J Neurophysiol 101: 773–788, 2009.
First published November 26, 2008; doi:10.1152/jn.90843.2008.

Different Neural Frequency Bands Integrate Faces and Voices Differently in the Superior Temporal Sulcus

Chandramouli Chandrasekaran and Asif A. Ghazanfar
Neuroscience Institute and Department of Psychology, Princeton University, Princeton, New Jersey

Submitted 1 August 2008; accepted in final form 20 November 2008

Chandrasekaran C, Ghazanfar AA. Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *J Neurophysiol* 101: 773–788, 2009. First published November 26, 2008; doi:10.1152/jn.90843.2008. The integration of auditory and visual information is required for the default mode of speech–face-to-face communication. As revealed by functional magnetic resonance imaging and electrophysiological studies, the regions in and around the superior temporal sulcus (STS) are implicated in this process. To provide greater insights into the network-level dynamics of the STS during audiovisual integration, we used a macaque model system to analyze the different frequency bands of local field potential (LFP) responses to the auditory and visual components of vocalizations. These vocalizations (like human speech) have a natural time delay between the onset of visible mouth movements and the onset of the voice (the "time-to-voice" or TTV). We show that the LFP responses to faces and voices elicit distinct bands of activity in the theta (4–8 Hz), alpha (8–14 Hz), and gamma (>40 Hz) frequency ranges. Along with single neuron responses, the gamma band activity was greater for face stimuli than voice stimuli. Surprisingly, the opposite was true for the low-frequency bands—auditory responses were of a greater magnitude. Furthermore, gamma band responses in STS were sustained for dynamic faces but not so for voices (the opposite is true for auditory cortex). These data suggest that visual and auditory stimuli are processed in fundamentally different ways in the STS. Finally, we show that the three bands integrate faces and voices differently: theta band activity showed weak multisensory behavior regardless of TTV, the alpha band activity was enhanced for calls with short TTVs but showed little integration for longer TTVs, and finally, the gamma band activity was consistently enhanced for all TTVs. These data demonstrate that LFP activity from the STS can be segregated into distinct frequency bands which integrate audiovisual communication signals in an independent manner. These different bands may reflect different spatial scales of network processing during face-to-face communication.

INTRODUCTION

Neuroimaging studies in humans suggest that cortical regions in and around the superior temporal sulcus (STS) are involved in the integration of faces and voices and numerous other classes of audiovisual signals (Beauchamp et al. 2004; Calvert 2001; Calvert et al. 2000; Ethofer et al. 2006; Kreifelts et al. 2007; Noesselt et al. 2007; van Atteveldt et al. 2004; Wright et al. 2003). Similar studies using electro- and magnetoencephalography (EEG and MEG) suggest that evoked responses, localized approximately to the superior temporal gyrus, show suppressive responses for multisensory stimuli compared with auditory stimuli (Besle et al. 2004; Klucharev et al. 2003; van Wassenhove et al. 2005). Single-unit studies in monkeys have also identified the STS as a nexus for inputs from different sensory modalities (Baylis et al. 1987; Bruce et al. 1981; Hikosaka et al. 1988; Schroeder and Foxe 2002), and two studies examined how different sensory signals are integrated by single neurons in this region (Barraclough et al. 2005; Benevento et al. 1977). There is, however, a large epistemic void between integrative processes that occur at the single neuron level with those occurring at the level measured by MEG/EEG/functional magnetic resonance imaging (fMRI) methods. The local field potential (LFP) represents activity at an intermediate spatial scale that can provide a scaffold between these two extremes (Varela et al. 2001).

LFPs predominantly reflect the input processes in a given cortical region (Logothetis 2003). This signal can be decomposed into discrete frequency bands (e.g., theta, alpha, or gamma band activity), each of which represents the synchronous activity of an oscillating network. These oscillating networks represent the "middle ground" linking the spiking activity of single neurons to behavior (Buzsaki and Draguhn 2004). Only recently, however, have studies identified the possible dynamics of these different frequency bands, their sources, and their possible functions (Belitski et al. 2008; Henrie and Shapley 2005; Kayser and Konig 2004; Lakatos et al. 2005; Liu and Newsome 2006; Pesaran et al. 2002). With regard to multisensory integration, the role(s) of different neural frequency bands is unknown (Senkowski et al. 2008).

Using a monkey model system, here we investigated whether discrete neural frequency bands integrated conspecific faces and voices differently by recording LFP activity in the upper bank of the STS. Vocal communication in macaque monkeys shows several parallels with human speech reading. These monkeys can match faces to voices based on expression type and indexical cues (Ghazanfar and Logothetis 2003; Ghazanfar et al. 2007), segregate competing multisensory vocal gestures (Jordan et al. 2005), and use similar eye-movement strategies as humans when viewing vocalizing faces (Ghazanfar et al. 2006). Furthermore, such behaviors are mediated by neural circuits that are similar to those activated by audiovisual speech in the human brain. Single neurons in the monkey STS integrate audiovisual biological motion, including vocalizations (Barraclough et al. 2005), as do neurons in the ventrolateral prefrontal cortex (Sugihara et al. 2006). Auditory cortex also integrates faces and voices (Ghazanfar et al. 2005), and this integration is mediated at least in part by interactions with the STS (Ghazanfar et al. 2008).

Address for reprint requests and other correspondence: A. A. Ghazanfar, Neuroscience Institute and Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08540 (E-mail: asifg@princeton.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

 0022-3077/09 $8.00 Copyright © 2009 The American Physiological Society

Downloaded from jn.physiology.org on February 6, 2009

In the current study, our data show that both faces and voices elicit responses in discrete frequency bands, including theta (4–8 Hz), alpha (8–14 Hz), and gamma band (>40 Hz), in the STS. By exploiting the fact that, just as in human speech, there is a natural and variable delay between the timing of the visible mouth movement and the onset of the voice component in monkey vocalizations (time-to-voice or TTV) (Ghazanfar et al. 2005), we show that these three different frequency bands integrate faces and voices differently. By *integration*, we mean that power in a given neural frequency band is significantly enhanced or suppressed in face+voice conditions relative to face- and voice-alone conditions. Theta band activity shows weak multisensory behavior and no dependence on the TTV, alpha band activity shows enhancement for short TTVs, but no consistent effects for longer TTVs, and finally the gamma band shows only enhancement—it is independent of the TTV. We interpret these results within a framework that suggests that the integration observed in these frequency bands reflect the different spatiotemporal scales of multiple networks each involving the STS.

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 superior temporal sulcus and to model a three-dimensional (3-D) skull reconstruction. From these skull models, we constructed custom-designed, form-fitting titanium headposts and recording chambers (see Logothetis et al. 2002 for details). 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 and sulcus (Pfingst and O'Connor 1980). All experiments were performed in compliance with the guidelines of the local authorities (Regierungspraesidium) and the European Community (EU VD 86/609/EEC) for the care and use of laboratory animals.

Stimuli

The 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 while seated in a primate 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. The vocalizations were coos and grunts. Videos were acquired at 30 frames/s (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 from the beginning of the first mouth movement to the mouth closure at the end of vocalization. The duration of the video clips and auditory onset relative to the initial mouth movement—the TTV—varied according to the vocalization (Fig. 1).

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 chair in front of a 21-in 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, and 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 the appearance of a video sequence with the audio track, the appearance of the video alone (no audio), or 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 conditions with a video component, the monkeys were required to restrict their eye movements to within the video frame for the duration of the video (Ghazanfar et al. 2005, 2008; Sugihara et al. 2006). Successful completion of a trial resulted in a juice reward. Eye position signals were digitized at a sampling rate of 200 Hz. Ten trials were presented for each condition: face + voice, face alone, and voice alone for each vocalization.

It is possible that restricting the subjects' eye movements could influence the neural activity in an abnormal fashion. We think that it probably does not for two reasons. First, natural eye movements (that is, unconstrained by any task demands) directed at audiovisual communication signals reveal stereotypical fixation patterns restricted to the eyes and the mouth. This is true for both monkeys and human observers (Ghazanfar et al. 2006; Vatikiotis-Bateson et al. 1998). Second, the video clips were so short that only very few (maximum: 3) eye movements could be made within a trial. It should also be noted that if we required strict central fixation, this may have led to a

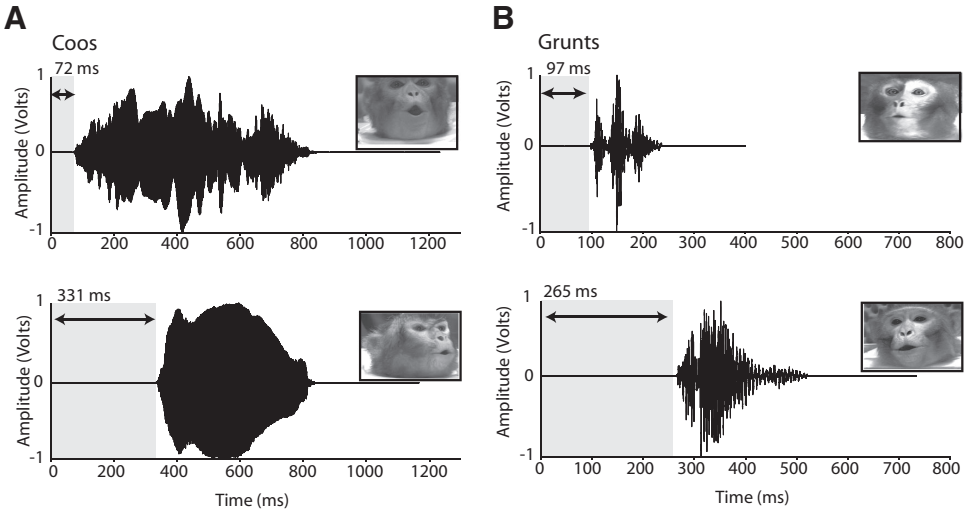

FIG. 1. Exemplars of the vocalization stimuli used in the study. *A*: example of 2 coo calls used in the study. *Top*: waveform of a coo call with a short time-to-voice (72 ms); *inset*: the frame with the maximum mouth opening. Gray shading, the period before the onset of the sound during which only the visual signal is present. *Bottom*: example of a coo call with a long time-to-voice (331 ms). *x* axes depict time in milliseconds. *y* axes depict amplitude of the signal in volts. *B*: example of 2 grunt calls used in the study. *Top*: waveform of a grunt call with a short time-to-voice (97 ms); *inset*: the frame with the maximum mouth opening. The gray shading denotes the period before the onset of the sound during which only the visual signal is present. *Bottom*: example of a grunt call with a long time-to-voice (265 ms). Figure conventions as in *A*.

Downloaded from jn.physiology.org on February 6, 2009

suppression of integrative responses (Bell et al. 2003) as well as to a suppression or elimination of auditory responses (Gifford and Cohen 2004).

