EJN fens EUROPEAN JOURNAL OF NEUROSCIENCE

European Journal of Neuroscience, Vol. 31, pp. 1807–1817, 2010 doi:10.1111/j.1460-9568.2010.07209.x

Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: implications for the evolution of audiovisual speech

Asif A. Ghazanfar,[1,2,3] Chandramouli Chandrasekaran[1,2] and Ryan J. Morrill[3]
[1]Neuroscience Institute, [2]Department of Psychology, and [3]Department of Ecology & Evolutionary Biology, Princeton University, Princeton NJ 08540, USA

Keywords: dynamic faces, face processing, facial expressions, jaw movement, monkey vocalizations, multisensory integration

Abstract

Audiovisual speech has a stereotypical rhythm that is between 2 and 7 Hz, and deviations from this frequency range in either modality reduce intelligibility. Understanding how audiovisual speech evolved requires investigating the origins of this rhythmic structure. One hypothesis is that the rhythm of speech evolved through the modification of some pre-existing cyclical jaw movements in a primate ancestor. We tested this hypothesis by investigating the temporal structure of lipsmacks and teeth-grinds of macaque monkeys and the neural responses to these facial gestures in the superior temporal sulcus (STS), a region implicated in the processing of audiovisual communication signals in both humans and monkeys. We found that both lipsmacks and teeth-grinds have consistent but distinct peak frequencies and that both fall well within the 2–7 Hz range of mouth movements associated with audiovisual speech. Single neurons and local field potentials of the STS of monkeys readily responded to such facial rhythms, but also responded just as robustly to yawns, a nonrhythmic but dynamic facial expression. All expressions elicited enhanced power in the delta (0–3Hz), theta (3–8Hz), alpha (8–14Hz) and gamma (> 60 Hz) frequency ranges, and suppressed power in the beta (20–40Hz) range. Thus, STS is sensitive to, but not selective for, rhythmic facial gestures. Taken together, these data provide support for the idea that that audiovisual speech evolved (at least in part) from the rhythmic facial gestures of an ancestral primate and that the STS was sensitive to and thus 'prepared' for the advent of rhythmic audiovisual communication.

Introduction

Audiovisual speech has a stereotypical rhythm that is between 2 and 7 Hz (Ohala, 1975; Munhall & Vatikiotis-Bateson, 1998; Greenberg *et al.*, 2003; Chandrasekaran *et al.*, 2009). This frequency range is related to the rate of syllable production, and disrupting the auditory component of this rhythm significantly reduces intelligibility (Drullman *et al.*, 1994; Shannon *et al.*, 1995; Saberi & Perrott, 1999; Smith *et al.*, 2002), as does disrupting the visual component (Vitkovitch & Barber, 1994, 1996; Kim & Davis, 2004; Campbell, 2008). In light of these data, recent neural theories of speech perception noted that the temporal modulations in speech are well matched to brain rhythms in the same frequency range (Poeppel, 2003; Schroeder *et al.*, 2008). Schroeder *et al* suggested that fast neocortical oscillations are phase-amplitude coupled to slower oscillations, and that these slower oscillations are entrained by the rhythmic structure of speech (Schroeder *et al.*, 2008). In a related theory, Poeppel *et al* suggest that the syllable rate is preferentially processed in a time window of ~150 to 300 ms and is mediated by the theta (3–8 Hz) rhythm in auditory cortex (Poeppel, 2003; Giraud *et al.*, 2007; Luo & Poeppel, 2007).

Correspondence: Dr Asif A. Ghazanfar, Neuroscience Institute, Green Hall, Princeton University, Princeton, NJ 08540, USA.
E-mail: asifg@princeton.edu

Received 30 September 2009, accepted 2 February 2010

Given the importance of the 2–7 Hz rhythm in audiovisual speech, understanding how audiovisual speech evolved requires investigating the origins of its rhythmic structure. Some have suggested that the cyclical basis of human speech (i.e., syllable production) evolved *de novo* in humans (Pinker & Bloom, 1990). An alternative account is that the rhythm of speech evolved through the modification of some pre-existing cyclical mandibular (jaw) movements in ancestral primates (MacNeilage, 1998, 2008). For example, while mandibular cyclicities are relatively rare during vocal production by nonhuman primates, they are extremely common as facial communicative gestures. Gestures such as lipsmacks and teeth-grinds of macaque monkeys involve cyclical movements of the mouth and are not accompanied by any vocal fold adduction (Hinde & Rowell, 1962; Redican, 1975). Thus, during the course of human evolution, as the theory goes, these nonvocal rhythmic facial expressions were coupled to vocalizations (MacNeilage, 1998, 2008).

Tests of such evolutionary hypotheses are difficult. However, if the idea that rhythmic speech evolved through the rhythmic facial expressions of ancestral primates has any validity then there are two predictions that can be tested using the comparative approach. The first is that, like speech, these rhythmic facial expressions in extant primates should occur within the 2–7 Hz frequency range. It is important to note that there are other important temporal scales in the production and perception of speech. Speech contains cues on

timescales as short as 50 ms and long utterances can contain units of information several seconds in length (Lieberman & Blumstein, 1988). However, as Old World monkey vocalizations do not appear to have important acoustic events on these very short and very long timescales, the focus on mandibular cycles seems appropriate (MacNeilage, 1998, 2008). The second prediction is that neocortical structures that are sensitive to faces and audiovisual communication signals in humans and other primates should also be responsive to these rhythmic nonvocal facial expressions. Such sensitivity would indicate that the nonhuman primate brain was, in some sense, 'prepared' for the evolution of audiovisual speech.

In the current study, we tested these hypotheses by examining the temporal structure of lipsmacks and teeth-grinds of macaque monkeys and the neural responses to these facial gestures in the superior temporal sulcus (STS), a region widely implicated in the processing of audiovisual communication signals in both humans (Calvert *et al.*, 2000; Callan *et al.*, 2003; Calvert & Campbell, 2003; Wright *et al.*, 2003; Reale *et al.*, 2007; Schroeder *et al.*, 2008) and monkeys (Barraclough *et al.*, 2005; Chandrasekaran & Ghazanfar, 2009; Dahl *et al.*, 2009). To test whether the STS was selective for rhythmic facial gestures as opposed to just being sensitive to them, we included 'yawns', which are dynamic, temporally-extended but nonrhythmic facial expressions.

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 (4.7T magnet, 500-μm slices) to identify the stereotaxic coordinates of the STS and to model a 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. Isofluorane anesthesia (1–2% in air via intubation tube) was used during the surgery. Buprenorphine (0.01 mg/kg) was used as the analgesic, pre- and postoperatively. 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 & O'Connor, 1980). The animals were given 3 months recovery time, and acclimatisation to head fixation was 1–2 weeks. All experiments were performed in compliance with the guidelines of the local authorities (Regierungspraesidium Tuebingen) 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 facial gestures spontaneously produced by rhesus monkeys in the same colony as the subject monkeys. The stimuli were filmed while monkeys were 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.

Analysis of mouth dynamics

In order to determine the frequency of rhythmic facial expressions, we analyzed video clips of lipsmacks and teeth-grinds using MATLAB to measure vertical mouth displacement as a function of time. Video was 30 frames per second, noninterlaced. Nyquist frequency for the data is 15 Hz ($f_{Nyquist}$ = ½ v, where v = sampling rate). Mouth displacement was measured frame-by-frame for the period of the relevant facial gesture. For teeth-grinds, during which the lips typically separate, displacement was measured by manually specifying a point in the middle of the upper lip and a point in the middle of the bottom lip. Lipsmacks are variable for mouth opening, so inter-lip distance does not always capture mouth displacement. During lipsmacks, the lips commonly pucker while moving up and down, but do not separate (Fig. 1A, top row). In these closed-mouth lipsmacks, displacement was measured as the distance between the lower lip and the nasion (the middle of the forehead where the bridge of the nose begins), an easily identifiable point on the face that does not move significantly during the target communicative gestures. During open-mouth lipsmacks, mouth displacement was measured in the same manner as teeth-grinds, using inter-lip distance. One oscillatory cycle is defined either as one mouth opening from closed to closed position, or one vertical displacement of the lips beginning and terminating at natural position. The data set presented here consists of 15 lipsmack bouts, mean video clip duration of (mean ± SD) 5.2 ± 1.9 s, and 21 teeth grind bouts, mean clip duration of 1.7 ± 0.6 s.

