Behavioral/Systems/Cognitive

Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex

Asif A. Ghazanfar, Joost X. Maier, Kari L. Hoffman, and Nikos K. Logothetis
Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany

In the social world, multiple sensory channels are used concurrently to facilitate communication. Among human and nonhuman primates, faces and voices are the primary means of transmitting social signals (Adolphs, 2003; Ghazanfar and Santos, 2004). Primates recognize the correspondence between species-specific facial and vocal expressions (Massaro, 1998; Ghazanfar and Logothetis, 2003; Izumi and Kojima, 2004), and these visual and auditory channels can be integrated into unified percepts to enhance detection and discrimination. Where and how such communication signals are integrated at the neural level are poorly understood. In particular, it is unclear what role "unimodal" sensory areas, such as the auditory cortex, may play. We recorded local field potential activity, the signal that best correlates with human imaging and event-related potential signals, in both the core and lateral belt regions of the auditory cortex in awake behaving rhesus monkeys while they viewed vocalizing conspecifics. We demonstrate unequivocally that the primate auditory cortex integrates facial and vocal signals through enhancement and suppression of field potentials in both the core and lateral belt regions. The majority of these multisensory responses were specific to face/voice integration, and the lateral belt region shows a greater frequency of multisensory integration than the core region. These multisensory processes in the auditory cortex likely occur via reciprocal interactions with the superior temporal sulcus.

Key words: **crossmodal; speech; vocalization; bimodal; superior temporal sulcus; temporal lobe**

Introduction

Multisensory integration refers to the influence of one sensory modality over another in the form of enhancement or suppression relative to the strongest "unimodal" response (Stein and Meredith, 1993). This definition applies to both behavioral and neural responses. Our current knowledge of bimodal integration of visual and auditory primate vocal signals in the brain is derived almost exclusively from human neuroimaging studies of audiovisual speech. The superior temporal sulcus (STS) and superior temporal gyrus of the temporal lobe are consistently activated by bimodal speech signals and often show enhanced activity over unimodally induced signals (Calvert et al., 2000; Callan et al., 2003; Wright et al., 2003). Furthermore, a recent study in rhesus monkeys has confirmed such integration in the STS at the level of single units for biologically meaningful actions (Barraclough et al., 2005). Beyond the STS, there are conflicting reports with regard to multisensory speech processing. It is unclear, for example, to what extent auditory cortical regions within the superior temporal plane integrate bimodal speech information. Some studies indicate that the auditory cortex plays a role in multisensory speech integration (Sams et al., 1991; Calvert et al., 1999; Callan et al., 2003), whereas others suggest that it does not (Bernstein et al., 2002; Olson et al., 2002; Wright et al., 2003). Furthermore, most studies report only response enhancement to congruent audiovisual speech tokens (Calvert et al., 1999, 2000; Callan et al., 2003), whereas others suggest a mixture of enhancement and suppression (Wright et al., 2003) or only response suppression (Besle et al., 2004; van Wassenhove et al., 2005).

In light of the ambiguous human neuroimaging data and the lack of relevant data from animal models, we examined the issue of dynamic face/voice integration in the presumptive unimodal areas of the rhesus monkey auditory cortex using the natural communication signals of the species. There are several characteristics about rhesus monkey vocalizations that make them interesting for multisensory integration studies. First, unlike pairings of artificial stimuli, audiovisual vocalizations are ethologically relevant and thus may tap into specialized neural mechanisms (Ghazanfar and Santos, 2004) or, minimally, integrative mechanisms for socially learned audiovisual associations. Second, the spatiotemporal complexity of facial and vocal signals may uncover principles of multisensory integration that cannot be revealed with the use of simple, static stimuli. Third, rhesus monkey vocalizations are characterized by temporal dynamics common to human speech. In particular, the onset of facial movement related to articulation occurs before the onset of the auditory signal. Thus, the use of species-typical vocal signals allows the assessment of neural homologies between nonhuman primate vocal communication systems and human speech.

While our monkey subjects viewed unimodal and bimodal versions of their species-typical vocalizations, we recorded the mean extracellular field potential (i.e., unit and subthreshold neural activity) using intracranial electrodes placed in the core region (which includes primary and primary-like auditory areas)

Received March 1, 2005; revised March 29, 2005; accepted April 12, 2005.

This work was supported by the Max Planck Society. We thank Christoph Kayser, David Leopold, Yusuke Murayama, Kristina Nielsen, Chris Petkov, and Gregor Rainer for their helpful advice during the course of this study.

Correspondence should be addressed to Asif A. Ghazanfar, Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544. E-mail: asifg@princeton.edu.

DOI:10.1523/JNEUROSCI.0799-05.2005

and the lateral belt area of the auditory cortex, a higher-order region (Hackett, 2002). From this neural signal, we focused on the local field potential (LFP) activity. Our reasons for doing so are twofold: (1) LFPs allow direct comparisons to be made with human studies using blood oxygenation level-dependent imaging (Mathiesen et al., 1998; Lauritzen, 2001; Logothetis et al., 2001; Kayser et al., 2004) or evoked potentials; and (2) robust LFP responses to "extra"-sensory signals seen in unimodal cortical areas are frequently unaccompanied by corresponding increases/decreases in the spiking of neurons recorded from the same cortical site (Schroeder et al., 2001; Schroeder and Foxe, 2002; Fu et al., 2004; Gail et al., 2004), perhaps because of the sparseness of the "integrating" neurons and/or because of laminar variations in neural signal processing.

Materials and Methods

Subjects and surgery. Two adult male rhesus monkeys (*Macaca mulatta*) were used in the experiments. For each monkey, we used preoperative whole-head magnetic resonance imaging (4.7 T magnet, 500 μm slices) to identify the stereotaxic coordinates of the auditory cortex and to model a three-dimensional skull reconstruction. From these skull models, we constructed custom-designed, form-fitting titanium head posts and recording chambers (Logothetis et al., 2002). The monkeys underwent sterile surgery for the implantation of a scleral search coil, head post, and recording chamber. The inner diameter of the recording chamber was 19 mm and was vertically oriented to allow an approach to the superior surface of the superior temporal gyrus (Pfingst and O'Connor, 1980; Recanzone et al., 2000). All experiments were performed in compliance with the guidelines of the local authorities (Regierungspraesidium, Tuebingen, Germany) and the European Union (European Communities Council Directive 86/609/EEC) for the care and use of laboratory animals.

Stimuli. The naturalistic stimuli were digital video clips of vocalizations produced by rhesus monkeys in the same colony as the subject monkeys. The stimuli were filmed while monkeys spontaneously vocalized in a primate restraint chair placed in a sound-attenuated room. This ensured that each video had similar visual and auditory background conditions and that the individuals were in similar postures when vocalizing. Vocalizations were four coos and four grunts. Videos were acquired at 30 frames per second (frame size, 720 × 480 pixels), whereas the audio tracks were acquired at 32 kHz and 16 bit resolution in mono. Across the vocalizations, the audio tracks were matched in average rms energy. The clips were cropped to the beginning of the first mouth movement to the mouth closure at the end of vocalization (see Fig. 1). The duration of the video clips varied according to the vocalization.