Data collection

Recordings were made from the upper bank of the STS. We employed a custom-made electrode drive that allowed us to move multiple electrodes independently. 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 (1–5,000 Hz), and acquired at 20.2-kHz sampling rate. Electrodes were lowered first into the auditory cortex until multiunit cortical responses could be driven by auditory stimuli. Search stimuli included pure tones, FM sweeps, noise bursts, clicks, and vocalizations. Using the analog multiunit signal (MUA; 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). Initially, in both monkeys, we discerned a coarse tonotopic map representing high-to-low frequencies in the caudal-to-rostral direction (Hackett et al. 1998). Such a map is identified as primary auditory cortex (A1) and gives an indication of the anterior-posterior location of the STS region (which lies just below auditory cortex) we recorded from. Thus upon the identification of primary auditory cortex, locating the upper bank of the STS was straightforward—it was the next section of gray matter below the superior temporal plane. Electrodes would be lowered until auditory cortical activity ceased, followed by a short silent period representing the intervening white matter. The cortical activity following this silent period arises from the upper bank of the STS. Its visual responses were tested with faces and a variety of visual motion stimuli (Bruce et al. 1981). Given the identification of primary auditory cortex in the superior temporal plane in every recording session (Ghazanfar et al. 2005, 2008) and the very slow, careful advancement of electrodes subsequently, the most likely location of our STS recordings was the TPO region of the upper bank. This is supported by the response properties of single neurons recorded in this region (see RESULTS). We recorded activity from 36 cortical sites over 15 different sessions. A maximum of four electrodes were lowered into the STS in a given session; the interelectrode distance was never <3 mm.

Data processing and analyses

LFPs (the low-frequency range of the mean extracellular field potential) were extracted off-line by band-pass filtering the signal between 1 and 300 Hz using a 4-pole, bidirectional Butterworth filter. LFPs were examined to ensure that the signal was not contaminated by 50-Hz line noise or other ambient noise. Basic response properties to each stimulus condition (face+voice, face alone and voice alone) were assessed following either band-pass filtering in the relevant frequency ranges bands or with spectral analyses. Data from both monkeys were largely similar and therefore pooled together.

Spectral analyses

All the spectral analyses were based on wavelet spectra using modified scripts based on the Chronux suite of Matlab routines (www.chronux.org) and Matlab scripts provided to us courtesy of Daeyeol Lee (Lee 2002; see also Ghazanfar et al. 2008 for details).

Band-pass filters

We used a two-way, least-squares finite impulse response (FIR) filter to band-pass filter the signal in the different frequency bands. We then applied a Hilbert transform to the band-pass signal to obtain an analytic representation of the signal. The absolute value of this complex number provides an estimate of the envelope. We averaged this envelope across trials to estimate the mean envelope and the SD.

Baseline correction

For the wavelet spectrograms, we estimated the baseline activity as the mean signal in the −300 to −200-ms range of the wavelet spectrogram across frequencies. We divided the signal during the stimulus period in each time frequency bin by this baseline activity. Values equal to 1 indicate that the stimulus activity is the same as the baseline activity. Values >1 indicate enhancement, and <1 indicate suppression.

For the Hilbert transformed band-pass signals, we again divided the signal by the mean activity in the – 300 to – 200-ms region of the baseline period. This region is far away from the onset of the stimulus and from the transients introduced by sharp filtering at the edges. We then subtracted "1" from the result to set all the baseline values around 0. This manipulation is the same as expressing the signal as a percent change from baseline.

Statistical inference

To analyze the LFP signals, we adapted an approach developed for the analyses of spiking data. The LFP signal in different frequency bands was averaged in 15-ms bins, and *t*-tests were performed to identify whether there were significant differences between the face+voice and the unisensory conditions (face alone and voice alone). In the case of a single cortical site, *t*-tests were performed with 10 trials for each condition. For the population, *t*-tests were performed with the mean estimates from the 36 cortical sites. Gray shading in Figs. 7 and 8 denote regions of significant difference between the face+voice condition and the two-unimodal conditions across the population of electrodes.

Multisensory index

To identify the relative enhancement of the face+voice response relative to the face- and voice alone conditions, we computed a multisensory index as the percent difference between the response to the multisensory condition and the unisensory conditions.

RESULTS

We recorded LFP activity from 36 cortical sites in the upper bank of the STS from two monkeys while the subjects viewed and/or heard conspecifics producing two classes of vocalizations: coos and grunts. These vocalizations possess unique auditory and visual components. Coos are longer in duration and are typically produced with protruded lips, whereas grunts are harsh noisy calls produced with a more subtle mouth opening and no lip protrusion and are generally much shorter in duration. We picked these two call types for the following reasons. First, they are both affiliative calls that are produced in largely the same, but multiple contexts (e.g., as a greeting toward other conspecifics, the anticipation of food, etc). Thus no distinct "meaning" can be associated with either of them except for "friendliness" broadly defined. Second, although they sound distinct, both are broadband calls with largely overlapping spectral profiles. Third, and finally, both coos and grunts are produced frequently (more so than other calls in the rhesus monkey repertoire), and thus both our subjects had extensive experience with them. We used a set of four coos and four grunts produced by different individuals to provide us with a set of communication signals with natural signal variations. One such variation is the TTV. As in human speech, the

Downloaded from jn.physiology.org on February 6, 2009

mouth begins to move before the auditory component is voiced by vocal fold activity. This TTV delay is variable across different call exemplars (Fig. 1). In our stimulus set, the eight calls provided a range of short and long TTVs: 66, 72, 75, 85, and 97 ms for short TTVs and 219, 265, and 331 ms for long TTVs. This natural variation allowed us to study the effect of temporal disparities in congruent social signals on multisensory integration in the STS. Two coos and three grunts comprised the short TTV call category, and two coos and one grunt the long TTV call category.

To confirm that we are in fact recording in the upper bank of the STS, we examined the response properties of a set of single neurons. We recorded from 61 single neurons in the upper bank of the STS. Figure 2 shows four representative single units responding to the different calls. Consistent with previous reports, these neurons are very sensitive to visual stimuli, respond weakly to auditory stimuli, and show integration effects when the two modalities are presented together (Barraclough et al. 2005; Baylis et al. 1987; Benevento et al. 1977; Bruce et al. 1981; Hikosaka et al. 1988). Figure 2, *A* and *B,* shows examples of multisensory enhancement and suppression, respectively. The neuron in Fig. 2*A* responded at multiple points to the dynamic face and in addition showed enhanced responses to the face+voice stimulus compared with the face-alone stimulus. Specifically, at 450 ms (see ↓) after onset of the face, a robust visual response is observed relative to baseline [$t(18) = 2.18$, $P = 0.04$]. In addition, an enhanced multisensory response is observed relative to baseline [$t(18) = 2.51$, $P = 0.021$]. Despite the integration [$F(2, 29) = 4.41$, $P = 0.02$, post hoc tests $P < 0.05$ for both face+voice vs. face and face+voice vs. voice], the neuron did not respond to the voice alone stimulus [$t(18) = 0.595$, $P = 0.54$]. Figure 2*B* shows another neuron that displayed a suppressed multisensory response. At ~350 ms after visual onset, visual [$t(18) = 2.42$, $P = 0.026$] and multisensory [$t(18) = 2.09$, $P = 0.05$] responses were enhanced relative to baseline, and the auditory response is suppressed relative to baseline [$t(18) = -4.19$, $P \ll 0.0001$]. This unit also integrated faces and voices [$F(2, 29) = 5.46$, $P = 0.01$, post hoc tests $P < 0.05$ for both face+voice vs. face and face+voice vs. voice].

The neurons in Fig. 2, *C* and *D,* show a different pattern: multisensory enhancement as well as robust responses to both unimodal visual and auditory stimuli. The neuron in Fig. 2*C* responded robustly to the face [$t(18) = 2.73$, $P = 0.013$], face+voice [$t(18) = 2.79$, $P = 0.01$] and a robust auditory off response [$t(18) = 2.19$, $P = 0.041$] to the voice at 600 ms after visual onset. Multisensory responses were observed for this neuron at ~400 ms after onset of the visual signal [$F(2,29) = 6.99$, $P = 0.003$, post hoc tests $P < 0.05$ for both face+voice vs. face and face+voice vs. voice]. Figure 2*D* shows another example of an STS neuron responsive to both visual and auditory stimuli. At ~175 ms after the onset of the visual signal, responses to both visual [$t(18) = 2.25$, $P = 0.03$] and auditory stimuli [$t(18) = 2.11$, $P = 0.049$] are enhanced above baseline. The multisensory response was also enhanced relative to the unimodal responses [$F(2, 29) = 5.39$, $P = 0.01$, post hoc tests $P < 0.05$ for both face+voice vs. face and face+voice vs. voice].

Although it was not our intention to fully characterize the integrative properties of single STS neurons (see Barraclough et al. 2005), we shall briefly describe the response properties for the 61 neurons we isolated. Fourteen of our 61 neurons (23%) were responsive to faces, voices, or their combination.

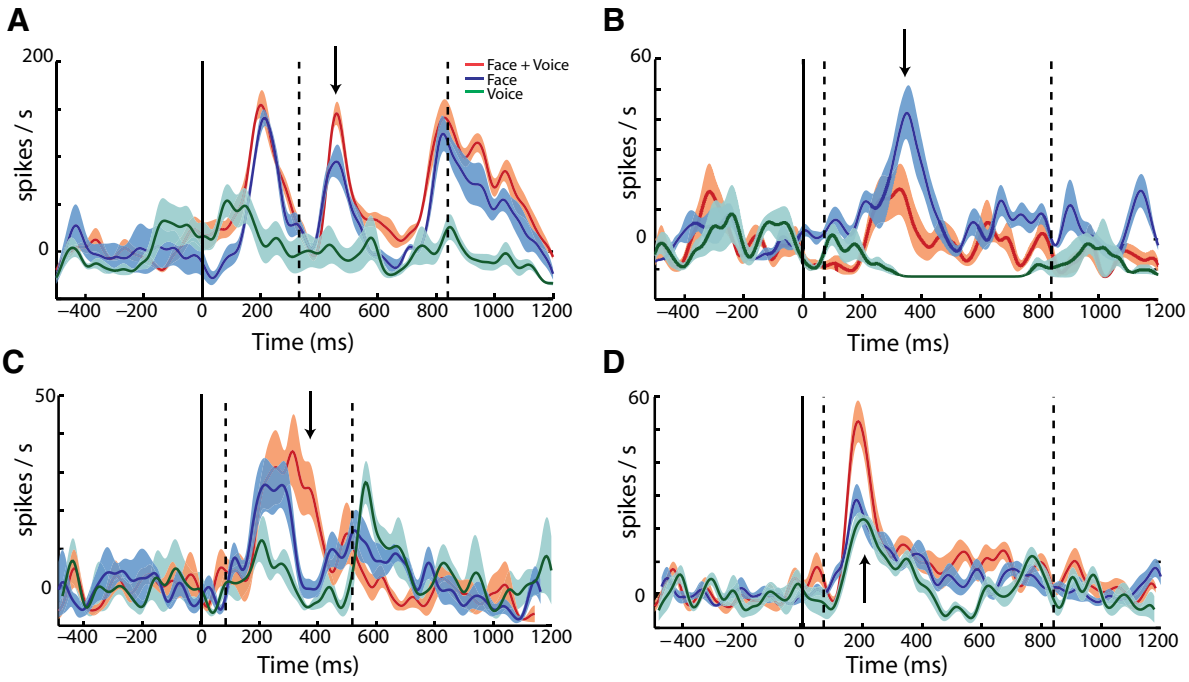

FIG. 2. Responses of single neurons to coo calls in the superior temporal sulcus. *A*: peristimulus time histogram (PSTH) showing that the response to the face+voice is significantly enhanced relative to the face alone response after the onset of the voice signal (- - -, on- and off-set of voice). Note the lack of an auditory response. x axes depict time in milliseconds, *y* axes depict the firing rate of the neuron in spikes per second. Shaded regions denote the SE. ↓, region where a significant multisensory response relative to both unisensory responses was observed. *B*: another single neuron showing multisensory suppression. Note the lack of an auditory response and the suppression of the multisensory response relative to the visual response alone. Figure conventions as in *A*. *C*: single neuron showing both unimodal auditory and visual responses as well as a multisensory enhancement when the 2 are combined. Figure conventions as in *A*. *D*: another single neuron showing almost equivalent auditory and visual responses and a strong multisensory enhancement effect. Figure conventions as in *A*.