Temporal frequency modulation of the mouth oscillation was estimated using a multi-taper Fourier transform (Chronux Toolbox; http://www.chronux.org). A power spectrum was generated for each gestural bout. To minimize extreme low-frequency noise and because of the Nyquist limit frequency, frequency pass band for the Fourier analysis was $0.5 = f_{pass} = 12$ Hz. For each spectrum, the frequency of peak spectral density was measured and considered to be the average rate of mouth oscillation for the corresponding gestural bout.

Behavioral apparatus and paradigm

Neurophysiology 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-inch color monitor at a distance of 94 cm. 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. The videos were displayed centrally at $10 \times 6.6°$. 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; Chandrasekaran & Ghazanfar, 2009). 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 two exemplars of lipsmacks and two exemplars of teeth-grinds. Each exemplar was produced by a different monkey (i.e., there were four different identities).

Data collection

Recordings were made from the upper bank of the STS in the same region in which we've previously reported integration of dynamic faces and voices (Chandrasekaran & Ghazanfar, 2009) and functional interactions with auditory cortex during this process (Ghazanfar *et al.*, 2008). 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 (1–5000 Hz)

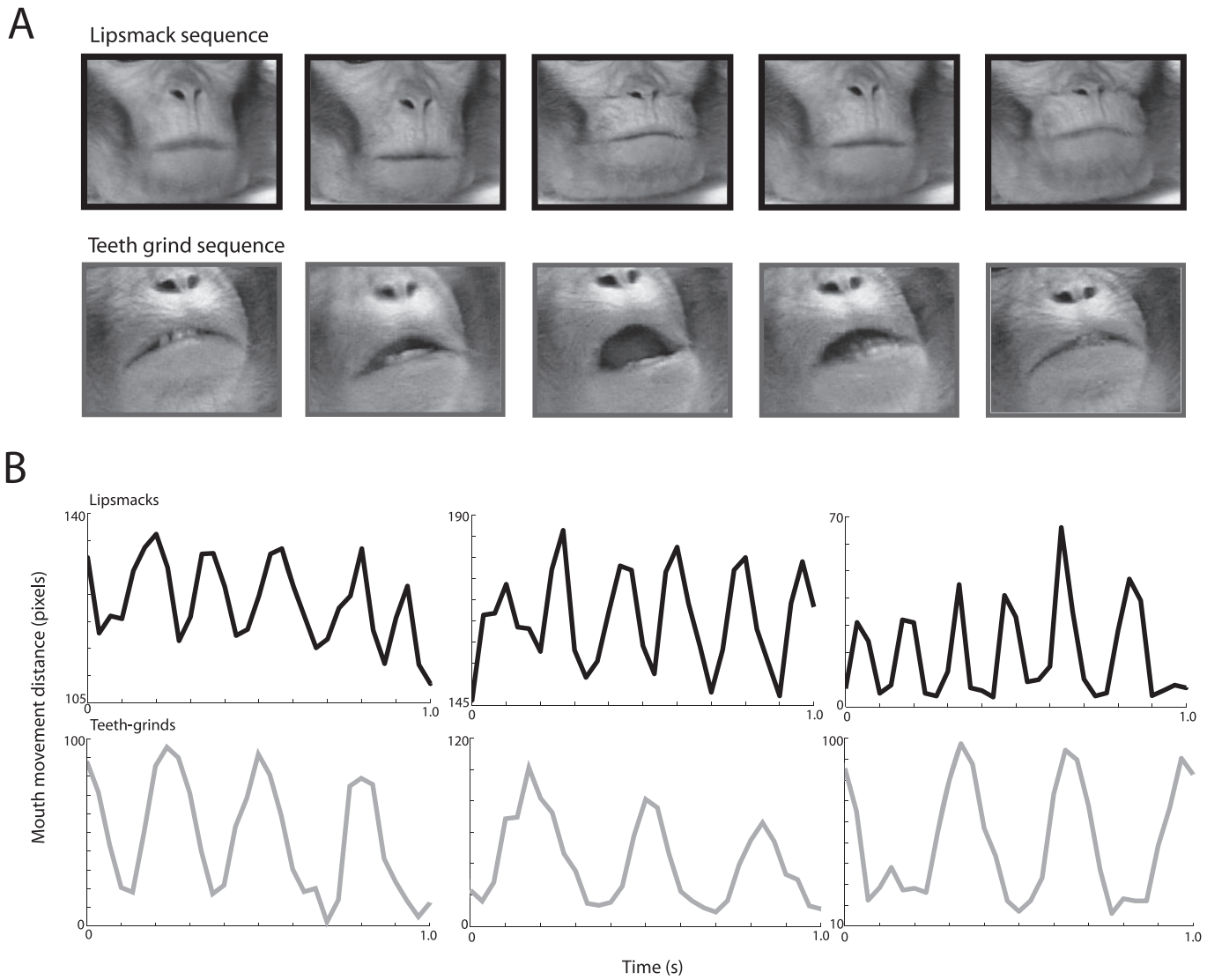

FIG. 1. Lipsmack and teeth-grind exemplars and time-series. (A) Frames of the orofacial region during two types of visuofacial communicative gestures. Above, example lipsmack, two cycles shown. Lipsmacks are variable in mouth opening. As in the example above, often the lips remain closed and touching while moving up and down vertically in a highly stereotyped fashion. Vertical displacement of the mouth can be noted by the relative position of the lips to the nose, reaching a minimum mid-cycle. Below, example teeth-grind, one cycle shown. Teeth-grinds involve an opening of the mouth with a separation of the lips. Mid-cycle there may be lateral displacement of the jaw, variable in its horizontal direction (see middle frame). (B) Above, three example lipsmack time-series graphs and, below, three example teeth-grind time-series graphs. Graphs depict mouth displacement in pixels as a function of time in seconds. *X*-axis is time in seconds; *Y*-axis depicts mouth movement distance in pixels; the distinct temporal modulation of lipsmacks and teeth grinds can be seen, with lipsmacks modulated at a higher rate. Note the rhythmic nature of both lipsmacks and teeth grinds.

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 (high-pass filtered at 500 Hz), frequency-tuning curves were collected for each site using 25 pure tone pips (100–21 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 and local field potentials recorded in this region (Chandrasekaran & Ghazanfar, 2009). We recorded activity from 45 cortical sites over 18 different sessions. A maximum of four electrodes were lowered into the STS in a given session; the inter-electrode distance was never less than 3 mm.

Data processing and analyses

Single units were extracted from the raw neural signal using principle component-based off-line spike-sorting in combination with time–voltage window thresholds. Only well-isolated neurons were included in the analyses (a minimum 6:1 signal-to-noise ratio). The time series of spikes was averaged across trials and then convolved with a Gaussian kernel of a particular width to produce a spike density function (Szucs, 1998). For our data, spike density functions were calculated by averaging spike trains and filtering with a 10-ms Gaussian kernel.