To test for the possibility that any multisensory integration that we observed was specific to faces and not just any arbitrary visual stimulus paired with the voice, we ran a control condition. Because there are many possible control stimuli for faces (none of which are ideal), we decided to use controls for which there are behavioral data. These were videos that mimicked the dynamics of the mouth movements in our natural video stimuli. In a psychophysical study, Bernstein et al. (2004) compared speech detection thresholds in noise for human subjects viewing the corresponding face versus visual control stimuli. The control stimuli consisted of an arbitrary shape on a gray background for which diameter in one axis varied dynamically with the amplitude of the speech signal. Therefore, the shape diameter mimicked very closely the opening of the mouth. Human subjects could enhance their auditory speech detection with such control stimuli, but not as well as with the face stimuli (Bernstein et al., 2004). Thus, we adopted a similar control strategy for our neural responses.

Our artificial mouth-movement videos were generated in Matlab (MathWorks, Natick, MA) using the Psychophysics Toolbox extensions (www.psychtoolbox.org). They consisted of expanding/contracting circular black disks on a gray background and mimicked the dynamics (opening, closing, and displacement) of the mouth in the natural videos. For each frame of each natural video, the position and size of the mouth were estimated and an approximately matching still frame of a disk was generated. This frame was compared with the corresponding frame of the natural video by overlaying the two frames using Adobe Premiere 6.0 software (Adobe Systems, San Jose, CA). The position and size of the disk were then adjusted until it approximated the diameter of the mouth in the corresponding frame of the natural video. This procedure was repeated frame by frame, and movies were generated by creating sequences of such frames.

Another possible control would have been to include "incongruent" conditions, whereby, for example, a coo face was paired with a grunt voice. Similar controls have been applied in the human imaging literature (Calvert et al., 2000). The use of such stimuli is problematic for us, because coos and grunts are of vastly different durations (Fig. 1), and we would thus have to compress or expand signal components to generate temporally matched stimuli; such a manipulation would generate species-atypical signals. Furthermore, coos and grunts do not have different meanings; they are both affiliative calls produced in many contexts, and mismatching them would not produce a semantically incongruent stimulus of the kind typically used in human multisensory studies (Calvert et al., 2000).

Behavioral apparatus and paradigm. Experiments were conducted in a double-walled, sound-attenuating booth lined with echo-attenuating foam. The monkey sat in a primate restraint chair in front of a 21 inch color monitor at a distance of 94 cm. On either side of the monitor were two speakers placed in the vertical center. Two speakers were used to reduce the spatial mismatch between the visual signals and the auditory signals.

The monkeys performed in a darkened booth. A trial began with the appearance of a central fixation spot. The monkeys were required to fixate on this spot within a 1 or 2° radius for 500 ms. This was followed by (1) the appearance of a video sequence with the audio track, (2) the appearance of the video alone (no audio), or (3) the audio track alone (black screen). The videos were displayed centrally at 10 × 6.6°, and the audio track was played at ~72 dB (as measured by a sound-level meter at 94 cm; C-weighted). In the visual conditions, the monkeys were required to view the video for its duration by restricting their eye movements to within the video frame. Successful completion of a trial resulted in a juice reward. Eye-position signals were digitized at a sampling rate of 200 Hz.

Data collection. Recordings were made from the core and lateral belt regions of the left auditory cortex using standard electrophysiological techniques. We used a custom-made multielectrode drive that allowed us to move up to eight electrodes independently. The minimum interelectrode distance was >2.0 mm. Guide tubes were used to penetrate the overlying tissue growth and dura. Electrodes were glass-coated tungsten wire with impedances between 1 and 3 MΩ (measured at 1 kHz). The stainless-steel chamber was used as the reference. Signals were amplified, filtered (1–5000 Hz), and acquired at a 20.2 kHz sampling rate. Electrodes were lowered until multiunit cortical responses could be driven by auditory stimuli. Search stimuli included pure tones, frequency-modulated sweeps, noise bursts, clicks, and vocalizations. Using the analog multiunit activity (MUA) signal (high-pass filtered at 500 Hz), frequency tuning curves were collected for each site using 25 pure tone pips (100 Hz to 21 kHz) delivered at a single intensity level (72 dB). Peak tuning is identical for both MUA and LFP signals in the auditory cortex (Norena and Eggermont, 2002). In both monkeys, we discerned a coarse tonotopic map representing high-to-low frequencies in the caudal-to-rostral direction. Such a map is identified as primary auditory cortex (A1). Lateral belt areas are collinear with tonotopic areas in the core region (Hackett, 2002). The lateral belt area adjacent to A1 is the "middle lateral belt area." This area was distinguished from A1 by its greater sensitivity to complex sounds than to pure tones, as reported in previous studies in both anesthetized (Rauschecker et al., 1995) and awake (Barbour and Wang, 2003) monkeys. These physiological criteria serve only as a rough guide, and it is likely that some of our electrodes were placed in rostrally adjacent regions in both the core and belt. We therefore make reference to only core or lateral belt throughout.

Data analysis. LFPs (the low-frequency range of the mean extracellular field potential) were extracted off-line by bandpass filtering the signal between 1 and 300 Hz using a four-pole bidirectional Butterworth filter.

Because we did not control for laminar location in our recordings, the signals were then full-wave rectified to allow an unambiguous assessment of power changes within the signal, according to stimulus conditions. Results from the two monkey subjects were similar; therefore, they are considered together in all analyses.

Responses to each stimulus condition (face plus voice, voice alone, and face alone) were assessed based on mean modulation in microvolts evoked. The responses were based on the averages of 10 trials per stimulus, and all stimuli were presented in random order. The LFP responses to these time-varying auditory and visual signals were phasic; they were not typically sustained for the duration of the stimulus. Furthermore, because of the variability of voice-onset times, we had no a priori knowledge of the latencies of any multisensory responses. For these reasons, a 20 ms window around peak responses in any condition was extracted between the onset and offset of the auditory stimulus, and the responses for the two other conditions were taken within this window and compared statistically with baseline activity and with each other using Wilcoxon sign–rank tests. If a response was significantly different from baseline activity (300 ms before video onset), we then determined whether the multisensory condition was significantly different from the strongest unimodal response. This unimodal response was invariably the auditory response, as expected. A similar neural analysis protocol was applied to multisensory neural responses in the superior colliculus of awake monkeys (Bell et al., 2003). We then used a well established multisensory integration index to calculate the magnitude of enhancement or suppression (Meredith and Stein, 1986): $[(\mathrm{MM} - \mathrm{SM}_{\mathrm{max}})/(\mathrm{SM}_{\mathrm{max}})] \times 100 = \%$ interaction, where MM is the mean response evoked by the multimodal stimulus and $\mathrm{SM}_{\mathrm{max}}$ is the mean response evoked by the most effective unimodal stimulus, which was invariably in the auditory domain. Although there are alternative methods for calculating the multisensory integration index [particularly in the imaging literature; for discussion, see Beauchamp et al. (2004)], we adopted the method developed by Stein and colleagues (Meredith and Stein, 1986; Wallace et al., 1996). This method has been frequently applied across different animal models, neural structures, and behavioral paradigms by different research groups (Meredith et al., 1987; Frens and Van Opstal, 1998; Bell et al., 2003; Barraclough et al., 2005). Thus, our findings are best interpreted within the existing physiology literature by adopting the same methodology.