Downloaded from jn.physiology.org on February 6, 2009

Of these 14 neurons, 8 (57%) displayed some form of integration as defined by a significant difference between the multisensory response and both unisensory responses. Seven of the 14 neurons (50%) responded to visual stimulation alone, 4 of these 7 (57%) showed a multisensory response. One of the 14 responsive neurons was sensitive to auditory alone stimulation (7%); this neuron also showed multisensory integration. Finally, 6 of the 14 neurons (43%) responded to both auditory and visual stimulation; of this subset, 3 of 6 neurons (50%) possessed a multisensory response. These numbers concur well with the more extensive reports of properties of single units in the upper bank of the superior temporal sulcus (Barraclough et al. 2005; Baylis et al. 1987; Benevento et al. 1977; Bruce et al. 1981; Hikosaka et al. 1988). Overall, the single-unit data give considerable credence to the notion that we are recording from the same region as these previous reports.

Dynamic faces and voices elicit complex spectrotemporal responses in the STS

We used wavelet-based spectral analyses and band-pass filters to decompose the LFP responses to faces and voices into different time-frequency components. Figure 3*A* shows the normalized spectrogram of the LFP response to the face- and voice-alone components of a single coo call aligned to the visual onset, averaged over 10 trials, for a single site in STS (*left*) as well as the population of 36 cortical sites (*right*). Colors denote the strength of the response within a given frequency band. These spectrograms reveal that faces and voices representing the coo expression elicit distinct low frequency activity at the onset of the response, as well as activity in a high-frequency band. This high-frequency activity is punctate for voices, but sustained for faces (see also Figs. 4 and 5). Several other features are apparent in these spectrograms. Theta (4–8 Hz)- and alpha (8–14 Hz)-band activity are suppressed below baseline at ~250 ms after the onset of the face, and this suppression persists throughout the remaining duration of the dynamic facial expression. In contrast, the gamma band activity seems to be robust for the entire duration of the moving face (see following text). For both faces and voices, average power spectra of the LFP response from the 0- to 200-ms poststimulus time interval reveal the distinct nature of these frequency bands. We separated out responses to faces and voices into the low- and high-frequency ranges to better illustrate these bands. Figure 3*B, left,* shows two clear peaks, one in the theta range and the other in the alpha range, whereas the *right panel* shows that the high-frequency power is a gamma band response (>40 Hz). A nearly identical pattern of results is seen for a grunt call (Fig. 3, *C* and *D*).

Note that although an even lower frequency band is readily apparent (1–4 Hz, the delta band) in Fig. 3, *A* and *C,* the long period of this frequency range combined with our short trial times limits our analytical resolution and therefore precludes further descriptions of this frequency band. Interestingly, while the overall shape of the spectra was the same for responses elicited by both faces and voices, the responses to voices were larger in the theta and alpha bands relative to dynamic face responses [theta band: face vs. voice, $t(574) = -10.68$, $P <<< 0.001$; alpha band: face vs. voice, $t(574) = -10.388$, $P <<< 0.001$]. This was surprising given that the STS has long been considered a predominantly visual area.

Gamma band responses to dynamic faces are sustained for the duration

Analysis of the spectral structure of STS LFP activity suggested that gamma band responses to visual stimuli are long and sustained (Fig. 3, *A* and *C, top*). To better characterize these responses, we analyzed the relationship between the duration of the visual signal and the duration of the gamma band responses. Figure 4*A* shows the baseline-normalized population-averaged ($n = 36$ cortical sites) gamma band response aligned to visual onset to the face component of a single coo (*top*) and a single grunt vocalization (*bottom*). The color bar denotes the normalized power. Figure 4*B* shows the Hilbert transformed gamma band activity (60–95 Hz) across all electrodes for the two calls shown in Fig. 4*A*. The gamma band responses to dynamic faces are elevated above the baseline throughout the duration of the facial expression, consistent with the visual sensitivity of this region. Figure 4*C* summarizes the near monotonic relationship between the duration of the gamma band activity and the duration of the facial expression.

The gamma band response to faces is more robust and sustained relative to voice responses. It is possible task variables could explain this difference. In the task, the monkey fixates within the video frame during the face-alone condition but has no such constraints in the voice-alone condition for two reasons: a visual stimulus used to behaviorally constrain (such as a frame or a fixation point) the monkeys' eye movements during the voice alone could "integrate" with the voice over a number of trials and strict and arbitrary fixation (in contrast to scanning a face) suppresses multisensory (Bell et al. 2003) and auditory responses (Gifford and Cohen 2004). We don't believe that this task variable is driving these differences, but contend that the auditory signals in a predominantly visual area are processed in a fundamentally different way than visual signals. Conversely, auditory signals in an auditory area are processed differently than visual signals. To illustrate this, we compared gamma band activity recorded in the STS with auditory cortical activity (the middle lateral belt region) recorded in a previous study under identical conditions and with the same stimuli (Ghazanfar et al. 2005). Figure 5 shows the gamma band activity from these two regions in response to faces and voices. For the face-alone stimuli, gamma band activity is sustained and greater in magnitude in the STS (Fig. 5, *A* and *B*) but almost nonexistent in response to voice-alone stimuli. In contrast, gamma band activity in auditory cortex is sustained and greater in magnitude for voice-alone stimuli but relatively weak for face-alone stimuli (Fig. 5, *C* and *D*). These differential patterns of activity suggest that the nature of gamma band activity is modality- and cortical region-dependent and not necessarily task-dependent (though, of course, we cannot eliminate that possibility).

Theta band activity shows weak multisensory effects and no clear dependence on the TTV

In all subsequent analyses, we defined multisensory integration as those face+voice responses that were significantly different from the strongest unimodal responses (Meredith et al. 1987; Wallace et al. 2004). In the theta band, the voice condition elicited the strongest unimodal responses. Figure 6*A* shows the baseline-corrected responses in the theta band across