Basic response properties of neurons were assessed using the firing rate of the neuron binned in 20-ms bins stepped by 1 ms. We used a two-sample *t*-test to compare whether firing rate in this 20-ms bin was significantly enhanced or suppressed relative to the firing rate in the 200 ms period before stimulus onset. This allowed us to investigate for each neuron whether it responded or not to a particular stimulus.

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

Spectral analyses

To examine the different frequency bands in the LFP that may be modulated by dynamic faces, we performed spectral analyses. All the spectral analyses were based on wavelet spectra using modified scripts based on the Chronux suite of Matlab routines (http://www.chronux.org) and Matlab scripts provided to us courtesy of Daeyeol Lee (Lee, 2002; see also Ghazanfar *et al.*, 2008 for details).

In all spectral analyses, we had to determine whether changes in signal power were increased, decreased or stayed the same during the presentation of dynamic faces. To do so, we compared the neural signal during stimulus presentation with the signal during the absence of stimulation (baseline). 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. A value of 1 indicates that the stimulus activity was the same as the baseline activity. Values > 1 indicate enhancement, and < 1 indicate suppression.

Results

Rhythmic facial expressions in macaques and their temporal structure

Macaque monkeys frequently produce two types of rhythmic facial expressions that are not accompanied by vocalizations: lipsmacks and teeth-grinds (Hinde & Rowell, 1962). Lipsmacks are kissing-like movements made by rapidly, but subtly, puckering and unpuckering the lips (Fig. 1A, top row); jaw and tongue movements are also sometimes apparent in this expression, but the teeth do not meet. Lipsmacks occur in a number of social contexts, but always involve face-to-face encounters and positive social interactions. Teeth-grinds, also known as 'teeth-chatters', involve lateral movements of the lower jaw with a grinding of the teeth (Fig. 1A, bottom row). Teeth-grinds seem to be nonspecific in that there are a number of situations that elicit them; these situations are usually accompanied by a high arousal state.

We analyzed the rate of mouth movements in these two rhythmic facial expressions by measuring the inter-lip distance as a function of time (Chandrasekaran *et al.*, 2009). Figure 1B shows time-series of each of three exemplars of lipsmacks (top row) and teeth-grinds (bottom row). The rhythmic nature is readily apparent in both facial expressions, and lipsmacks seem to be produced at a higher frequency than teeth-grinds. To quantify this, we performed a spectral analysis on these mouth movement dynamics. Figure 2A reveals that lipsmacks are consistently produced with a higher peak mouth movement frequency (5.82 ± 0.90 Hz, mean ± SD; $n = 15$) than teeth-grinds (3.19 ± 0.54 Hz, $n = 21$). The peak frequencies differed significantly between these two expressions ($t_{34} = 10.962$, $P < 0.001$; Fig. 2B).

Taken together, these data reveal that rhythmic facial expressions of macaque monkeys are produced with different stereotypical mouth movement frequencies, and both frequencies fall within the range seen for the mouth movement frequencies reported from human audiovisual speech (2–7Hz; Chandrasekaran *et al.*, 2009).

Single-neuron responses to dynamic facial expressions in the STS

We presented two exemplars each of lipsmacks, teeth-grinds and yawns to monkeys performing a fixation task while we recorded from the upper bank of the STS, from a region just below primary auditory cortex and whose neurons both integrate faces and voices (Chandrasekaran & Ghazanfar, 2009) and interact with the lateral belt auditory cortex during that process (Ghazanfar *et al.*, 2008). Yawns were included as control stimuli to determine whether the region of STS we recorded from was selective for rhythmic facial expressions or sensitive to dynamic faces more generally. Like human yawns, macaque monkey yawns consist of opening the mouth to its maximal extent over a period of hundreds of ms; it is typically a single ballistic movement (Hinde & Rowell, 1962). Unlike humans, they are produced in situations of mild stress and in aggressive contexts.

We found that STS neurons were equally sensitive to all expression types. Figure 3A and B shows single neurons responding to the two lipsmack exemplars. As seen in Fig. 3A, there were multiple points of enhancement and suppression relative to baseline during the stimulus period, suggesting that this neuron was sensitive to the dynamics of the mouth. For example, at 80 ms after stimulus onset, firing rate was enhanced relative to baseline ($t_{18} = 2.71$, $P = 0.01$) while at 480 ms there was a significant suppression ($t_{18} = -2.84$, $P = 0.01$; Fig. 3A). Similarly, Figure 3C and D shows a similar pattern of neural firing in response to teeth-grinds. As illustrated in Fig. 3C, there was a significant enhancement of firing rate relative to baseline ($t_{18} = 2.24$, $P = 0.03$) at 90 ms and suppression at 590 ms ($t_{18} = -2.66$, $P = 0.01$). Finally, Fig. 3E and F show spiking activity in response to yawns. There was a significant and sustained enhancement of firing relative to baseline($t_{18} = 2.18$, $P = 0.04$) starting at 94 ms after stimulus onset, bracketed by periods of significant suppression relative to baseline (at 32 ms; $t_{18} = -2.49$, $P = 0.02$; Fig. 3E).

As a population, 33 out of 87 neurons (38%) responded to at least one of the six exemplars (Fig. 4A). For the 33 responsive neurons, Fig. 4B shows a Venn diagram which reports the percentage of neurons selective to one or more of the facial expressions. The largest

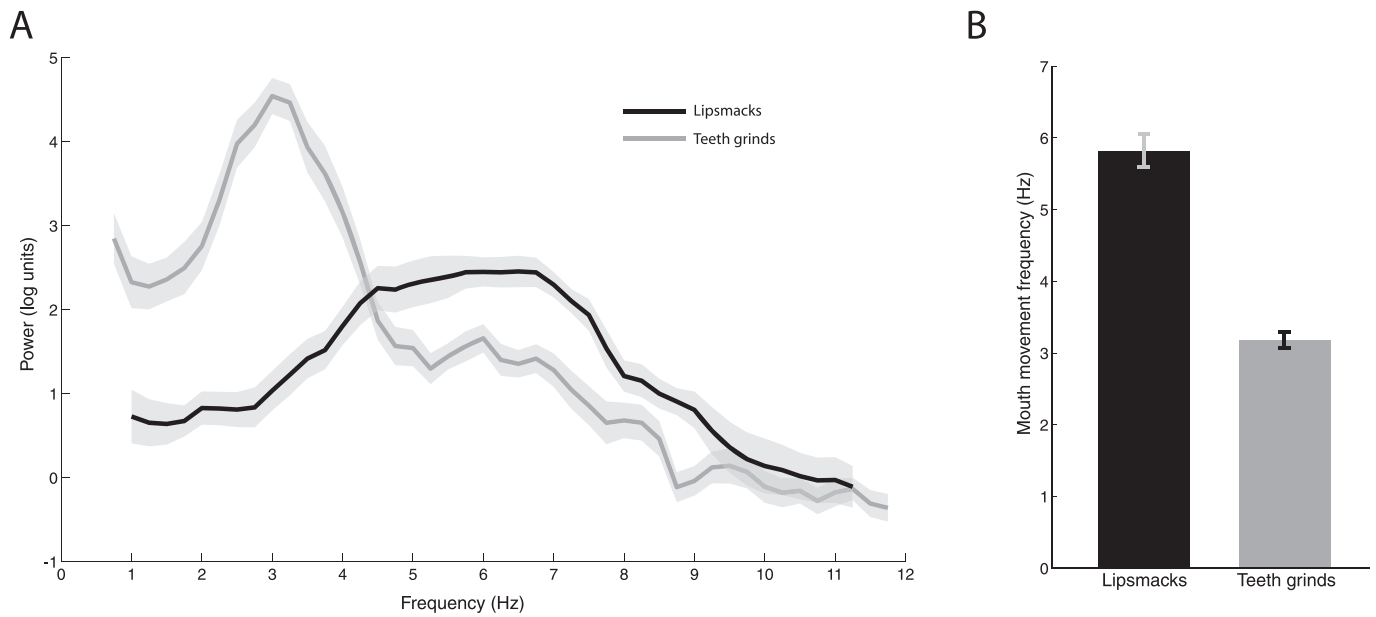

FIG. 2. Temporal modulation of visuofacial communicative signals. (A) Mean Fourier spectra of mouth displacement for all lipsmacks (blue, $n = 15$) and teeth-grinds ($n = 21$). *X*-axis shows frequency in Hz and *Y*-axis shows power in natural log units. Shaded regions denote SEM. (B) Mean peak density frequencies for lipsmacks (5.817 ± 0.233 Hz; SEM) and teeth grinds (3.185 ± 0.117 Hz, SEM). Error bars denote SEM.

