, Averbeck, & Romanski, 2006) and the amygdala (Gothard, Battaglia, Erickson, Spiteri, & Amaral, 2007; Kuraoka & Nakamura, 2007).

**PHASE Resetting AS A Mechanism For Audiovisual Integration In Auditory Cortex**

The auditory cortical data suggest that, even at the “earliest” cortical stage of vocalization processing, visual cues can modulate auditory responses. One candidate mechanism for this modulation is “phase resetting” (Lakatos, Chen, O’Connell, Mills, & Schroeder, 2007). Phase resetting occurs when an ongoing neural oscillation, in essence, starts over after the presentation of a stimulus (figure 22.7A). Schroeder and colleagues (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008), based on results from somatosensory-auditory integration (Lakatos et al., 2007), hypothesized that, during audiovisual speech, the onset of mouth motion prior to the voice could lead to a phase reset of ongoing oscillations in the auditory cortex. Subsequent auditory inputs falling on high-excitability peaks of this reset oscillation will be amplified, whereas auditory inputs falling on the low-excitability peaks of this oscillation would be suppressed (figure 22.7B). Schroeder and colleagues originally proposed it for LFPs and multiunit activity, but this idea could be extended to single neurons quite readily. For example, if LFPs are thought to reflect synaptic input and thereby membrane potential dynamics (Logothetis, 2002; Okun, Naim, & Lampl, 2010), then

**Figure 22.7** Phase resetting as a mechanism for enhancement and suppression. (A) Before the onset of the visual stimulus, the phase of the auditory cortical oscillation is random, and thus, their mean across trials is flat, and the phase distribution in the circular plot is widely dispersed. However, once the face comes on, the oscillations are reset and thus aligned with each other, leading to a consistent phase distribution over trials. (B) In the overall mechanism of multisensory integration, the visual signal resets the phase of the ongoing oscillation (as in A), and when the auditory input arrives in a high-excitability phase, then the response is enhanced. In contrast, when the auditory input arrives in the low-excitability region of the reset oscillation, then the response is suppressed.
one hypothesis is that the onset of mouth motion depolarizes or hyperpolarizes the membrane potential of auditory neurons, and the time of arrival of auditory inputs would lead to multisensory responses being either enhanced or suppressed relative to the auditory-alone responses. If one assumes that membrane potential dynamics track the LFP (Logothetis, 2002; Okun et al., 2010), and that visual inputs are weak and modulatory, this change in the membrane potential would in principle be observed as a phase reset in the LFP. More modeling and experimental studies are, however, needed to verify whether phase resetting is the result of a modulation of membrane potential.

We would like to note one potential problem with the phase-resetting hypothesis and its applicability to communication signals. The original phase-resetting hypothesis claim was made based on physiological data generated by using punctate stimuli were <10 msec in duration) auditory and somatosensory stimuli. Therefore, the prediction is not straightforward with regard to how continuous mouth motion seen during the production of vocalizations would modify such processing in primary auditory cortex or how shifting eye movements of observers (Ghazanfar et al., 2006; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998) would modify the phase resetting of ongoing oscillations (Ghazanfar & Chandrasekaran, 2007). A dynamic, vocalizing face is a complex sequence of sensory events, but one that elicits fairly stereotypical eye movements: we and other primates fixate on the eyes but then saccade to mouth when it moves before saccading back to the eyes (Ghazanfar et al., 2006; Lansing & McConkie, 2003). Eye position influences single-neuron and local field potential activity in multiple regions of auditory cortex (Fu et al., 2004; Werner-Reiss, Kelly, Trause, Underhill, & Groh, 2003). Therefore, one possibility is that the eye fixations at the onset of mouth movements (figure 22.2B) send a signal to the auditory cortex, which resets the phase of an ongoing oscillation. Such effects have, for example, already been seen in the primary visual cortex (Rajkai, Lakatos, & Schroeder, 2008). This eye movement signal thus primes the auditory cortex to amplify or suppress (depending on the timing) the neural response to a subsequent auditory signal originating from the mouth. Because mouth movements precede the voiced components of both human (Abry et al., 1996; Chandrasekaran & Ghazanfar, 2009; Chandrasekaran et al., 2009) and monkey vocalizations (Chandrasekaran & Ghazanfar, 2009; Ghazanfar et al., 2005), the temporal order of visual to auditory signals is consistent with this idea. This hypothesis is also supported (though indirectly) by the finding that suppressed versus enhanced responses to face/voice signals in the auditory cortex and the STS are influenced by the timing of mouth movements relative to the onset of the voice (Chandrasekaran & Ghazanfar, 2009; Ghazanfar et al., 2005). Finally, as STS neurons seem to be involved in saccades to visual targets (Scalaidhe, Albright, Rodman, & Gross, 1995) and respond to mouth movements and biological motion (Barraclough et al., 2005; Ghazanfar et al., 2010; Ghazanfar et al., 2005; Perrett, Rolls, & Caan, 1982; Puce & Perrett, 2003), they would be a perfect candidate again to modulate auditory cortex during this process (Ghazanfar et al., 2008).

SUMMARY AND CONCLUSIONS

Nonhuman primate vocalizations and speech comprise multisensory signals. Both afford orofacial movements with acoustic, spectrally rich vocalizations. Such orofacial movements are thought to assist in improved perception, especially in noisy and unstructured auditory environments, helping the receiver better segment auditory cues from the cacophony of signals around (Heller & Freyman, 2005; Ross et al., 2007; Sumby & Pollack, 1954). Several cortical areas in the temporal, parietal, and frontal lobes integrate visual and auditory cues. Our analysis here focused on the multisensory properties of regions in the temporal lobe such as auditory cortex and the upper bank of the STS. In particular, auditory neurons are modulated by the visual cues present in communication signal, and one source of such visual modulation is the upper bank of the STS. Phase resetting seems to be an attractive mechanism for explaining the enhancement and suppression of multisensory responses (see also van Wassenhove et al., chapter 9, in this volume).

In the nonhuman animal literature, the wealth of signal structure and behavioral and physiological data related to multisensory processing have rarely come from the same experiments. As a result, very little is known with regard to how multisensory neurophysiological phenomena directly relate to the structure of sensory signals and, more importantly, to behavior. Obviously though it may seem, it is time to combine behavior and physiology in a single multisensory communication experimental paradigm. We could then carefully test hypothetical multisensory mechanisms that the independent behavioral and physiological data suggest. The use of synthetic agents (avatars) may be critical in this regard (Steckenfinger & Ghazanfar, 2009), not only allowing us to manipulate the acoustic signal in the many sophisticated ways that we already do but also providing us with the means to parametrically manipulate facial dynamics.
REFERENCES