Results

We analyzed rectified LFPs from 46 sites in the core region and 35 sites in the lateral belt region of the auditory cortex while the subjects watched and/or heard conspecifics producing two types of affiliative vocalizations: coos and grunts (Hauser and Marler, 1993). Both types of vocalizations

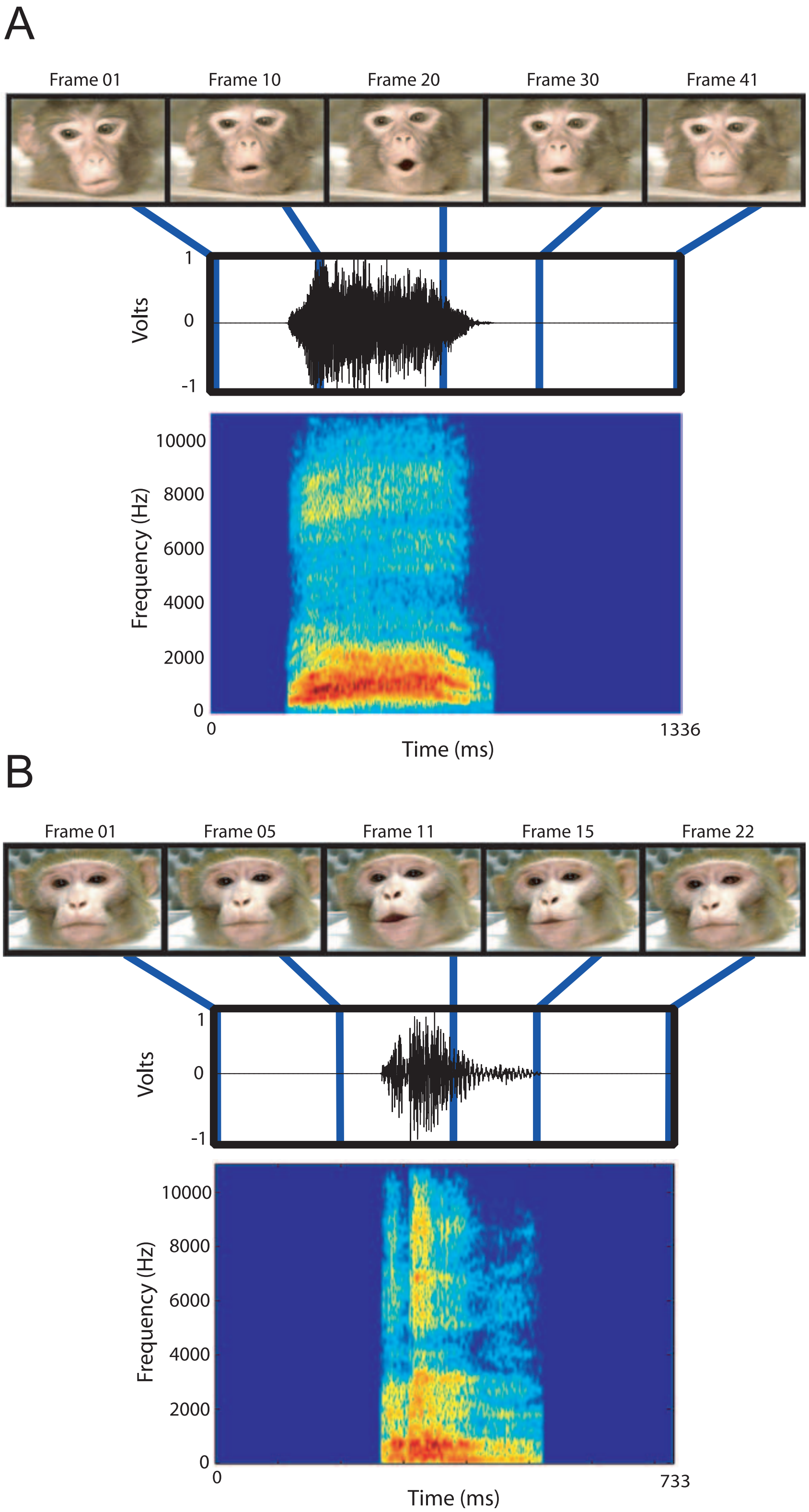

Figure 1. Exemplars of the visual and auditory components of the two types of vocalizations used in this study. Top panels show representative frames at five intervals from the start of the video (the onset of mouth movement) until the end of mouth movement. Middle panels display the time waveform of the auditory component of the vocalization, in which the blue lines indicate the temporally corresponding video frames. Bottom panels show the spectrogram for the same vocalization. ***A***, The coo vocalization. Coos are long-duration, tonal calls produced with protruded lips. ***B***, The grunt vocalization. Grunts are short-duration, noisy calls produced with a subtle mouth opening relative to coos. For both vocalizations, the mouth-movement onset precedes the auditory component.