proportion of neurons was responsive to all three dynamic facial expressions (42%). The next largest proportion was for neurons that responded to both lipsmacks and yawns (15%). Other proportions for different types of selectivity ranged from 6 to 12%. For example, four (12%) of the total number of face-sensitive neurons responded only to lipsmacks, while three (9%) responded only to teeth-grinds and another 9% only to yawns. Figure 4C shows the percentage of neurons that responded to each of the six gestures grouped according to the different expression categories. Figure 4D shows that, for these six exemplars, the majority of neurons were responsive to more than one facial gesture or identity.

These single-neuron data reveal that this region of the monkey STS, that shows face–voice integration and multisensory interactions with auditory cortex, is also sensitive, but not selective, to rhythmic and dynamic nonvocal facial expressions.

Local field potential responses to dynamic facial expressions

There are numerous studies that examine human fMRI-BOLD responses and event-related potential (ERP) responses to dynamic facial stimuli (Puce *et al.*, 1998, 2000, 2007; Miki *et al.*, 2004). To what extent local neuronal populations in the STS contribute to these responses is not known. LFP responses represent activity at an intermediate spatial scale that can provide a scaffold between the single neurons reported above and the data from human studies. Thus, we examined the raw LFP responses to these same gestures as well as the spectral structure of these LFP responses.

As the single-neuron data would imply, we found robust LFP responses to these dynamic facial expressions (Fig. 5A–C). The left panel of Fig. 5A shows LFP activity from a single STS site, averaged over 10 trials, in responses to a lipsmack. A triphasic response with significant deflections at 80 ms ($t_{18} = -2.40$, $P < 0.02$), 160 ($t_{18} = -7.36$, $P < 0.001$) and 250 ms ($t_{18} = 7.04$, $P < 0.\ 001$) was observed. The right panel (Fig. 5A) shows the average across all 45 cortical sites. The gray histogram shows the percentage of sites with significant deviations from baseline across post-stimulus time. Twenty-five per cent of cortical sites showed significant deviations at 80 ms, 70% at 160 ms and 75% at 250 ms. For teeth-grinds, the left panel of Fig. 5B also shows a triphasic response with the same temporal characteristics as responses to lipsmacks (80 ms, $t_{18} = -5.96$, $P < 0.001$; 160 ms, $t_{18} = 5.21$, $P < 0.001$; 250 ms, $t_{18} = 6.84$, $P < 0.001$). The right panel shows the population response of all cortical sites to teeth-grinds: significant deviations at 28% (80 ms), 49% (160 ms) and 75% (250 ms). For yawns, Fig. 5C shows activity at similar time points as the lipsmacks and teeth-grinds: at 80 ms, $t_{18} = -2.23$, $P = 0.03$; at 160 ms, $t_{18} = -7.33$; and at 250 ms, $t_{18} = 7.28$, $P < 0.001$. The left panel of Fig. 5C shows a single exemplar of an LFP site responding to a yawn with a triphasic response pattern, while the right panel shows the percentage of cortical sites with significant deviations from baseline across post stimulus time. For yawns, 31% of cortical sites showed significant deviations at 80 ms, 53% at 160 ms and 53% at 250 ms.

These data show that many more cortical sites show sensitivity in the LFP than in the single-neuron (spiking) data. The time course of these deflections also seem to match up reasonably well with those reported for the human ERP responses to dynamic faces in humans (after accounting for the 3/5 timing rule for human–monkey comparisons (Schroeder *et al.*, 2004)). Importantly, the temporal structure of these raw LFP responses do not seem to be related to the temporal structure of the facial expressions as rhythmic lipsmacks and teeth-grinds elicited the same temporal profile as the nonrhythmic yawns.

We next analyzed the frequency-band structure of LFP responses. Figure 6A–C shows the population-level wavelet spectrogram of responses to lipsmacks. For low-frequency ranges, enhancements were observed in the delta (0–3Hz), theta (3–8Hz) and alpha (8–14Hz) bands, while suppression was observed in the beta (20–40Hz) range (Fig. 6A). Gamma band activity was enhanced in a broad swath, ranging from 60 to 130Hz, and was sustained over the duration of the facial expression (Fig. 6B). Figure 6C shows the percentage of cortical sites with significant differences between stimulus-period activity (0 to 400 ms) and baseline (−200 to 0 ms) as a function of