Downloaded from jn.physiology.org on February 6, 2009

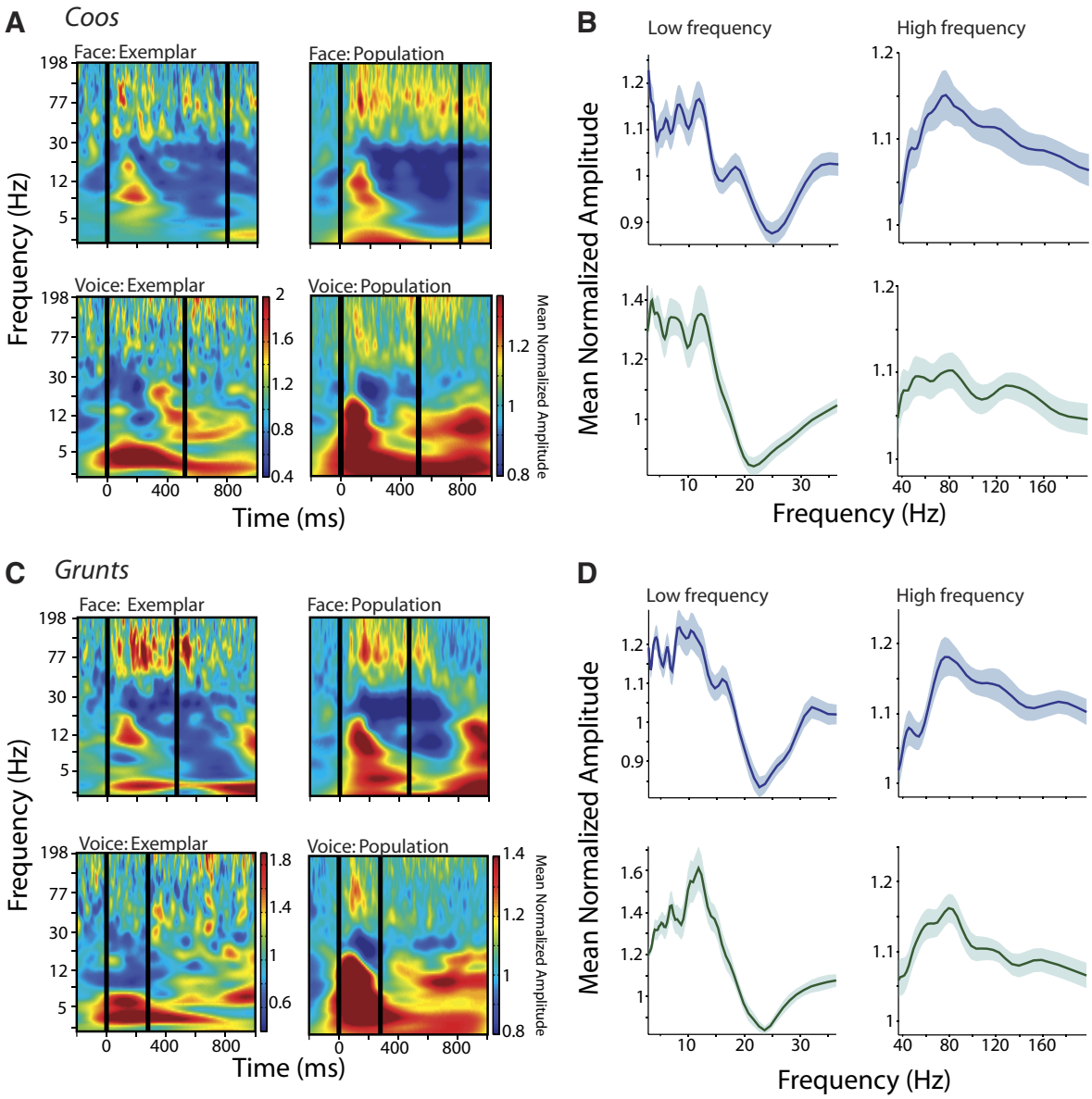

FIG. 3. Time-frequency representation of the local field potential responses in the superior temporal sulcus (STS) to the visual and auditory components of coo and grunt vocalizations. *A*: spectrograms show the average spectra for a single cortical site (*left*) in response to face (*top*) and voice (*bottom*) components of a coo call. Spectra are averages of 10 trials for the call and aligned to the onset of the visual signal for faces and to the onset of the auditory signal for voices. *Right*: the population spectrograms (n = 36 cortical sites) for the face (*top*) and voice (*bottom*) components of the same coo call. *x* axes depict the time in milliseconds. Black lines denote the on- and off-set of the face or voice signal. *y* axes depict the frequency of the oscillations in hertz. Color bar indicates the amplitude of these signals normalized by the baseline mean. *B*: population spectra as a function of frequency for the low-frequency (1–37 Hz) responses shown in *A* for the face (*top*) and voice (*bottom*) components (*left*). Population spectra for the high-frequency (37–200 Hz) responses are shown on the *right*. *x* axes depict frequency in hertz. *y* axes depict the average normalized spectrum from 0 to 200 ms after onset of the respective signal. Shaded regions denote the SE. *C*: spectrograms on the left for a single cortical site in response to face (*top*) and voice (*bottom*) components of a grunt call. Population spectrograms are on the *right*. Conventions as in *A*. *D*: population spectra as a function of frequency for the low-frequency (1–37 Hz) responses for the face (*top*) and voice (*bottom*) components of the grunt call in *C*. Population spectra for the high frequency (37–200 Hz) responses are shown on the right. Conventions as in *B*.

all cortical sites aligned to onset of the visual stimulus for the three conditions: face+voice, face- and voice-alone conditions for a grunt call with a 97-ms TTV (*top*) and a coo call with a 219-ms TTV (*bottom*). For the grunt call, the peak theta band response to the face+voice is robustly enhanced relative to the face-alone condition [$t(70) = 6.05$, $P < 0.0001$] but is not significantly different from the voice-alone condition [$t(70) = 0.061$, $P = 0.951$]. Across the population, the theta band was enhanced for the face+voice relative to the voice-alone condition for 6 of 36 cortical sites (16%) and suppressed in 5 (14%) for this grunt call. For the coo call, the face+voice condition was enhanced relative to the face-alone condition [$t(70) = 2.87$, $P = 0.005$] but not significantly different from the voice-alone response [$t(70) = 0.95$, $P = 0.63$]. Across the population, the theta band face+voice responses were enhanced in 4 of 36 (11%) and suppressed in 4 of 36 (11%) of cortical sites when compared with the voice-alone response for this coo call. Thus multisensory effects in the theta band were weak.

Downloaded from jn.physiology.org on February 6, 2009

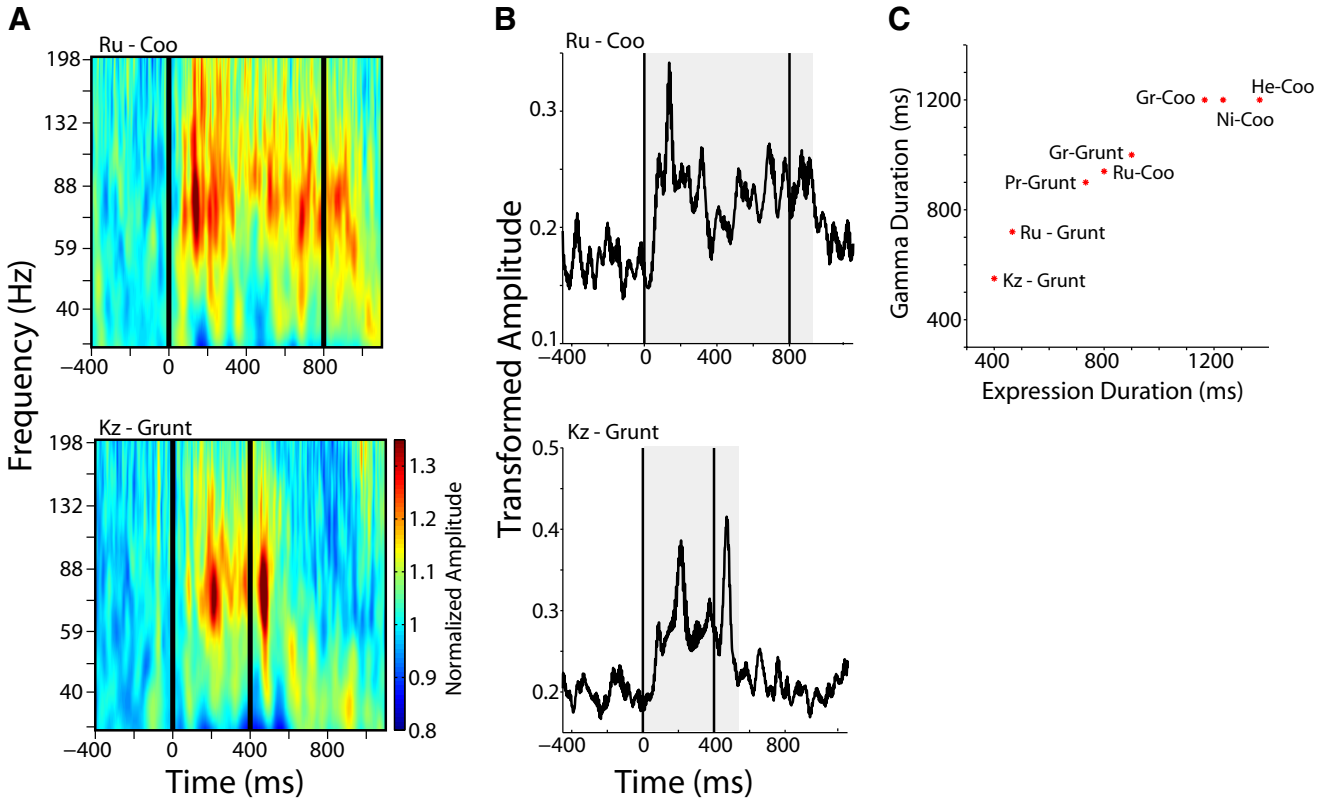

FIG. 4. Gamma activity is enhanced throughout the duration of the dynamic facial expression. *A*: spectrogram of the response to the face component of a coo call (*top*) averaged across cortical sites as well as a spectrogram of the response to the face component of a grunt call (*bottom*). *x* axes depict time in milliseconds, *y* axes depict frequency in hertz. Colorbars denote the activity normalized by the baseline mean. Black lines denote the onset and offset of the visual signal. *B*: Hilbert transformed gamma band activity (60–95 Hz) in response to the face component of the coo call and grunt call shown in *A*. *x* axes depict time in milliseconds. *y* axes depict the Hilbert transformed amplitude. Black lines denote onset and offset of the visual signal. *C*: relationship between the duration of gamma band activity and the duration of the dynamic visual stimulus in the STS. *x* axes depict stimulus duration in milliseconds. *y* axes depict gamma band duration in milliseconds. Each point denotes 1 call in the stimulus set. The last 3 calls are placed at 1,200 ms because our trial durations stopped at 1,200 ms after visual onset.

Figure 6, *B* and *C,* shows the response differences for all eight calls in terms of theta amplitude differences and percent integration in the population of cortical sites, respectively. In only two cases, did the theta band show integrative effects: suppression for the 75-ms TTV [$t(70) = -2.281$, $P = 0.026$] and the 331-ms TTV [$t(70) = -2.05$, $P = 0.044$; Fig. 6*B*]. This result precludes the use of a simple timing rule to explain the theta band integration. That is, the suppression did not occur exclusively for those calls with short TTVs versus those with long TTVs. To put it another way, calls with very similar short TTVs (66, 72, 75, and 85 ms) could be either suppressed (in 1 case) or show no significant difference in the theta band activity between the face+voice and both the unimodal conditions. The long TTV calls showed a similar mixture of suppression in one case (331 ms) and no difference in the two other TTVs (219 and 265 ms). For the five short TTV calls, theta band face+voice responses were enhanced in 16% of cortical sites (range: 8–28%), and suppressed in 13% of cortical sites (range: 5–20%) relative to the voice-alone response. For the three long TTV calls, 12% of cortical sites (range: 6–20%) showed enhanced responses, whereas 16% (range: 11–20%) showed suppressed responses. Figure 6*C* reveals that there was no systematic relationship between the multisensory integration index and the TTV for the theta band response. No significant linear relationship ($R^2 = 0.29$, $P = 0.16$) was observed between the amount of integration and the TTV.

Alpha band activity shows multisensory enhancement that is dependent on the TTV

Like the theta band, our wavelet analysis revealed that alpha band responses were larger in response to voices than faces. Multisensory responses were therefore defined again as face+voice responses that were significantly different from the voice-alone responses. Figure 7*A* shows the normalized alpha band activity aligned to visual onset for the grunt call with a 97-ms TTV (*top*) and the coo call with a 219 ms TTV(*bottom*). Gray regions denote 15-ms bins where face+voice responses were significantly different from the face- and voice-alone responses. For the grunt call, a robust enhancement of the face+voice response was observed relative to both the face- and voice-alone responses at multiple time points in the response. For example, at 66 ms after onset of the auditory signal, the face+voice response was enhanced relative to the voice-alone response [$t(70) = 2.58$, $P = 0.011$] and the face-alone response [$t(70) = 6.51$, $P < 0.0001$]. For this particular grunt call, the alpha band face+voice responses were enhanced in 16 of 36 cortical sites (44%) and suppressed in 1 site (2.7%) relative to the voice alone response.

Downloaded from jn.physiology.org on February 6, 2009

In contrast, for the 219-ms coo call, the face+voice alpha band response was not significantly different from the voice-alone response [$t(70) = -0.3012$, $P = 0.76$]. For this particular coo call, the alpha band face+voice responses were enhanced in only 6 of 36 cortical sites (16%) and suppressed in 5 (14%).