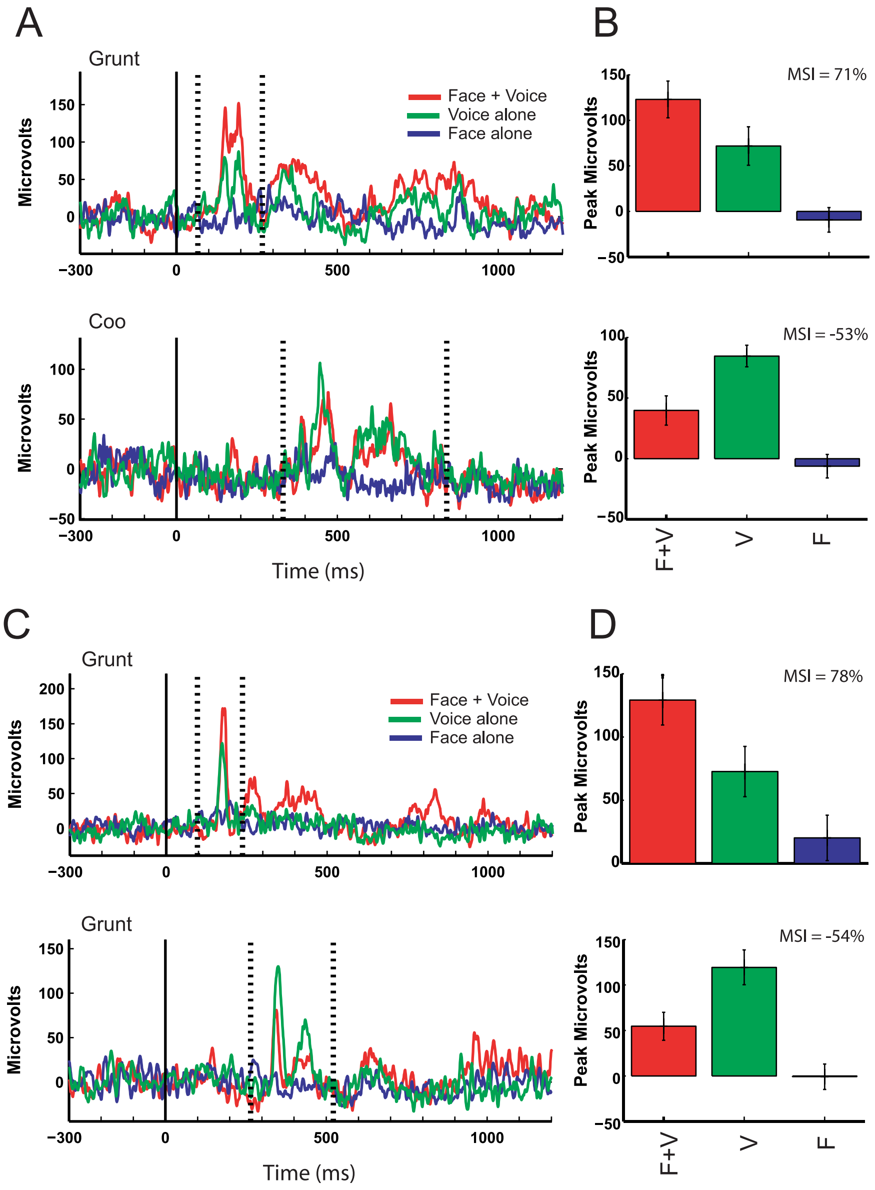

Figure 2. Auditory cortical responses to multimodal vocalizations. Rectified local field potential responses to face plus voice (F+V), voice alone (V), and face alone (F) components of coos and grunts were compared. The solid vertical line indicates the onset of the face signal. Dotted vertical lines indicate the onset and offset of the voice signal. Graphs represent the mean of 10 repetitions with the mean baseline activity subtracted on a trial-by-trial basis. Bar graphs show the mean and SEM of the maximum response (face plus voice or voice alone using a 20 ms window; see Materials and Methods) between the voice onset and offset. This response was then compared statistically with the responses for the other conditions. A multisensory integration (MSI) index was computed using these responses and is indicated at the top right of each bar graph. ***A***, ***B***, One enhanced response and one suppressed response from the auditory core region. ***C***, ***D***, One enhanced response and one suppressed response from the lateral belt region.

have unique auditory and visual components (Hauser et al., 1993; Partan, 2002). Coos are long-duration, tonal calls produced with the lips protruded (Fig. 1*A*). Grunts are short-duration, noisy calls produced with a more subtle mouth opening with no lip protrusion (Fig. 1*B*). Both call types are spectrally rich and wideband. As in human speech (Abry et al., 1996), the onsets of mouth movements during production of rhesus monkey vocal signals precede the auditory component. During the presentation of video conditions, monkeys were required to maintain fixation within the video frame (see Materials and Methods).

LFPs in both auditory cortical regions showed robust multisensory integration of dynamic faces and voices. This integration took the form of either enhancement or suppression. Figure 2, *A* and *B*, shows a significantly enhanced response to a grunt (top) and a significantly suppressed response to a coo (bottom) from cortical sites in the core region. These responses occurred after the onset of the voice signal. Similarly, responses to coos and grunts from cortical sites in the lateral belt region also showed robust multisensory enhancement (Fig. 2*C*,*D*, top) and suppression (Fig. 2*C*,*D*, bottom).

The auditory core and lateral belt regions differed in the extent to which they expressed multisensory integration (Fig. 3*A*). Cortical sites in both regions could show enhancement only, suppression only, or both enhancement and suppression together, depending on the stimulus, but the percentage of sites showing multisensory integration was significantly greater in the lateral belt region (88.24%) than in the auditory core (71.74%) (χ^2 test for independence; df = 1; p = 0.005). Among sites that showed multisensory integration, the lateral belt also had a trend toward a greater percentage of sites that expressed both enhancement and suppression (depending on the stimulus) than the core: 41.18 versus 19.57%, respectively (Fig. 3*A*). The distribution of the peak latencies of these multisensory responses is shown in Figure 3*B*. The median peak latency for the auditory core was 84 ms; in the lateral belt, the median latency was 93.5 ms. These values likely correspond to the auditory N100 peak reported in the event-related potential literature.

Grunts versus coos: enhancement and suppression

Across both areas (i.e., all sites that showed some form of multisensory integration), there were significantly more instances of enhancement than suppression (Fig. 4). A two-way ANOVA, with enhancement versus suppression as one factor and grunts versus coos as a second factor, revealed a significant main effect for frequency of enhancement versus suppression ($F_{(1,236)}$ = 27.65; $p < 0.0001$) (Fig. 4, dark vs light bars) and a significant main effect for grunts versus coos ($F_{(1,236)}$ = 4.34; p = 0.0383). There was also a significant interaction between these two factors ($F_{(1,236)}$ = 4.34; p = 0.0383), indicating that the frequency of enhancement was greater for grunts than for coos (p = 0.004) (Fig. 4, dark bars). For a given site, there was a 23.35% chance of observing an enhanced multisensory response to any given grunt but only a 14.75% chance that a coo would elicit such a response.

Temporal factors

In multisensory paradigms using artificial visual–auditory stimuli, the magnitude of multisensory enhancement or suppression

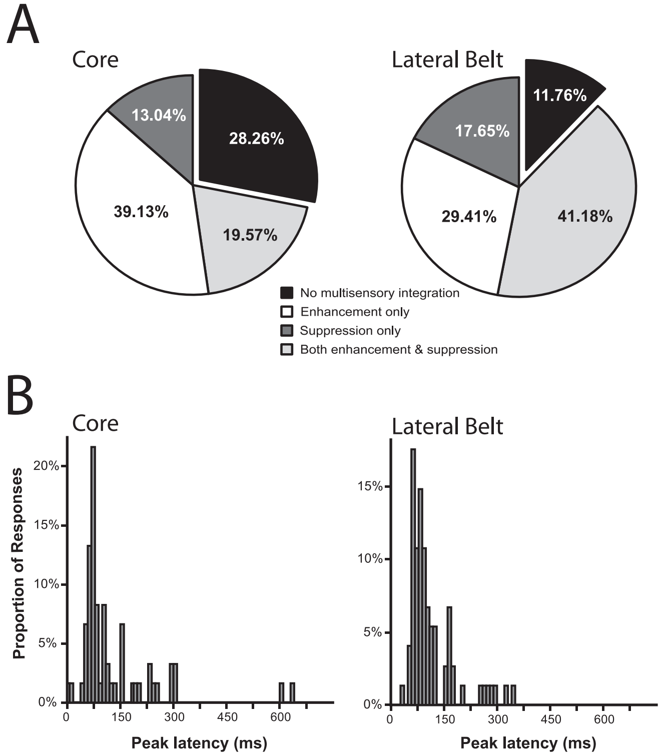

Figure 3. Multisensory integration across two auditory cortical regions. ***A***, The relative amounts of multisensory integration seen across cortical sites. The percentages represent the fraction of the total number of sites in the auditory core region ($n = 46$ sites) and the lateral belt region ($n = 35$ sites). The lateral belt had significantly more sites demonstrating multisensory integration. ***B***, The distribution of peak latencies for the core and lateral belt regions. These data represent the peak amplitude of statistically significant multisensory responses in the LFP signals. Responses were assessed between the onset and offset of the auditory component of the vocalizations; thus latencies are relative to the auditory onset.