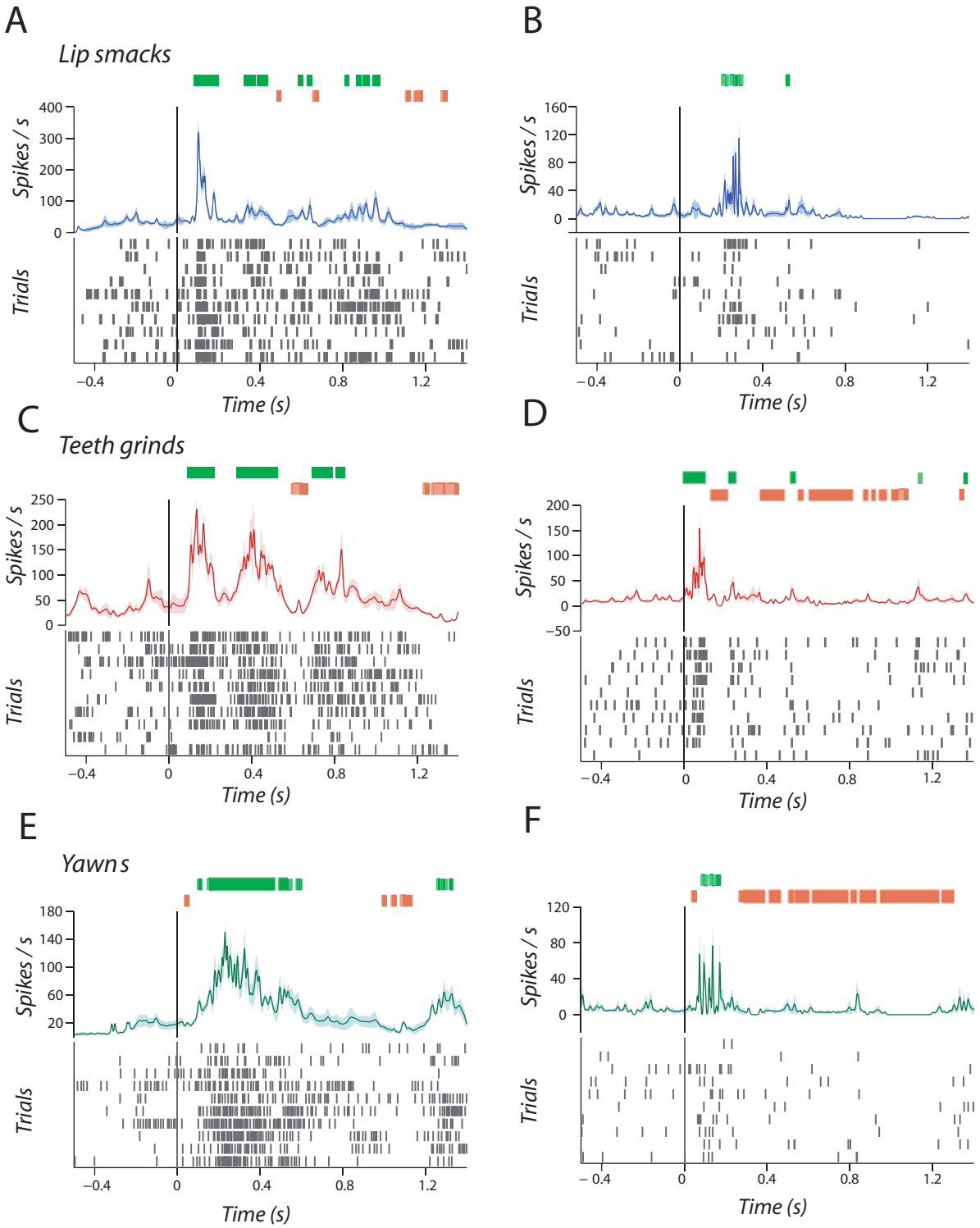

FIG. 3. Single-unit responses to rhythmic facial expressions. (A and B) Two examples of single units responding to lipsmacks in the form of peristimulus spike density functions. (C and D) Two examples of single units responding to teeth-grinds. (E and F) Two examples of single units responding to yawns. *X*-axis is time in seconds and *Y*-axis is spikes per second. Spike rasters are displayed below for each of the 10 trials. Green hatches above indicate significantly enhanced firing rates above baseline; orange hatches indicate significantly suppressed firing rates below baseline.

frequency. Power in delta, theta, alpha and gamma bands was enhanced for several cortical sites, whereas suppression was observed for the beta band. Teeth-grinds (Fig. 6E–H) and yawns (Fig. 6G–I) showed virtually identical patterns of responses.

Discussion

To investigate the evolutionary origins of audiovisual speech, we examined the temporal structure of rhythmic facial expressions in

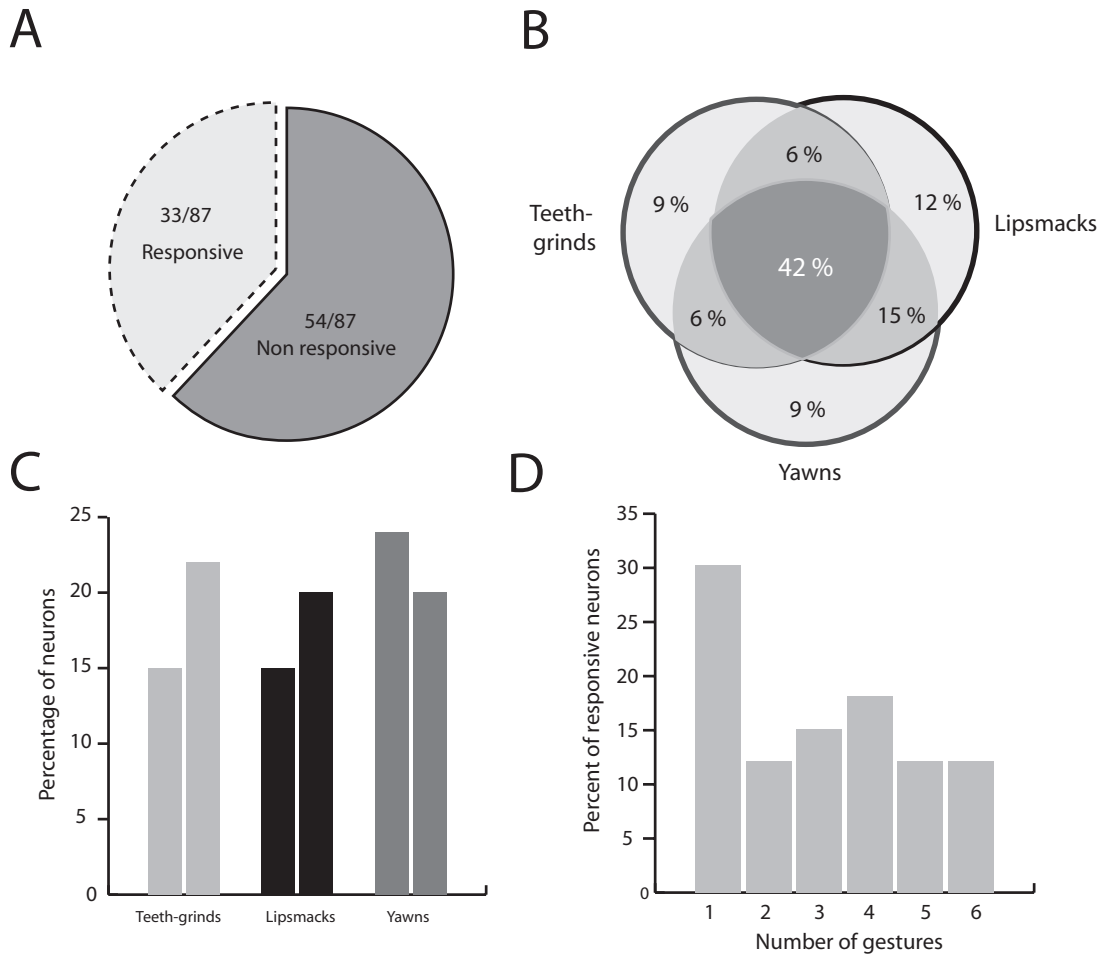

FIG. 4. Population data for single unit responses. (A) Proportion of neurons responding to the facial expressions. (B) Venn diagram of the percentage of responsive neurons showing selectivity of different types. (C) Percentage of neurons responding to each of the six exemplars of rhythmic facial expressions. (D) Selectivity of the neuronal population in terms of the percentage of neurons responding to 1–6 gestures.

macaque monkeys and the responses of STS neurons to such expressions. We found that both lipsmacks and teeth-grinds had consistent but distinct peak frequencies and that both fell well within the range of mouth movements associated with audiovisual speech. Single neurons and LFPs of the STS of monkeys responded to such facial dynamics, but were not selective for rhythmic facial expressions as they just as readily responded to yawns, a dynamic but nonrhythmic gesture. All expressions elicited enhanced power in the delta (0–3Hz), theta (3–8Hz), alpha (8–14Hz) and gamma (> 60 Hz) frequency ranges, and suppressed power in the beta (20–40Hz) range. We discuss these findings below.

Our hypothesis from the outset was that, if the rhythmic nature of audiovisual speech evolved from the rhythmic nonvocal facial expressions of ancestral primates (MacNeilage, 1998, 2008), then the closely-related macaque monkey should produce such facial expressions with mouth dynamics in the same frequency range: 2–7 Hz. We found this to be the case. Lipsmacks were produced with mouth movements occurring at a rate of ∼6Hz, and teeth-grinds at a rate of ∼3Hz. Remarkably, the rate of mouth movements between these two facial expressions were significantly different, suggesting different underlying physiological mechanisms. Lipsmacks may be produced at a faster rate than teeth-grinds for two reasons which are not mutually exclusive: differences in musculature (Waller *et al.*, 2008; Burrows *et al.*, 2009) and/or differences in subcortical and cortical motor sources of muscle activity (Sherwood *et al.*, 2004a,b, 2005; Sherwood, 2005). One important point to note is that the number and innervation of facial muscles in the macaque are nearly identical to that of both chimpanzee and human (Burrows *et al.*, 2009); thus, at least at the level of periphery, macaque monkeys have the potential to produce the same facial (but perhaps not lingual) articulatory movements as humans during speech.