The responses to the face+voice for short TTV calls were enhanced relative to the voice-alone conditions for all five

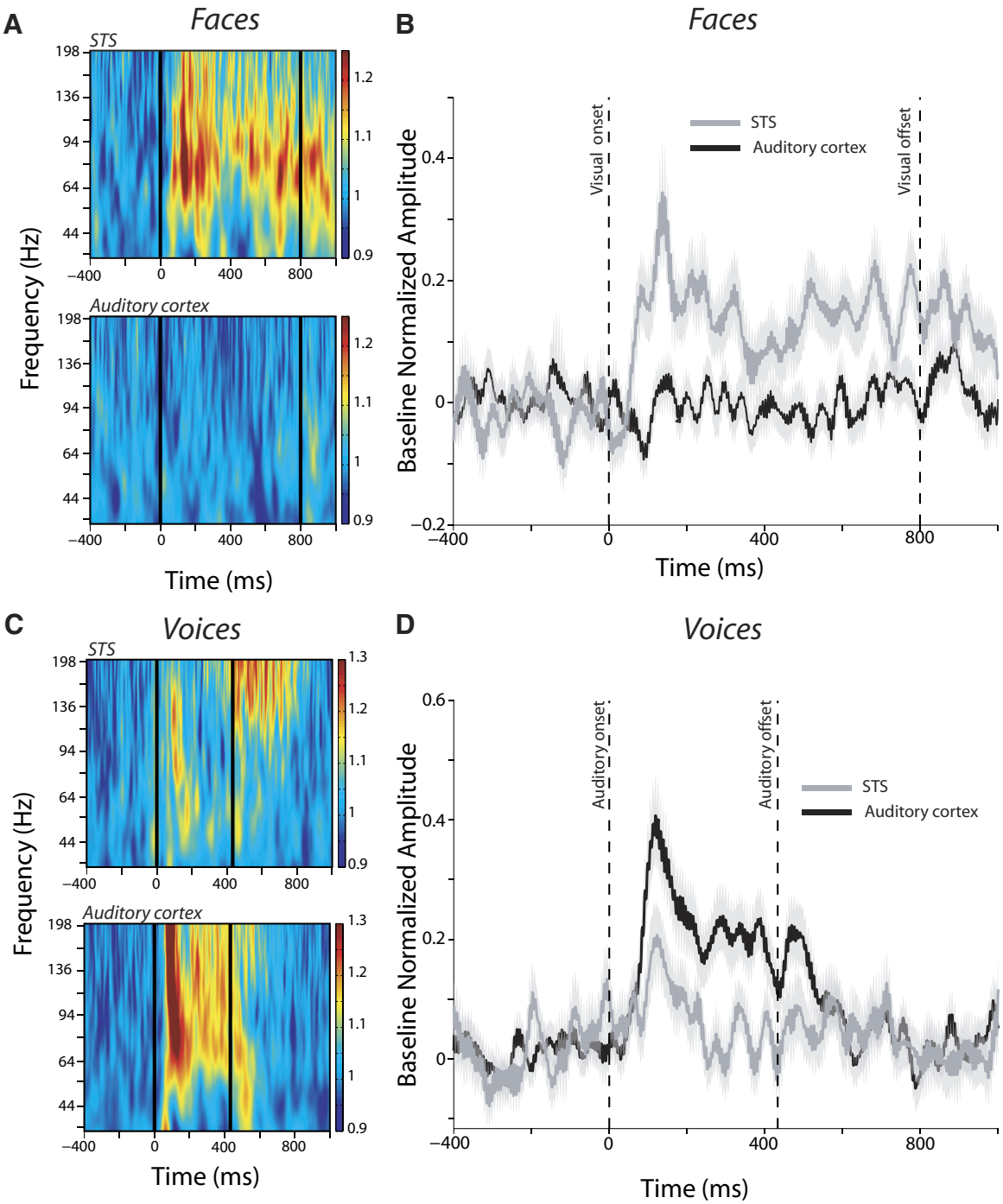

FIG. 5. Gamma activity in STS and auditory cortex display different dynamics in response to faces and voices. *A*: spectrograms show the mean normalized gamma band response in both STS ($n = 36$ cortical sites) and lateral belt auditory cortex ($n = 66$ cortical sites) to the face component of a coo call aligned to visual onset. *x* axes depict time in milliseconds, *y* axes depict frequency in hertz. Colorbar denotes the activity normalized by the baseline mean. Black lines denote the onset and offset of the face. *B*: Hilbert transformed gamma band activity (60–95 Hz) in STS and auditory cortex in response to the face component of the coo call shown in *A*. *x* axes depict time in milliseconds. *y* axes depict the Hilbert transformed amplitude in the gamma band. Dashed lines denote onset and offset of the face signal. *C*: spectrograms show the mean normalized gamma band response in both STS ($n = 36$ cortical sites) and lateral belt auditory cortex ($n = 66$ cortical sites) to the voice component of a coo call aligned to auditory onset. *x* axes depict time in milliseconds, *y* axes depict frequency in hertz. Colorbar denotes the activity normalized by the baseline mean. Black lines denote the onset and offset of the voice. *D*: Hilbert transformed gamma band activity (60–95 Hz) in STS and auditory cortex in response to the voice component of the coo call shown in *C*. *x* axes depict time in milliseconds. *y* axes depict the Hilbert transformed amplitude in the gamma band. Dashed lines denote onset and offset of the voice signal.

Downloaded from jn.physiology.org on February 6, 2009

calls: 66 ms: $t(70) = 2.68$, $P = 0.009$; 72 ms: $t(70) = 3.37$, $P = 0.0012$; 75 ms: $t(70) = 2.92$, $P = 0.004$; and 85 ms: $t(70) = 3.83$, $P = 0.0003$ (Fig. 7, *B* and *C*). In contrast, the responses to face+voice and voice-alone conditions were not significantly different for the three long TTV calls: 265 ms: $t(70) = 0.095$, $P = 0.92$; 331 ms: -0.24, $P = 0.81$. Enhanced alpha band face+voice responses relative to voice-alone responses were observed on average in 32% of cortical sites (range: 25–44%) for short TTV calls but only for 15% (range: 13 – 17%, $n = 3$ calls) for long TTV calls. Suppressed alpha band face+voice responses were observed in 7% of cortical sites for the short TTV calls (range: 2–11%), and 17% (range: 13–22%) for long TTV calls. Thus a much greater proportion of cortical sites displayed enhanced face+voice responses for short TTV calls versus long TTV calls. To quantify this difference, Fig. 7*C* plots the multisensory integration index for the alpha band responses according to TTV length. Relative to the voice-alone condition, short TTV calls showed enhanced alpha band face+voice responses, while the long TTV calls showed no such enhancement. The amount of enhancement was well correlated with the TTV ($R^2 = 0.77$, $P = 0.004$) suggesting that multisensory integration in the alpha band is sensitive to the time delay between face and voice stimuli, with shorter TTV calls showing more enhanced responses relative to the long TTV calls.

Gamma band activity shows multisensory enhancement only, independent of TTV

Figure 8*A* shows the normalized gamma band response (60–95 Hz) across all electrodes, aligned to the visual onset, for the 97-ms TTV grunt call (*top*) and the 331-ms TTV coo call (*bottom*). Gray regions denote 15-ms bins that showed significant differences between both the face+voice and face- and voice-alone responses. For the grunt call, the peak gamma band response to the face+voice condition was significantly enhanced relative to the face alone [$t(70) = 2.03$, $P = 0.0462$] and voice-alone conditions [$t(70) = 2.72$, $P = 0.0081$]. Face+voice responses were enhanced relative to the face- and voice-alone responses in 12 of 36 cortical sites (33%) for this grunt call. Similarly, for the coo call, the peak gamma band response to face+voice condition was enhanced relative to both the face alone [$t(70) = 3.10$, $P = 0.003$] and voice alone [$t(70) = 3.11$, $P = 0.003$] conditions. Such enhanced responses for this coo call were observed in 25% of cortical sites. Unlike alpha band responses (Fig. 7), which showed enhancement for all short TTV calls and no integration for long TTV calls, gamma band responses were nearly always enhanced independent of the TTV. Save for one, the 66-ms TTV, all other short TTV calls elicited enhanced face+voice responses relative to both the face alone [72 ms: $t(70) = 2.59$, $P = 0.011$; 75 ms: $t(70) = 3.39$, $P = 0.001$; 85 ms: $t(70) = 2.56$, $P = 0.012$] and voice alone conditions [72 ms: $t(70) = 3.30$, $P = 0.002$; 75 ms: $t(70) = 2.65$, $P = 0.009$; 85 ms: $t(70) = 2.84$, $P = 0.005$]. Similarly, face+voice gamma band responses to all long TTV calls were enhanced relative to both face-alone [219 ms: $t(70) = 4.38$, $P < 0.0001$; 265 ms: $t(70) = 4.84$, $P < 0.0001$] and voice-alone conditions [219 ms: $t(70) = 3.58$, $P = 0.001$; 265 ms: $t(70) = 4.43$, $P < 0.0001$]. Figure 8*B* plots the peak response in the gamma band during the auditory on and off period for each of the stimulus conditions. Thirty percent (range: 22–36%) of cortical sites showed enhanced responses to the face+voice condition for short TTV calls and 31% (range: 27–33%) for the long TTV calls. Suppression of the face+voice responses were <10% for both short and long TTV calls.

Across the population of cortical sites, seven of eight calls in our stimulus set elicited a robust enhancement for the face+voice condition relative to both faces and voices. Figure 8*C* plots the multisensory integration index relative to the voice-alone condition as a function of the TTV (note: the same pattern holds true when plotted relative to the face-alone condition). Unlike theta and alpha bands, population gamma band responses were consistently enhanced. No correlation was found between the TTV and the magnitude of enhancement ($R^2 = 0.21$, $P = 0.24$).

DISCUSSION

Using the macaque monkey as a model system and natural, dynamic vocalizations as stimuli, we tested the integration of faces and voices in the STS both as a function of different frequency bands of the LFP responses and as function of natural time delay between the initial facial movement and the onset of the voice component (the TTV). We found that faces and voices elicit distinct and concurrent activity in distinct frequency ranges of the STS LFP: theta (4–8 Hz), alpha (8–14 Hz), and gamma (>40 Hz). These three frequency bands integrated faces and voices differently. The theta band showed no consistent relationship between the multisensory response and the TTV, whereas the alpha band showed enhancement of power depending on the TTV. Short TTVs elicited enhancement, whereas long TTVs elicited no integration. The gamma band invariably showed enhanced responses to face-voice pairings, and these responses were independent of the TTV parameter.