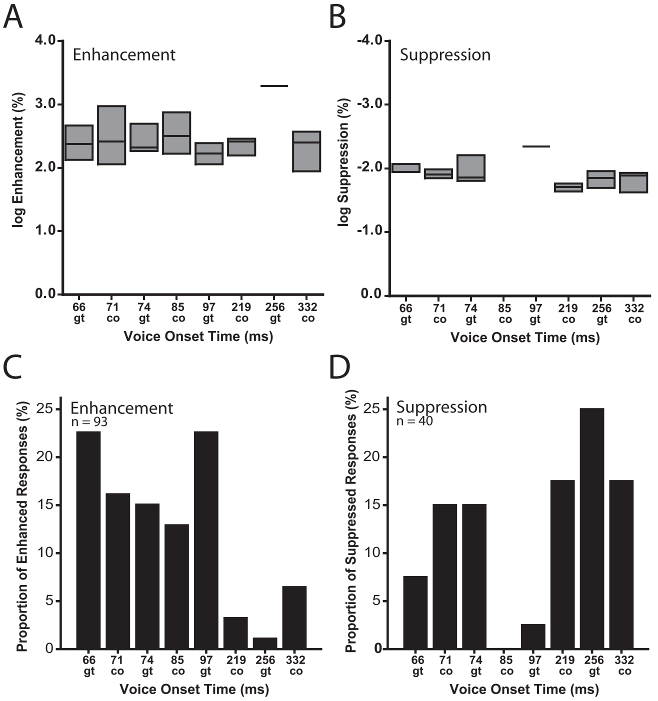

Figure 5. Relationship between voice-onset time and multisensory integration. ***A***, ***B***, Median (black lines) and interquartile ranges (gray boxes) of enhancement and suppression relative to voice-onset time. The *x*-axis represents voice-onset time; the *y*-axis represents the log10-base percentage of the multisensory integration index value. gt, Grunts; co, coos. Note that in ***A***, there was only one enhancement response in the "256 ms/gt" category, whereas in ***B***, there was no response in the "85 ms/co" category and only one response in the "97 ms/gt" category. The magnitude of multisensory effects was not related to voice-onset time. ***C***, ***D***, Proportion of enhanced ($n = 93$) and suppressed ($n = 40$) responses across the different voice-onset time categories. Note that enhancement was more frequently observed for short voice-onset times, whereas suppression was more common at longer voice-onset times.

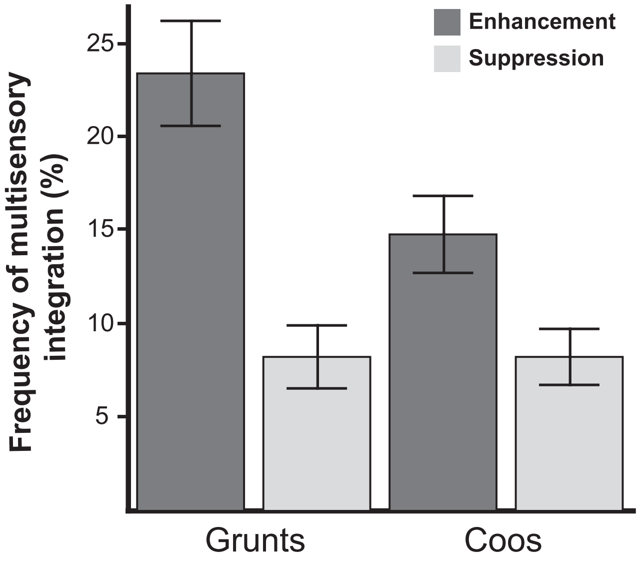

Figure 4. The average frequency of multisensory integration seen across all sites. For both cortical regions, there were more instances of enhancement than suppression, and grunts more frequently elicited enhancement than did coos. Error bars represent SEM.

for a given neuron is related to the temporal disparity between the visual and auditory stimuli (Meredith et al., 1987; Wallace et al., 1996). If the multisensory voice stimuli were processed as artificial stimuli are in the superior colliculus, one would predict that, across a population of responses, the magnitude of enhancement or suppression would covary with the voice-onset time relative to the initiation of mouth movement. Across our eight stimuli, there was a range of voice-onset times, from 66 to 332 ms, and voice-onset times for grunts and coos overlapped throughout the range. Figure 5, *A* and *B*, reveals that there is no correlation between voice-onset time and the magnitude of multisensory enhancement ($r = -0.038$; $p = 0.719$) (Fig. 5*A*) or suppression ($r = -0.307$; $p = 0.054$) (Fig. 5*B*). Nevertheless, the overall proportion of enhanced and suppressed responses is strongly influenced by voice-onset time. Figure 5, *C* and *D*, shows the percentage of enhanced or suppressed responses according to the voice-onset time. A greater number of enhanced responses occurred for shorter voice-onset times (Fig. 5*C*). In contrast, suppressed responses primarily occurred when voice-onset times were longer (Fig. 5*D*). The differences between these distributions were highly significant (χ^2 test for independence; df = 7; $p = 2.98 \times 10^{-13}$).

Law of inverse effectiveness

One of the hallmarks of multisensory integration is the principle of inverse effectiveness (Stein and Meredith, 1993): the level of multisensory enhancement is inversely related to the strength of the unimodal responses. Thus, the weaker the unimodal responses, the greater the multisensory enhancement is likely to be. We tested this rule in our data by correlating the level of enhanced multisensory integration, as measured by the multisensory integration index (see Materials and Methods), with the corresponding response magnitude in the auditory alone condition. In the

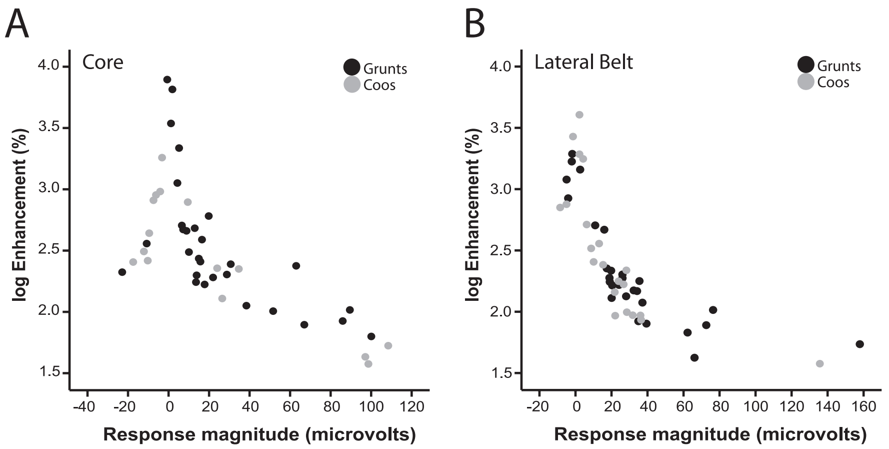

Figure 6. Auditory cortical LFP enhancements obey the law of inverse effectiveness, whereby the degree of multisensory enhancement is inversely related to the magnitude of the unimodal response. The *y*-axes depict the log10-base percentage enhancement calculated from the multisensory integration index (see Materials and Methods). The *x*-axes depict the corresponding response magnitude of the auditory alone response. Gray dots represent coo responses; black dots represent grunt responses. ***A***, Responses from the auditory core region. ***B***, Responses from the lateral belt region.