The STS has long been established as a cortical node in the network involved in the integration of audiovisual communication signals, including human speech (Calvert *et al.*, 2000; Callan *et al.*, 2003; Calvert & Campbell, 2003; Wright *et al.*, 2003; Barraclough *et al.*, 2005; Reale *et al.*, 2007; Chandrasekaran & Ghazanfar, 2009; Dahl *et al.*, 2009). More recently, it's been suggested that STS is at least one of the sources driving multisensory responses to communication signals in the auditory cortex (Calvert, 2001; Ghazanfar *et al.*, 2005), and this hypothesis is supported by recent studies examining the physiological functional connectivity between the STS and auditory cortex in monkeys (Ghazanfar *et al.*, 2008; Kayser & Logothetis, 2009). In light of these data, it is important to establish (if our evolutionary hypothesis is true) whether the STS responds to rhythmic facial expressions, as that would establish that it was 'ready' for the advent of rhythmic audiovisual speech.

Previous single-unit and fMRI studies of monkey neocortex established that there are multiple regions in and around the STS

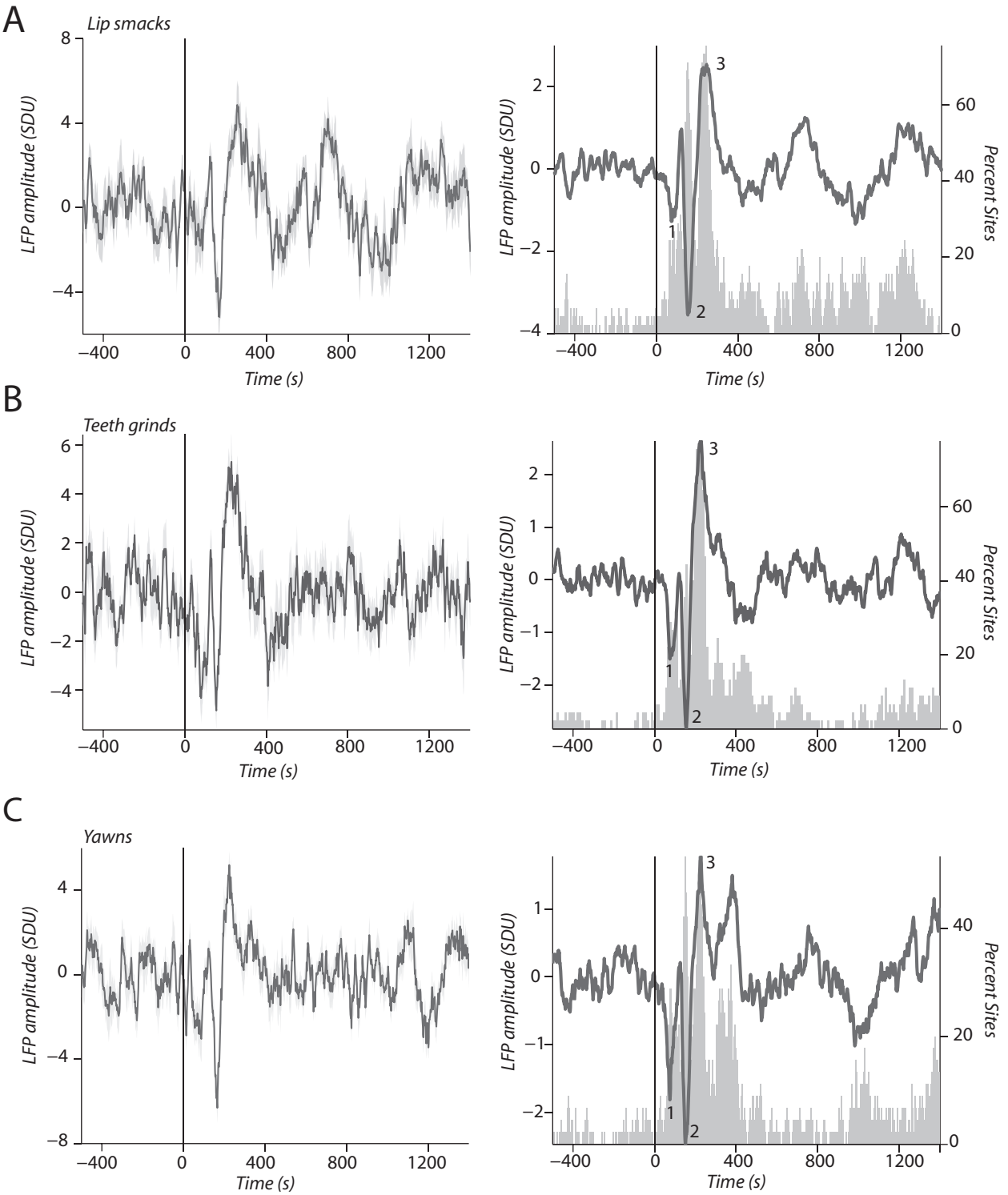

FIG. 5. Raw LFP responses to dynamic facial expressions. (A) LFP response to a lipsmack from a single cortical site (left panel) and population LFP response to lipsmacks (right panel). (B) LFP response to a teeth-grind from a single cortical site (left panel) and population LFP response to teeth-grinds (right panel). (C) LFP response to a yawn from a single cortical site (left panel) and population LFP response to yawns (right panel). *X*-axis is time in ms and *Y*-axis is amplitude in SD units. Gray histograms in right panels indicate the percentage of cortical sites (right *y*-axis) that showed significant deviations from baseline during that time interval. Labeled points 1, 2 and 3 indicate 80, 160 and 250 ms post-stimulus (see Results).

that are responsive to static faces, and these responses can be modulated by identity, expression, eye gaze and head orientation (Bruce *et al.*, 1981; Perrett *et al.*, 1982, 1985; Hasselmo *et al.*, 1989; Harries & Perrett, 1991; Tsao *et al.*, 2003; Eifuku *et al.*, 2004; De Souza *et al.*, 2005; Pinsk *et al.*, 2005, 2009). However, no studies of monkey STS have investigated whether dynamic facial expressions elicit responses from neurons in this structure. Furthermore, it should not be presumed that a cortical region that responds to static faces should automatically respond to dynamic faces. Different face regions have differential sensitivities to particular