Spectral profile of LFP responses to communication signals

Prior monkey physiology work investigating multisensory responses in the STS focused on single neurons. These studies generally concluded that single units in the STS were very responsive to visual inputs, particularly visual motion or face stimuli, and that they are occasionally responsive to auditory and tactile inputs (Baylis et al. 1987; Bruce et al. 1981; Hikosaka et al. 1988). The results from our study support these findings, but extend them in important ways. We investigated the structure of synaptic input processes in the STS on multiple time scales by analyzing the LFP. The LFP is composed of signals of several different frequency bands ranging from very slow to fast fluctuations (Buzsaki and Draguhn 2004), and many of these frequency bands can be differentially modulated by sensory, motor, and perceptual processes (Henrie and Shapley 2005; Kayser and Konig 2004; Liu and Newsome 2006; Pesaran et al. 2002; Rickert et al. 2005; Wilke et al. 2006). In the STS, LFP responses to both faces and voices were robust in general, but in the single units (Baylis et al. 1987; Bruce et al. 1981; Hikosaka et al. 1988) and high-frequency gamma band, visual responses were stronger and of longer duration than auditory responses. The gamma band responses to dynamic faces were sustained throughout the duration of a facial expression, suggesting that this region is actively involved in the processing of facial kinematics. In contrast, lower frequency

Downloaded from jn.physiology.org on February 6, 2009

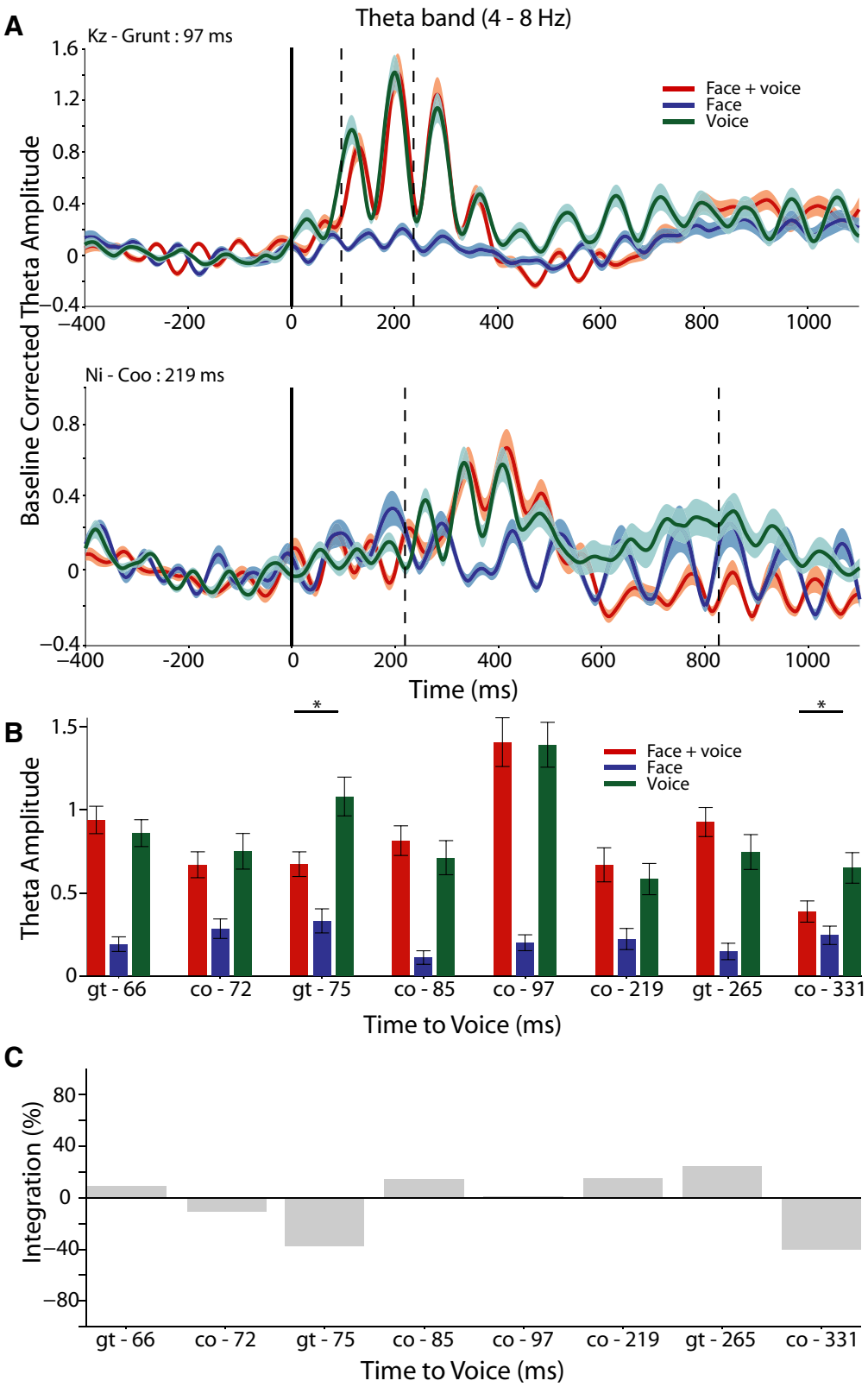

FIG. 6. Theta band activity shows little or no multisensory integration. *A*: theta band activity (4–8 Hz) for the face+voice, face-alone, and voice-alone conditions for a grunt call with a 97-ms time-to-voice (*top*) and a coo call with a 219-ms time-to-voice (*bottom*). *x* axes depict time in milliseconds. *y* axes depict the baseline corrected amplitude. Shaded regions denote SE. Solid line denotes the onset of the visual stimulus. Dashed lines denote the onset and offset of the auditory stimuli. Title denotes animal identity and the call type (coo or grunt) along with the corresponding time-to-voice. *B*: peak theta amplitude for the 8 calls in the stimulus set sorted as a function of time-to-voice. *x* axes depict time-to-voice (TTV) and call type (co – coo, gt – grunt). *y* axes depict peak amplitude. Asterisk denotes significant differences between face + voice and other unisensory responses. *C*: percent integration of the peak theta response to the face+voice condition relative to the voice-alone response as a function of the TTV. *x* axes depict TTV. *y* axes depict integration in percent.

Downloaded from jn.physiology.org on February 6, 2009

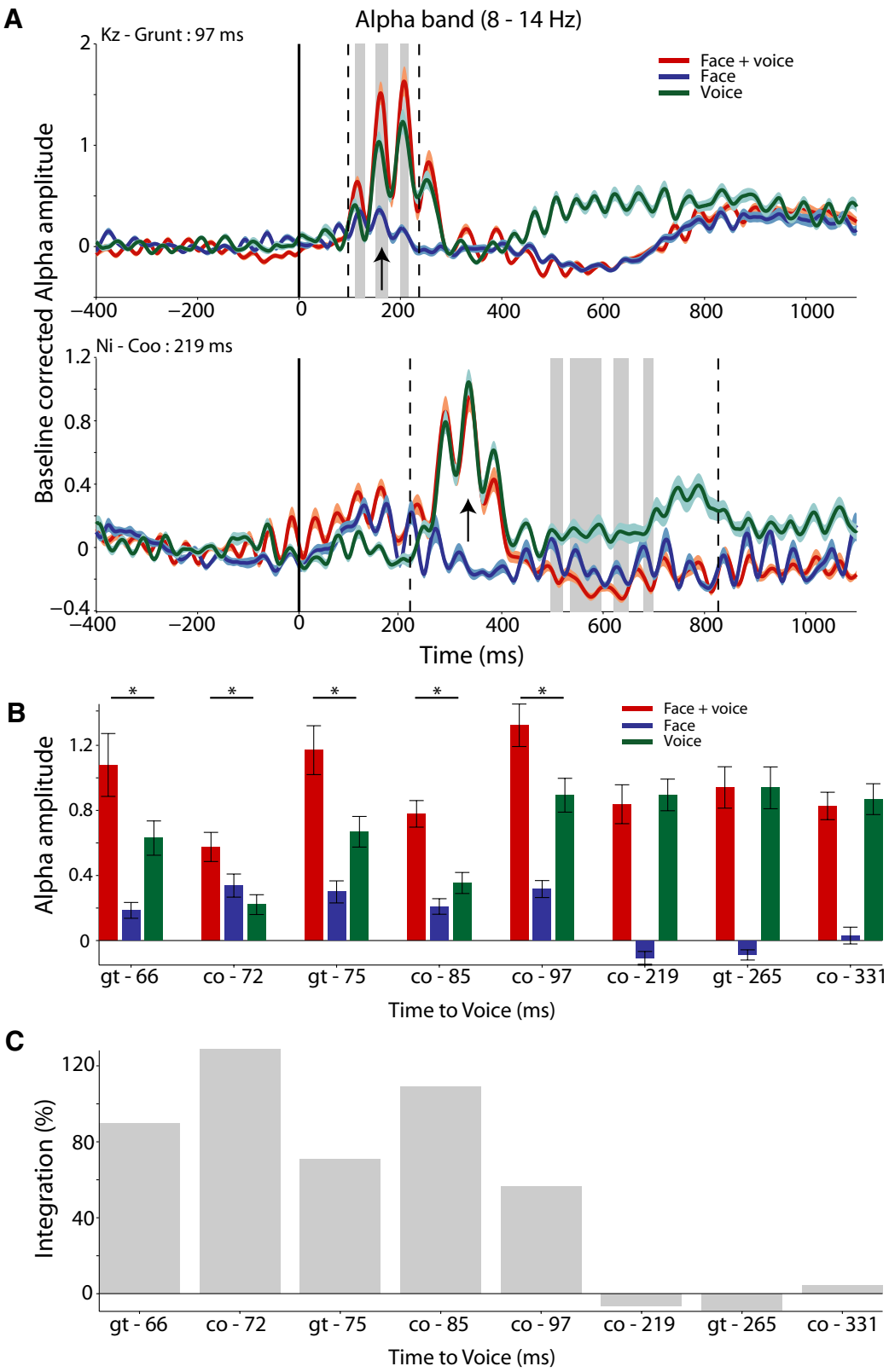

FIG. 7. Alpha band activity shows enhancement for short TTVs and suppression or no enhancement for long TTV calls. *A*: alpha band activity (8–14 Hz) for the face+voice, face-alone, and voice-alone conditions for a grunt call with a 97-ms TTV (*top*) and a coo call with a 219-ms TTV (*bottom*). ▩, time slices where significant differences were observed between the face + voice and face alone and voice alone conditions. ↑, the region where responses were taken for the graph in *B* and in *C*. Conventions as in Fig. 6*A*. *B*: baseline corrected peak amplitude in the alpha band for the 8 calls in the stimulus set sorted as a function of TTV. *x* axes depict TTV and call type (co – coo, gt – grunt). *y* axes depict peak amplitude. *, significant differences between face+voice and voice-alone responses. *C*: percent integration of the peak alpha response to the face+voice condition relative to the voice-alone response as a function of the TTV. *x* axes depict time, *y* axes depict integration in percent.