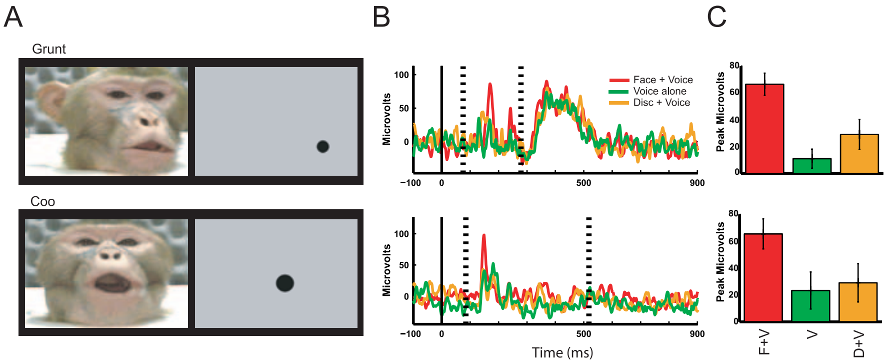

Figure 7. Face specificity in the multisensory responses of the auditory cortex. ***A***, One frame of a grunt and coo face at maximal mouth opening for one stimulus monkey and the corresponding frames from the disk control videos. ***B***, Examples of rectified LFP responses to face plus voice, voice alone, and disk plus voice conditions corresponding to the stimuli in ***A***. Conventions are as in Figure 2. ***C***, Bar graphs of peak responses corresponding to ***B***. F+V, Face plus voice; V, voice alone; D+V, disk plus voice. Error bars represent SEM.

core region, there was a strong negative correlation between multisensory enhancement and unimodal auditory response magnitudes for both grunts ($n = 29$; $r = -0.625$; $p < 0.001$) and coos ($n = 15$; $r = -0.838$; $p < 0.001$) (Fig. 6*A*). A similar pattern was seen in the lateral belt for both grunts ($n = 28$; $r = -0.729$; $p < 0.001$) and coos ($n = 21$; $r = -0.694$; $p < 0.001$) (Fig. 6*B*). Thus, multisensory LFP responses in the rhesus monkey auditory cortex adhere to the principle of inverse effectiveness.

Face-selective multisensory integration

It is possible that the enhancement and suppression that we observed could be induced by any visual stimulus and a vocalization. To control for this, we designed artificial movies based on those used to study the psychophysics of speech reading (Bernstein et al., 2004) (for justification, see Materials and Methods). Our stimuli consisted of a dynamic black disk on a light gray background. The diameter of the disk and the position of the disk mimicked the diameter of mouth opening and mouth position on a frame-by-frame basis (Fig. 7*A*) (see Materials and Methods). In essence, the dynamic disk stimuli control for both the onset and offset of a generic visual stimulus and visual motion in the mouth region.

We analyzed and compared the multisensory responses in the core and lateral belt regions for face plus voice integration and the disk plus voice integration. Figure 7, *B* and *C*, shows cortical sites with significant multisensory integration for the face plus voice condition but not for the corresponding disk plus voice condition. Across all of the significant multisensory responses to the face plus voice condition in the core region, 67% were specific to faces; that is, the disk plus voice condition also did not elicit a significant multisensory response. In the lateral belt region, 80% of face plus voice multisensory responses were specific and did not integrate to the disk plus voice condition. Figure 8*A* shows exemplars of the minority of "nonspecific" multisensory responses in which cortical sites integrated both the face plus voice and the disk plus voice in a similar manner. Figure 8*B* shows a response in which there was integration of disk plus voice but not of face plus voice. Such "disk-specific" multisensory responses comprised only a small fraction of responses. On a per-cortical-site basis in the core region, there were 0.46 multisensory responses for disk plus voice (but not face plus voice), compared with 0.89 face plus voice (but not disk plus voice) multisensory responses per site ($t_{(38)} = 2.54$; $p = 0.015$; Bonferroni's corrected, $p = 0.017$) (Fig. 9). Similarly, for cortical sites in the lateral belt, there were 0.31 disk plus voice multisensory responses per site but 1.71 face plus voice multisensory responses per site ($t_{(34)} = 5.45$; $p < 0.0001$; Bonferonni's corrected, $p = 0.017$) (Fig. 9). In addition, there were significantly more face plus voice multisensory responses in the lateral belt than in the core region ($t_{(71)} = 3.03$; $p = 0.003$; Bonferonni's corrected, $p = 0.017$). Thus, the vast majority of responses were specific to the association of faces and voices.

Field potentials versus spiking activity

Our LFP data showed robust multisensory integration of faces and voices. To assess to what extent such integration can be seen in spiking activity, we conducted identical analyses on our analog MUA signal. Only a small subset of cortical sites showed multisensory integration in the MUA signal. In the core region, only 35% of the cortical sites showed significant multisensory MUA responses, and only 40% did so in the lateral belt (data not shown). This is in contrast to 73 and 89% of core and lateral belt sites, respectively, for the LFP signal (Fig. 3*A*).

Discussion

Previous neurophysiological experiments of multisensory integration in animal models have primarily been confined to studies of spatial and temporal integration of artificial stimuli (for review, see Stein and Meredith, 1993). Although we have learned a great deal from such studies (indeed, they have laid the foundation on which all multisensory neuroscience is built), crossmodal

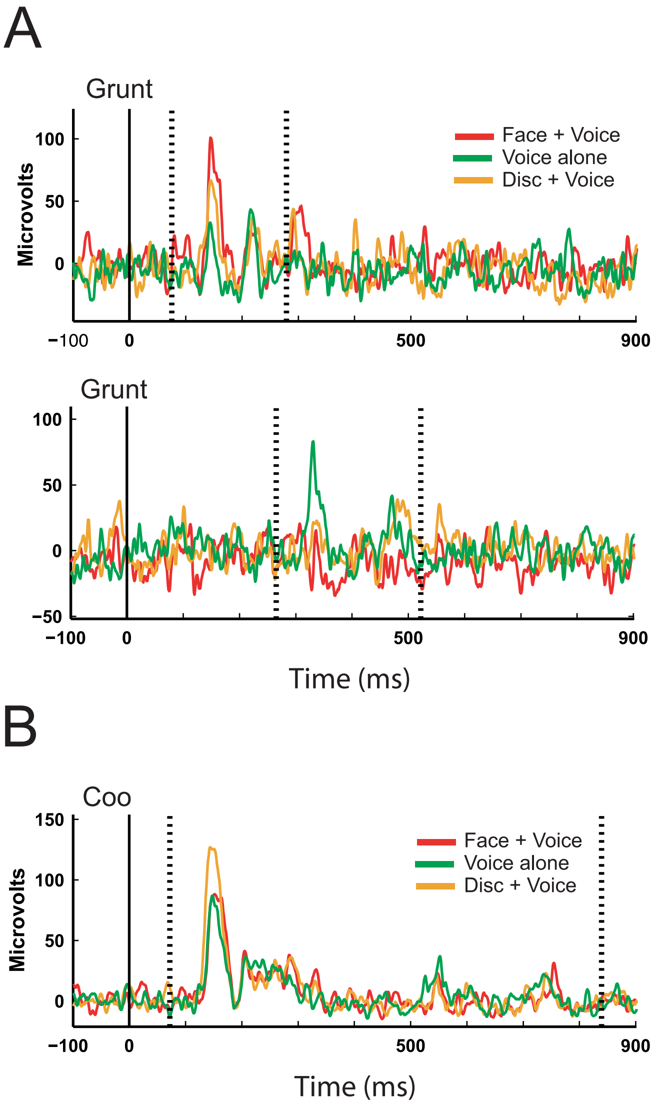

Figure 8. ***A***, Examples of responses integrating both face plus voice and disk plus voice. In both examples, the face plus voice and disk plus voice responses were significantly different from the voice alone condition ($p < 0.05$). ***B***, Example showing integration of disk plus voice only. The disk plus voice response was significantly different from the other two response conditions ($p < 0.05$).