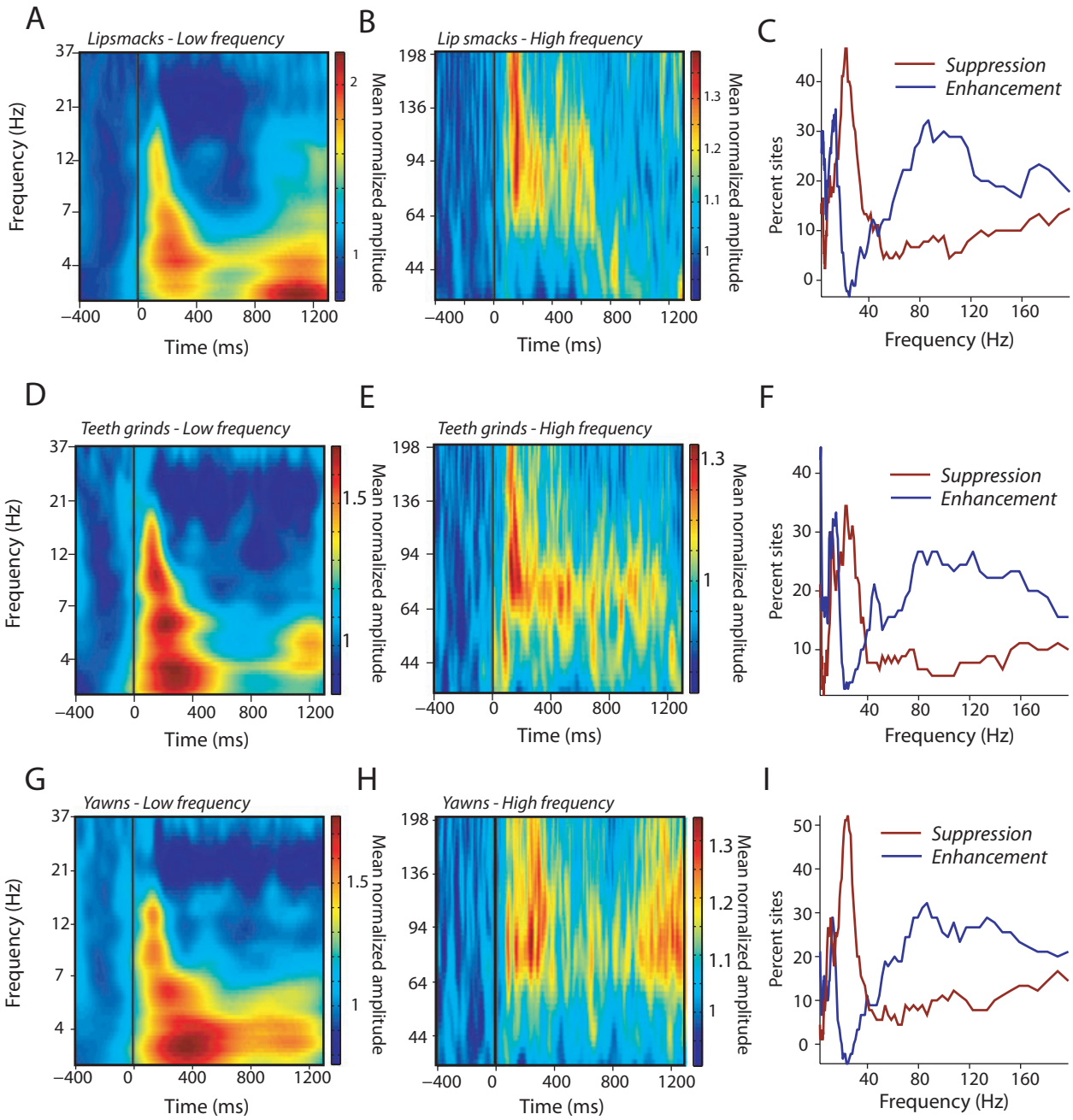

FIG. 6. Spectral analyses of population-level LFP responses to dynamic facial expressions (n = 45). (A) Low-frequency spectrograms of the mean LFP response to lipsmacks. (B) High-frequency spectrograms of the mean LFP responding to lipsmacks. *X*-axis is time in milliseconds and *Y*-axis is frequency in Hz; color bar indicates mean normalized amplitude. (C) Percentage of sites (*x*-axis) show suppression (red line) or enhancement (blue line) as a function of frequency (*y*-axis, in Hz) in response to lipsmacks. (D–I) Same type of plots as above, but for (D–F) responses to teeth-grinds and (G–I) responses to yawns.

features, expressions and their combination, etc. (Hasselmo *et al.*, 1989; Eifuku *et al.*, 2004; Freiwald *et al.*, 2009). Our data, recorded from the same region of STS that integrates faces and voices (Chandrasekaran & Ghazanfar, 2009) and interacts with auditory cortex during this process(Ghazanfar *et al.*, 2008), establish that a large proportion of single neurons do, indeed, respond to dynamic facial expressions and show little selectivity as to expression type (this is even more pronounced given that we used so few exemplars and three distinct expression types). In some cases, the spiking activity of these neurons seemed to have a rhythmic structure as well (see exemplars in Fig. 3A and C).

We observed a triphasic LFP response to moving faces. These data support prior studies of ERPs over temporal cortex in response to moving faces(Puce *et al.*, 1998, 2000, 2003). In response to dynamic facial gestures, we observed in the LFP a negative deflection at 160 ms after stimulus onset and a positive deflection at 250 ms after stimulus onset. ERP studies of moving faces have also found that two components, the N170 and P350, seem to be sensitive to facial motion (Puce *et al.*, 1998, 2000, 2003). In addition, robust ERP responses are observed even to degraded line drawings of faces (Puce *et al.*, 2003). Although our data do not precisely match the ERP data recorded in humans, they suggest that similar loci in both monkeys and humans

are involved in the processing of dynamic faces. Finally, our results show that LFPs are accompanied by robust spiking activity in some sites and thus confirms speculation in prior human studies (see Puce *et al.*, 2000, 2003) that suggested that the STS in monkeys should be sensitive to dynamic faces. Further experiments that disentangle the contribution of the onset and facial dynamics to both LFPs and spikes would help further our understanding dynamic face processing in the STS.

A spectral analysis of the LFP responses to dynamic faces revealed that they elicited enhanced power in the delta (0–3Hz), theta (3–8Hz), alpha (8–14Hz) and gamma (> 60 Hz) frequency ranges, and suppressed power in the beta (20–40Hz) range. Theta (4–8 Hz)- and alpha (8 –14 Hz)-band activity were suppressed below baseline at 250 ms after the appearance of the face, and this suppression persisted throughout the remaining duration of the dynamic facial expression. In contrast, the gamma band activity was robust and sustained for the entire duration of the moving face. This response pattern is exactly what we reported for STS responses to the facial components of vocal expressions (Chandrasekaran & Ghazanfar, 2009). For face–voice integration these different frequency bands have distinct properties: some show integration and others do not, depending on certain time variables (Chandrasekaran & Ghazanfar, 2009). In the present case, what different processes the distinct frequency bands elicited by dynamic facial expressions may be mediating is an open question. Oddly enough, to date we know of no human studies of band-limited EEG or MEG responses to dynamic faces.

In conclusion, the rhythmic, but nonvocal, facial expressions of macaque monkeys have a temporal structure that is similar to the structure of audiovisual speech (Munhall & Vatikiotis-Bateson, 1998; Chandrasekaran *et al.*, 2009). Though single-unit and LFP responses to rhythmic facial expressions were not selective (responding just as strongly to yawns), but merely sensitive to them, the data support the notion that the region of the STS from which we recorded is 'prepared' to process rhythmic audiovisual speech. This region, lying just below primary auditory cortex (see Materials and methods), is already known to integrate faces and voices and interact with lateral belt auditory cortex (Ghazanfar *et al.*, 2008; Chandrasekaran & Ghazanfar, 2009). It remains an open question whether this region of the STS is sensitive to mouth movements at a higher frequency, that is, whether it is constrained by its intrinsic circuitry or not. Testing this would require the use of avatars or synthetic faces (Steckenfinger & Ghazanfar, 2009), as the rhythmic structure of real monkey facial expressions do not move much faster than ~7 Hz. Taken together, the data from our study support the hypothesis that, during the course of human evolution, nonvocal rhythmic facial expressions were coupled to vocalizations (MacNeilage, 1998, 2008) and, further, that no brain-based structural elaborations or embellishments via natural selection were needed to process this newly-evolved rhythmic audiovisual communication signal.

Acknowledgements

The physiological data were collected at the Max Planck Institute for Biological Cybernetics in Tuebingen, Germany by A.A.G. This work was supported by the National Institutes of Health (NINDS) R01NS054898 (A.A.G.), the National Science Foundation BCS-0547760 CAREER Award (A.A.G.), Autism Speaks (A.A.G.) and Princeton University's Quantitative and Computational Neuroscience training grant NIH R90 DA023419-02 (C.C.).

Abbreviations

ERP, event-related potential; LFP, local field potential; STS, superior temporal sulcus.