Downloaded from jn.physiology.org on February 6, 2009

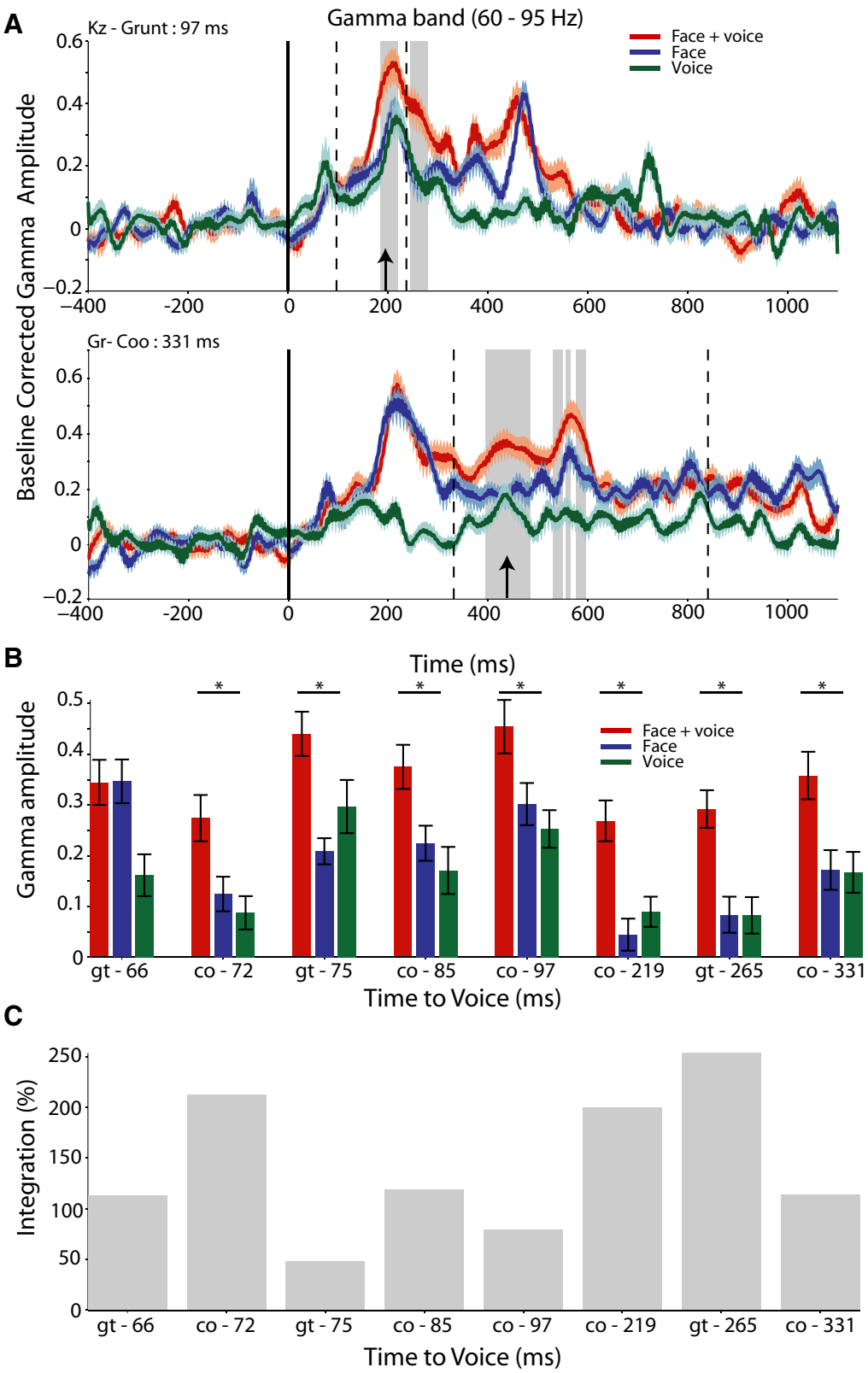

FIG. 8. Gamma band activity shows enhancement for both short and long TTVs. *A*: gamma band activity (60–95 Hz) for the face+voice, face-alone, and voice-alone conditions for a grunt call with a 97-ms TTV (*top*) and a coo call with a 331-ms TTV. ▢, time slices where significant differences were observed between the face+voice and face-alone and voice-alone conditions. ↑, the region where responses were taken for the graph in *B* and *C*. Conventions as in Fig. 6. *B*: peak amplitude in the gamma band for each of the 8 calls in the stimulus set as a function of TTV. *x* axes depict TTV and call type (co – coo, gt – grunt). *y* axes depict peak amplitude. *, significant differences between face+voice and voice-alone responses and between face+voice and face-alone responses. *C*: percent integration of the peak gamma response to the face+voice condition relative to the voice alone response as a function of the TTV. *x* axes depict time, *y* axes depict integration in percent.

Downloaded from jn.physiology.org on February 6, 2009

bands had initial onset responses followed by suppression for much of the duration of the facial expression. This suppression of the lower frequency bands has been repeatedly observed in the LFP response in other cortical areas during both perception of sensory stimuli (Liu and Newsome 2006) and movement (Scherberger et al. 2005). Further investigations are necessary to understand the relationship between suppression of the lower frequencies and relevant stimuli or task variables.

Remarkably, in the theta and alpha band, auditory inputs elicited stronger responses than visual inputs. This is somewhat surprising given that the STS, based on single-unit responses, has long been considered more visual than auditory. One possibility is that this discrepancy, at least in part, may be related to the dynamic nature of the visual stimuli used in our study as compared with the static stimuli used in the majority of single-unit investigations. For example, a recent event related potential (ERP) study compared the N170 response (a dominant, low-frequency response) to static versus dynamic faces and found that the static faces elicited a larger response (Puce et al. 2000). Thus our attenuated visual responses, relative to auditory responses, in the lower frequency bands may be influenced by our use of dynamic faces. Seemingly enhanced responses to voices therefore need to be interpreted with caution because it is being contrasted with responses to dynamic faces. Another reason for the observed differences in low-frequency responses could be due to the differential action of auditory and visual inputs on neurons in the STS. For example, in an intracellular study of neurons in the upper bank of the STS, Benevento and collaborators (1977) reported excitation by visual stimuli and primarily inhibition by auditory stimuli (similar to the responses seen in Fig. 2*B*). Consistent with our results, they found that auditory stimuli were only weakly effective in generating spiking activity in STS neurons. Thus the large low-frequency LFP responses to auditory stimuli accompanied by weak or no spiking activity could be due to the activation of a large inhibitory network—the LFP signal does not distinguish between excitatory and inhibitory synaptic inputs. Finally, a current source density analysis of visual and auditory processes in the STS found that auditory stimuli elicited robust current source density responses but weak spiking activity (Schroeder and Foxe 2002). Again, this is consistent with the present results.

Temporal principles of multisensory integration

The temporal principle of multisensory integration suggests that multisensory enhancement effects will be observed when auditory and visual stimuli are linked in time (Stein and Stanford 2008). The time window can be long, enabling integration to occur despite differences in response latencies, conduction speeds, and stimulus onset times. The magnitude of the integrated response is usually maximal when the peak periods of stimulus-induced activity coincide (Meredith et al. 1987; Stein and Stanford 2008). This principle was developed in the context of a spatial localization and an orienting paradigm (Jiang et al. 2002; Stein 1998), where temporal and spatial coincidence are the most salient cues for localization. Communication signals, such as speech and monkey vocalizations, possess a range of natural delays between their visual and auditory components. What temporal principle might be predictive of multisensory responses to these more complex stimuli? In a seminal study of multisensory responses in the primary auditory cortex, Lakatos et al. (2007) proposed phase resetting as a mechanism of interaction between multisensory stimuli. They observed that somatosensory stimuli, which are generally ineffective at eliciting suprathreshold responses in auditory cortex, nevertheless reset the phase of ongoing oscillations in auditory cortex. Auditory stimuli that arrive at the high excitability phase of the delta/theta/gamma components of this reset oscillation were enhanced, whereas responses to stimuli arriving at the low excitability phase of delta and theta cycles were suppressed. A similar pattern of results has also been reported for visual-auditory interactions in the belt auditory cortex (Kayser et al. 2008). Thus the tentative rule for multisensory integration in "unimodal" areas is that the nondominant modality resets the phase of the oscillation, and enhanced versus suppressed responses are determined by the timing of the dominant modality relative to the peaks or troughs of the oscillation.

Although we exploited only the natural time delays available in our stimulus and did not systematically vary them, our results are not easily explained by such a rule. Theta band responses showed no systematic integrative effects related to the TTV. In contrast, the alpha band showed enhancement for short TTV calls and no integration for calls with long TTVs. Finally, gamma band responses were consistently enhanced regardless of the TTV variable. These three patterns of integration (or lack thereof) occurred using the identical stimulus set for all three different neural frequency bands. Three reasons might account for the disparity of our STS data with the hypothesis proposed for unimodal sensory areas like auditory cortex (Lakatos et al. 2007). First, our natural stimuli are long in duration and lead to sustained gamma band activity and complex dynamics in the lower frequencies. Second, for such rich sensory signals such as faces and voices, there may be a complex interaction between the dynamics of the facial posture and the spectral and amplitude properties of the voice (Ghazanfar and Chandrasekaran 2007). Thus there may be more than one time point when visual and auditory information may synergize at the neural level or the dynamics of the face may reset the ongoing oscillation multiple times during an expression, complicating any straightforward prediction for when enhancement or suppression may occur. Third, the phase-resetting hypothesis was developed to explain multisensory integration in unimodal sensory cortices. STS is an association area receiving synaptic inputs from multiple sensory systems—robust enough to at least occasionally elicit suprathreshold responses from multiple modalities. In contrast, suprathreshold responses to visual or somatosensory inputs alone are not generally seen in auditory cortex (Ghazanfar et al. 2008; Lakatos et al. 2007). Thus the phase-resetting hypothesis may not apply to the STS (Ghazanfar and Schroeder 2006).

Different neural frequency bands may have different roles in multisensory integration

Previous studies investigated how single sensory stimuli or motor actions are encoded by the different frequency ranges (Liu and Newsome 2006; Rickert et al. 2005; Wilke et al. 2006). Likewise, our data suggest that different frequency

Downloaded from jn.physiology.org on February 6, 2009

bands in the STS signal may represent multiple underlying processes during audiovisual communication. In our paradigm, the multisensory low-frequency theta and alpha band responses show very different properties compared with the gamma responses. Such a result reinforces conclusions from a recent study that used signal and noise correlations to show that lower frequencies (theta and alpha ranges) are independent of activity in the gamma band (Belitski et al. 2008). The different frequency bands reflect different neuronal processes and are generated by distinct sources. One attractive hypothesis is that the frequency of a band is inversely proportional to the scale of the cortical network underlying that frequency (Buzsaki 2006). Gamma band activity is thought to reflect activity in a local cortical microcolumn (Liu and Newsome 2006), whereas alpha and theta band activity could putatively represent input processes, the sources of which may include multiple cortical areas acting in a coordinated fashion. In our case, this would mean that the areas with which the STS connects would have similar behavior in the lower frequencies but divergent behavior in higher frequencies (as seen in Fig. 5 comparing auditory cortex to STS). Putative sources of such communication-relevant inputs into STS include the ventrolateral prefrontal cortex (Romanski et al. 1999), frontal eye fields (Schall et al. 1995), and the belt region of auditory cortex (Seltzer and Pandya 1994).

Recordings from multiple structures simultaneously and joint analysis of these signals as has been done for auditory cortex and STS (Ghazanfar et al. 2008; Maier et al. 2008), will help shed light onto the mechanisms underlying activity in these different frequency ranges and their role in multisensory integration. Given that the integration of faces and voices appears to be experience-dependent (Lewkowicz and Ghazanfar 2006; Lewkowicz et al. 2008), it will also be important to investigate how experience shapes these different neural frequency bands and the network states that they represent (Fontanini and Katz 2008). Are they simultaneously affected during learning of face/voice associations, or do they change in sequential manner with changes in the higher frequency bands emerging before lower frequency bands as occurs during development (Mohns and Blumberg 2008)?