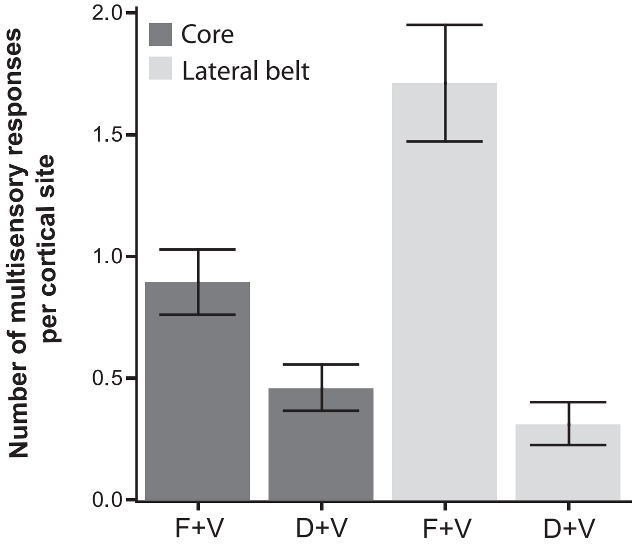

Figure 9. For a given cortical site, the frequency of face plus voice (F+V) multisensory responses exceeds that of disk plus voice (D+V) responses. The *y*-axis indicates the frequency of observing multisensory responses for a given cortical site. The maximum number of possible responses is eight (the number of stimuli). Dark gray bars represent the core region of the auditory cortex, whereas light gray bars represent the lateral belt. Error bars represent mean and SE.

“identification” (Calvert, 2001), like audiovisual speech, has not been explored neurophysiologically in animals. The current data unequivocally demonstrate that local field potentials in the auditory cortex are capable of multisensory integration of facial and vocal signals (i.e., “crossmodal identification,” in rhesus monkeys). The vast majority of responses were specific to face plus voice integration, and such integration could take the form of either enhancement or suppression, although enhanced responses were more common. These enhanced responses were biased toward one call type: the grunt.

It is important to note that, for a given cortical site, not all exemplars within a call category could elicit enhancement or suppression. For example, grunt A may elicit enhancement, whereas grunt B may not show any integration at all. The reason for the lack of category specificity is likely attributable to the diverse features of the individual exemplars. Within a call category, each exemplar is produced by a different individual. As a result, the facial component (e.g., head movement, mouth position, etc.) and vocal component (e.g., the spectral structure, duration, etc.) are not identical across the different exemplars within a call category. It is, therefore, not surprising that we observe a diversity of responses at a given cortical site. Our results, however, indicate one important component of variability between all of the call exemplars that may lead to predictable neural responses: the voice-onset time. There is a broad range of voice-onset times within a call category, and onset time seems to influence the probability (but not the magnitude) of seeing an enhanced or a suppressed multisensory response independent of the call type (Fig. 5). This suggests that the temporal relationship between the face and voice may be one predictor of the type of multisensory response observed.

Previous neuroimaging studies of multimodal speech suggested that suppression is especially prominent when the speech tokens from the two modalities are incongruent in identity (Calvert et al., 2000). The present data, along with recent human neuroimaging data (Wright et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005), suggest that identity incongruence is not a requirement for response suppression. Recently, two human evoked-potential studies have reported that face plus voice integration is represented only by suppressed auditory N100 responses (Besle et al., 2004; van Wassenhove et al., 2005). This is not supported by our LFP data, in which we found both suppression and enhancement (in fact, more frequently enhancement) to our congruent face plus voice stimuli relative to voice alone. We suggest that the consistently suppressed responses in the N100 component in these human studies are attributable to the very long time interval between the presentation of the face and the voice signal (Besle et al., 2004; van Wassenhove et al., 2005). In both studies, the time between the appearance of the face and the onset of the auditory signal typically exceeded 500 ms. In our data, enhanced responses were primarily seen when this time interval was <100 ms, and suppressed responses were primarily seen at intervals >200 ms.

Within the domain of enhanced responses, we found that grunt vocalizations were overrepresented relative to coos. Because their voice-onset times overlapped and because the frequency of their suppressed responses was indistinguishable, we suggest that this response difference likely reflects a behaviorally relevant distinction. Coos and grunts are both affiliative vocalizations produced in a variety of contexts (Hauser et al., 1993; Partan, 2002). They differ, however, in their direction of expression and range of communication. Coos are generally contact calls rarely directed toward any particular individual. In contrast, grunts are often directed toward individuals in one-on-one situations, often during social approaches, such as in baboons and vervet monkeys (Cheney and Seyfarth, 1982; Palombit et al., 1999). Given their production at close range and context, grunts may produce a stronger face/voice association than coo calls. This putative distinction appeared to be reflected in the pattern of multisensory responses across the two regions of the auditory cortex.

Along with Calvert (2001), we speculate that a major source of visual input into the auditory cortex may be the upper bank of the STS (uSTS). Our control experiments revealed that the majority of multisensory responses to faces and voices were specific to faces and voices; that is, they did not also integrate to the combination of our control videos and the voice signal. Given the presence of face-selective neurons in the uSTS (Harries and Perrett, 1991) and its direct connections with the superior temporal plane (which includes the core and lateral belt regions of the auditory cortex) (Seltzer and Pandya, 1994), the uSTS provides a likely source of face inputs into the auditory cortex. The “feedback” hypothesis is also supported by a current-source density study that demonstrated that the auditory belt area receives visual input in the supragranular and infragranular layers, a pattern consistent with feedback connectivity (Schroeder and Foxe, 2002). Given that there is a greater frequency of multisensory integration to faces/voices in the lateral belt than in the more medial core region, we predict that the pattern of uSTS inputs into the superior temporal plane should taper off in the lateral-to-medial direction. Although there are likely to be many other behaviorally relevant bimodal events represented in the auditory cortex, the present results suggest that the human auditory cortex and monkey auditory cortex play homologous roles in processing bimodal vocal signals.

We found that although the vast majority of cortical sites showed multisensory integration in the LFP signal, a much smaller proportion of sites showed such integration in the spiking activity (the analog MUA, in this case). Even in well established polysensory cortical areas, such as the superior temporal sulcus, only 23% of visually responsive single neurons are significantly influenced by auditory stimuli (Barraclough et al., 2005). Thus, an investigation solely focused on spiking activity will have likely missed many of the effects reported here. This discrepancy between LFPs and spiking activity is not surprising; it has long been known that the electroencephalogram signals and unit activity do not always correspond with each other (Bullock, 1997). Recently, for example, Schroeder and colleagues (Schroeder et al., 2001; Fu et al., 2004) demonstrated robust eye-movement and somatosensory-related field potential activity in the auditory cortex that did not necessarily have a counterpart in the analog MUA signals. Thus, establishing the relationship between unit activity and LFPs will be particularly important in revealing the cortical mechanisms of multisensory integration as it has in other domains of sensory and motor physiology (Bullock, 1997; Pesaran et al., 2002; Mehring et al., 2003; Henrie and Shapley, 2005).

References

Abry C, Lallouache M-T, Cathiard M-A (1996) How can coarticulation models account for speech sensitivity in audio-visual desynchronization? In: Speechreading by humans and machines: models, systems and applications (Stork D, Henneke M, eds), pp 247–255. Berlin: Springer.

Adolphs R (2003) The cognitive neuroscience of human social behaviour. Nat Rev Neurosci 4:165–178.

Barbour DL, Wang X (2003) Contrast tuning in auditory cortex. Science 299:1073–1075.

Barraclough NE, Xiao D, Baker CI, Oram MW, Perrett DI (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J Cogn Neurosci 17:377–391.

Beauchamp MS, Lee KE, Argall BD, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823.

Bell AH, Corneil BD, Munoz DP, Meredith MA (2003) Engagement of visual fixation suppresses sensory responsiveness and multisensory integration in the primate superior colliculus. Eur J Neurosci 18:2867–2873.

Bernstein LE, Auer ET, Moore JK, Ponton CW, Don M, Singh M (2002) Visual speech perception without primary auditory cortex activation. NeuroReport 13:311–315.

Bernstein LE, Auer ET, Takayanagi S (2004) Auditory speech detection in noise enhanced by lipreading. Speech Comm 44:5–18.

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234.

Bullock TH (1997) Signals and signs in the nervous system: the dynamic anatomy of electrical activity is probably information-rich. Proc Natl Acad Sci USA 94:1–6.

Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E (2003) Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport 14:2213–2218.

Calvert GA (2001) Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cereb Cortex 11:1110–1123.

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999) Response amplification in sensory-specific cortices during crossmodal binding. NeuroReport 10:2619–2623.

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649–657.

Cheney DL, Seyfarth RM (1982) How vervet monkeys perceive their grunts: field playback experiments. Anim Behav 30:739–751.

Frens MA, Van Opstal AJ (1998) Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. Brain Res Bull 46:211–224.

Fu K-MG, Shah AS, O’Connell MN, McGinnis T, Eckholdt H, Lakatos P, Smiley J, Schroeder CE (2004) Timing and laminar profile of eye-position effects on auditory responses in primate auditory cortex. J Neurophysiol 92:3522–3531.

Gail A, Brinksmeyer HJ, Eckhorn R (2004) Perception-related modulations of local field potential power and coherence in primary visual cortex of awake monkey during binocular rivalry. Cereb Cortex 14:300–313.

Ghazanfar AA, Logothetis NK (2003) Facial expressions linked to monkey calls. Nature 423:937–938.

Ghazanfar AA, Santos LR (2004) Primate brains in the wild: the neural bases for social interactions. Nat Rev Neurosci 5:603–616.

Hackett TA (2002) The comparative anatomy of the primate auditory cortex. In: Primate audition: ethology and neurobiology (Ghazanfar AA, ed), pp 199–226. Boca Raton, FL: CRC.

Harries MH, Perrett DI (1991) Visual processing of faces in temporal cortex—physiological evidence for a modular organization and possible anatomical correlates. J Cogn Neurosci 3:9–24.

Hauser MD, Marler P (1993) Food-associated calls in rhesus macaques (*Macaca mulatta*). I. Socioecological factors. Behav Ecol 4:194–205.

Hauser MD, Evans CS, Marler P (1993) The role of articulation in the production of rhesus monkey, *Macaca mulatta*, vocalizations. Anim Behav 45:423–433.

Henrie JA, Shapley R (2005) LFP power spectra in V1 cortex: the graded effect of stimulus contrast. J Neurophysiol, in press.

Izumi A, Kojima S (2004) Matching vocalizations to vocalizing faces in a chimpanzee (*Pan troglodytes*). Anim Cogn 7:179–184.

Kayser C, Kim M, Ugurbil K, Kim D-S, Konig P (2004) A comparison of hemodynamic and neural responses in cat visual cortex using complex stimuli. Cereb Cortex 14:881–891.

Lauritzen M (2001) Relationship of spikes, synaptic activity, and local changes of cerebral blood flow. J Cereb Blood Flow Metab 21:1367–1383.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150–157.

Logothetis NK, Merkle H, Augath M, Trinath T, Ugurbil K (2002) Ultra high-resolution fMRI in monkeys with implanted RF coils. Neuron 35:227–242.

Massaro DW (1998) Perceiving talking faces: from speech perception to a behavioral principle. Cambridge, MA: MIT.

Mathiesen C, Cesar K, Algoren N, Lauritzen M (1998) Modification of activity-dependent increases of cerebral blood flow by excitatory synaptic activity and spikes in rat cerebellar cortex. J Physiol (Lond) 512:555–566.

Mehring C, Rickert J, Vaadia E, Cardoso de Oliveira S, Aertsen A, Rotter S (2003) Inference of hand movements from local field potentials in monkey motor cortex. Nat Neurosci 6:1253–1254.

Meredith MA, Stein BE (1986) Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J Neurophysiol 56:640–662.

Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J Neurosci 7:3215–3229.

Norena A, Eggermont JJ (2002) Comparison between local field potentials and unit cluster activity in primary auditory cortex and anterior auditory field in the cat. Hear Res 166:202–213.

Olson IR, Gatenby JC, Gore JC (2002) A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Res Cogn Brain Res 14:129–138.

Palombit RA, Cheney DL, Seyfarth RM (1999) Male grunts as mediators of social interaction with females in wild chacma baboons (*Papio cynocephalus ursinus*). Behaviour 136:221–242.

Partan SR (2002) Single and multichannel signal composition: facial expressions and vocalizations of rhesus macaques (*Macaca mulatta*). Behaviour 139:993–1027.

Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen RA (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat Neurosci 5:805–811.

Pfingst BE, O'Connor TA (1980) Vertical stereotaxic approach to auditory-cortex in the unanesthetized monkey. J Neurosci Methods 2:33–45.

Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory-cortex. Science 268:111–114.

Recanzone GH, Guard DC, Phan ML (2000) Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. J Neurophysiol 83:2315–2331.

Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu S-T, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127:141–145.

Schroeder CE, Foxe JJ (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Brain Res Cogn Brain Res 14:187–198.

Schroeder CE, Lindsey RW, Specht C, Marcovici A, Smiley JF, Javitt DC (2001) Somatosensory input to auditory association cortex in the macaque monkey. J Neurophysiol 85:1322–1327.

Seltzer B, Pandya DN (1994) Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol 343:445–463.

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge, MA: MIT.

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 102:1181–1186.

Wallace MT, Wilkinson LK, Stein BE (1996) Representation and integration of multiple sensory inputs in primate superior colliculus. J Neurophysiol 76:1246–1266.

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb Cortex 13:1034–1043.