Resolving discrepancies between EEG and fMRI studies of audiovisual speech

Several studies investigating the neural bases of audiovisual speech in humans revealed that the auditory N1 component (from EEG or MEG) is suppressed for dynamic audiovisual speech (Besle et al. 2004; Klucharev et al. 2003; van Wassenhove et al. 2005). In contrast, hemodynamic studies of audiovisual integration have consistently observed enhancement of the multisensory response compared with the unisensory responses in the upper bank of the STS (Callan et al. 2003, 2004; Calvert 2001; Calvert et al. 2000; Skipper et al. 2005). Our dissociation between activity in low and high frequencies suggest some resolution to this apparent discrepancy. First, ERPs are dominated by low-frequency responses. Our results suggest that delays between auditory and visual stimuli play a crucial role in deciding whether low-frequency responses to multisensory stimuli are enhanced or suppressed. In support of these, a previous report from our group, using the same stimuli and task, demonstrated that enhanced versus suppressed LFP responses to audiovisual communication signals in the auditory cortex is determined by the TTV parameter: short delays lead to enhancement, longer delays to suppression (Ghazanfar et al. 2005). The temporal dynamics of audiovisual speech (as presented in an experimental paradigm) therefore need to be specified to interpret enhanced versus suppressed responses in the ERP localized to the superior temporal gyrus. This conclusion is well supported by a recent ERP study that suggests that the duration of anticipatory visual motion decides the level of suppression of the multisensory response relative to the auditory alone response (Stekelenburg and Vroomen 2007).

The enhanced hemodynamic response observed in studies of audiovisual speech can be explained by appealing to results that suggest that gamma band activity and the blood-oxygen-level-dependent (BOLD) response are well-correlated with each other (Goense and Logothetis 2008; Niessing et al. 2005; Nir et al. 2007). Thus the enhanced gamma band activity observed in this current study supports and extends findings that the BOLD signal localized to the STS is sensitive to audiovisual stimuli. The multisensory gamma band response was never suppressed relative to the unisensory conditions, and this was insensitive to the delay between visual and auditory stimuli. The enhanced BOLD responses in neuroimaging studies of audiovisual speech might therefore be explained by the enhanced gamma band activity we observed.

ACKNOWLEDGMENTS

The physiological data were collected at the Max Planck Institute for Biological Cybernetics in Tuebingen, Germany by A. A. Ghazanfar.

GRANTS

This work was supported by the National Institute of Neurological Disorders and Stroke Grant R01NS-054898 to A. A. Ghazanfar, the National Science Foundation BCS-0547760 CAREER Award to A. A. Ghazanfar, Autism Speaks to A. A. Ghazanfar, and Princeton University's Quantitative and Computational Neuroscience training grant National Institutes of Health R90 DA-023419-02 to C. Chandrasekaran.

REFERENCES

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

Baylis GC, Rolls ET, Leonard CM. Functional subdivisions of the temporal lobe neocortex. *J Neurosci* 7: 330–342, 1987.

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

Belitski A, Gretton A, Magri C, Murayama Y, Montemurro MA, Logothetis NK, Panzeri S. Low-Frequency local field potentials and spikes in primary visual cortex convey independent visual information. *J Neurosci* 28: 5696–5709, 2008.

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

Benevento LA, Fallon J, Davis BJ, Rezak M. Auditory-visual interactions in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Exp Neurol* 57: 849–872, 1977.

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

Downloaded from jn.physiology.org on February 6, 2009

Bruce C, Desimone R, Gross CG. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol* 46: 369–384, 1981.

Buzsaki G. *Rhythms of The brain.* New York: Oxford, 2006.

Buzsaki G, Draguhn A. Neuronal oscillations in cortical networks. *Science* 304: 1926–1929, 2004.

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

Callan DE, Jones JA, Munhall K, Kroos C, Callan AM, Vatikiotis-Bateson E. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J Cognit Neurosci* 16: 805–816, 2004.

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

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

Ethofer T, Pourtois G, Wildgruber D. Investigating audiovisual integration of emotional signals in the human brain. *Prog Brain Res* 156: 345–361, 2006.

Fontanini A, Katz DB. Behavioral states, network states, and sensory response variability. *J Neurophysiol* 100: 1160–1168, 2008.

Ghazanfar AA, Chandrasekaran CF. Paving the way forward: integrating the senses through phase-resetting of cortical oscillations. *Neuron* 53: 162–164, 2007.

Ghazanfar AA, Chandrasekaran C, Logothetis NK. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J Neurosci* 28: 4457–4469, 2008.

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

Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J Neurosci* 25: 5004–5012, 2005.

Ghazanfar AA, Nielsen K, Logothetis NK. Eye movements of monkeys viewing vocalizing conspecifics. *Cognition* 101: 515–529, 2006.

Ghazanfar AA, Schroeder CE. Is neocortex essentially multisensory? *Trends Cogni Sci* 10: 278–285, 2006.

Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. Vocal tract resonances as indexical cues in rhesus monkeys. *Curr Biol* 17: 425–430, 2007.

Gifford GW, Cohen YE. Effect of a central fixation light on auditory spatial responses in area LIP. *J Neurophysiol* 91: 2929–2933, 2004.

Goense JB, Logothetis NK. Neurophysiology of the BOLD fMRI signal in awake monkeys. *Curr Biol* 18: 631–640, 2008.

Hackett TA, Stepniewska I, Kaas JH. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Comp Neurol* 394: 475–495, 1998.

Henrie JA, Shapley R. LFP power spectra in V1 cortex: the graded effect of stimulus contrast. *J Neurophysiol* 94: 479–490, 2005.

Hikosaka K, Iwai E, Saito HA, Tanaka K. Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *J Neurophysiol* 60: 1615–1637, 1988.

Jiang W, Jiang H, Stein BE. Two corticotectal areas facilitate multisensory orientation behavior. *J Cognit Neurosci* 14: 1240–1255, 2002.

Jordan KE, Brannon EM, Logothetis NK, Ghazanfar AA. Monkeys match the number of voices they hear with the number of faces they see. *Curr Biol* 15: 1034–1038, 2005.

Kayser C, Konig P. Stimulus locking and feature selectivity prevail in complementary frequency ranges of V1 local field potentials. *Eur J Neurosci* 19: 485–489, 2004.

Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. *Cereb Cortex* 18: 1560–1574, 2008.

Klucharev V, Mottonen R, Sams M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognit Brain Res* 18: 65–75, 2003.

Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D. Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage* 37: 1445–1456, 2007.

Lakatos P, Chen C-M, O'Connell MN, Mills A, Schroeder CE. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53: 279–292, 2007.

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94: 1904–1911, 2005.

Lee D. Analysis of phase-locked oscillations in multi-channel single-unit spike activity with wavelet cross-spectrum. *J Neurosci Methods* 115: 67–75, 2002.

Lewkowicz DJ, Ghazanfar AA. The decline of cross-species intersensory perception in human infants. *Proc Natl Acad Sci USA* 103: 6771–6774, 2006.

Lewkowicz DJ, Sowinski R, Place S. The decline of cross-species intersensory perception in human infants: underlying mechanisms and its developmental persistence. *Brain Res* 1242: 291–302, 2008.

Liu J, Newsome WT. Local field potential in cortical area MT: Stimulus tuning and behavioral correlations. *J Neurosci* 26: 7779–7790, 2006.

Logothetis NK. The underpinnings of the BOLD functional magnetic resonance imaging signal. *J Neurosci* 23: 3963–3971, 2003.

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

Maier JX, Chandrasekaran C, Ghazanfar AA. Integration of bimodal looming signals through neuronal coherence in the temporal lobe. *Curr Biol* 18: 963–968, 2008.

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

Mohns EJ, Blumberg MS. Synchronous bursts of neuronal activity in the developing hippocampus: modulation by active sleep and association with emerging gamma and theta rhythms. *J Neurosci* 28: 10134–10144, 2008.

Niessing J, Ebisch B, Schmidt KE, Niessing M, Singer W, Galuske RAW. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science* 309: 948–951, 2005.

Nir Y, Fisch L, Mukamel R, Gelbard-Sagiv H, Arieli A, Fried I, Malach R. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Curr Biol* 17: 1275–1285, 2007.

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze H-J, Driver J. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J Neurosci* 27: 11431–11441, 2007.

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

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

Puce A, Smith A, Allison T. ERPs evoked by viewing facial movements. *Cognit Neuropsychol* 17: 221–239, 2000.

Rickert J, de Oliveira SC, Vaadia E, Aertsen A, Rotter S, Mehring C. Encoding of movement direction in different frequency ranges of motor cortical local field potentials. *J Neurosci* 25: 8815–8824, 2005.

Romanski LM, Bates JF, Goldman-Rakic PS. Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *J Comp Neurol* 403: 141–157, 1999.

Schall JD, Morel A, King DJ, Bullier J. Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J Neurosci* 15: 4464–4487, 1995.

Scherberger H, Jarvis MR, Andersen RA. Cortical local field potential encodes movement intentions in the posterior parietal cortex. *Neuron* 46: 347–354, 2005.

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

Seltzer B, Pandya DN. 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, 1994.

Senkowski D, Schneider TR, Foxe JJ, and Engel AK. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci* 31: 401–409, 2008.

Skipper JI, Nusbaum HC, Small SL. Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25: 76–89, 2005.

Stein BE. Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. *Exp Brain Res* 123: 124–135, 1998.

Stein BE, Stanford TR. Multisensory integration: current issues from the perspective of the single neurons. *Nat Rev Neurosci* 9: 255–266, 2008.

Stekelenburg JJ, Vroomen J. Neural correlates of multisensory integration of ecologically valid audiovisual events. *J Cognit Neurosci* 19: 1964–1973, 2007.

Downloaded from jn.physiology.org on February 6, 2009

Sugihara T, Diltz M, Averbeck BB, Romanski LM. Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *J Neurosci* 26: 11138–11147, 2006.

van Atteveldt N, Formisano E, Goebel R, Blomert L. Integration of letters and speech sounds in the human brain. *Neuron* 43: 271–282, 2004.

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

Varela F, Lachaux J-P, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. *Nat Rev Neurosci* 2: 229–239, 2001.

Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall KG. Eye movement of perceivers during audiovisual speech perception. *Percept Psychophys* 60: 926–940, 1998.

Wallace MT, Ramachandran R, Stein BE. A revised view of sensory cortical parcellation. *Proc Nat Acad Sci USA* 101: 2167–2172, 2004.

Wilke M, Logothetis NK, Leopold DA. Local field potential reflects perceptual suppression in monkey visual cortex. *Proc Natl Acad Sci USA* 103: 17507–17512, 2006.

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

Downloaded from jn.physiology.org on February 6, 2009