References

Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W. & Perrett, D.I. (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.*, **17**, 377–391.

Bruce, C., Desimone, R. & Gross, C.G. (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the Macaque. *J. Neurophysiol.*, **46**, 369–384.

Burrows, A.M., Waller, B.M. & Parr, L.A. (2009) Facial musculature in the rhesus macaque (Macaca mulatta): evolutionary and functional contexts with comparisons to chimpanzees and humans. *J. Anat.*, **215**, 320–334.

Callan, D.E., Jones, J.A., Munhall, K., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2003) Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, **14**, 2213–2218.

Calvert, G.A. (2001) Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex*, **11**, 1110–1123.

Calvert, G.A. & Campbell, R. (2003) Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.*, **15**, 57–70.

Calvert, G.A., Campbell, R. & Brammer, M.J. (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.*, **10**, 649–657.

Campbell, R. (2008) The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **363**, 1001–1010.

Chandrasekaran, C. & Ghazanfar, A.A. (2009) Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *J. Neurophysiol.*, **101**, 773–788.

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

Dahl, C.D., Logothetis, N.K. & Kayser, C. (2009) Spatial organization of multisensory responses in temporal association cortex. *J. Neurosci.*, **29**, 11924–11932.

De Souza, W.C., Eifuku, S., Tamura, R., Nishijo, H. & Ono, T. (2005) Differential characteristics of face neuron responses within the anterior superior temporal sulcus of macaques. *J. Neurophysiol.*, **94**, 1252–1266.

Drullman, R., Festen, J.M. & Plomp, R. (1994) Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, **95**, 2670–2680.

Eifuku, S., De Souza, W.C., Tamura, R., Nishijo, H. & Ono, T. (2004) Neuronal correlates of face identification in the monkey anterior temporal cortical areas. *J. Neurophysiol.*, **91**, 358–371.

Freiwald, W.A., Tsao, D.Y. & Livingstone, M.S. (2009) A face feature space in the macaque temporal lobe. *Nat. Neurosci.*, **12**, 1187–1196.

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

Ghazanfar, A.A., Chandrasekaran, C. & Logothetis, N.K. (2008) Interactions between the Superior Temporal Sulcus and Auditory Cortex Mediate Dynamic Face/Voice Integration in Rhesus Monkeys. *J. Neurosci.*, **28**, 4457–4469.

Giraud, A.L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S.J. & Laufs, H. (2007) Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, **56**, 1127–1134.

Greenberg, S., Carvey, H., Hitchcock, L. & Chang, S. (2003) Temporal properties of spontaneous speech–a syllable-centric perspective. *J. Phon.*, **31**, 465–485.

Hackett, T.A., Stepniewska, I. & Kaas, J.H. (1998) Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *J. Comp. Neurol.*, **394**, 475–495.

Harries, M.H. & Perrett, D.I. (1991) Visual processing of faces in temporal cortex - physiological evidence for a modular organization and possible anatomical correlates. *J. Cogn. Neurosci.*, **3**, 9–24.

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

Hinde, R.A. & Rowell, T.E. (1962) Communication by posture and facial expressions in the rhesus monkey (Macaca mulatta). *Proc. Zool. Soc. Lond.*, **138**, 1–21.

Kayser, C. & Logothetis, N.K. (2009) Directed interactions between auditory and superior temporal cortices and their role in sensory integration. *Front. Integr. Neurosci.*, **3**, 7.

Kim, J. & Davis, C. (2004) Investigating the audio-visual speech detection advantage. *Speech. Commun.*, **44**, 19–30.

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

Lieberman, P. & Blumstein, S.E. (1988) *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press, Cambridge.
Logothetis, N.K., Merkle, H., Augath, M., Trinath, T. & Ugurbil, K. (2002) Ultra high-resolution fMRI in monkeys with implanted RF coils. *Neuron*, **35**, 227–242.
Luo, H. & Poeppel, D. (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, **54**, 1001–1010.
MacNeilage, P.F. (1998) The frame/content theory of evolution of speech production. *Behav. Brain. Sci.*, **21**, 499–511.
MacNeilage, P.F. (2008) *The Origin of Speech*. Oxford University Press, Oxford, UK.
Miki, K., Watanabe, S., Kakigi, R. & Puce, A. (2004) Magnetoencephalographic study of occipitotemporal activity elicited by viewing mouth movements. *Clin. Neurophysiol.*, **115**, 1559–1574.
Munhall, K. & Vatikiotis-Bateson, E. (1998) The moving face during speech communication. In Campbell, R., Dodd, B. & Burnham, D. (Eds), *Hearing by Eye II*. Taylor and Francis, Sussex, pp. 123–139.
Ohala, J. (1975) Temporal Regulation of Speech. In Fant, G. & Tatham, M.A.A. (Eds), *Auditory Analysis and Perception of Speech*. Academic Press, London, pp. 431–453.
Perrett, D.I., Rolls, E.T. & Caan, W. (1982) Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, **47**, 329–342.
Perrett, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D. & Jeeves, M.A. (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond., B, Biol.*, **223**, 293–317.
Pfingst, B.E. & O'Connor, T.A. (1980) Vertical Stereotaxic Approach to Auditory-Cortex in the Unanesthetized Monkey. *J. Neurosci. Methods*, **2**, 33–45.
Pinker, S. & Bloom, P. (1990) Natural language and natural selection. *Behav. Brain. Sci.*, **13**, 707–784.
Pinsk, M.A., DeSimone, K., Moore, T., Gross, C.G. & Kastner, S. (2005) Representations of faces and body parts in macaque temporal cortex: a functional MRI study. *Proc. Natl Acad. Sci. USA*, **102**, 6996–7001.
Pinsk, M.A., Arcaro, M., Weiner, K.S., Kalkus, J.F., Inati, S.J., Gross, C.G. & Kastner, S. (2009) Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study. *J. Neurophysiol.*, **101**, 2581–2600.
Poeppel, D. (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech. Commun.*, **41**, 245–255.
Puce, A., Allison, T., Bentin, S., Gore, J.C. & McCarthy, G. (1998) Temporal cortex activation in humans viewing eye and mouth movements. *J. Neurosci.*, **18**, 2188–2199.
Puce, A., Smith, A. & Allison, T. (2000) ERPs evoked by viewing facial movements. *Cognitive Neuropsychology*, **17**, 221–239.
Puce, A., Syngeniotis, A., Thompson, J.C., Abbott, D.F., Wheaton, K.J. & Castiello, U. (2003) The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. *Neuroimage*, **19**, 861–869.
Puce, A., Epling, J.A., Thompson, J.C. & Carrick, O.K. (2007) Neural responses elicited to face motion and vocalization pairings. *Neuropsychologia*, **45**, 93–106.
Reale, R.A., Calvert, G.A., Thesen, T., Jenison, R.L., Kawasaki, H., Oya, H., Howard, M.A. & Brugge, J.F. (2007) Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience*, **145**, 162–184.
Redican, W.K. (1975) Facial expressions in nonhuman primates. In Rosenblum, L.A. (ed), *Primate Behavior: Developments in Field and Laboratory Research*. Academic Press, New York, pp. 103–194.
Saberi, K. & Perrott, D.R. (1999) Cognitive restoration of reversed speech. *Nature*, **398**, 760.
Schroeder, C.E., Molholm, S., Lakatos, P., Ritter, W. & Foxe, J.J. (2004) Human-simian correspondence in the early cortical processing of multisensory cues. *Cogn. Process.*, **5**, 140–151.
Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S. & Puce, A. (2008) Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci.*, **12**, 106–113.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.
Sherwood, C.C. (2005) Comparative anatomy of the facial motor nucleus in mammals, with an analysis of neuron numbers in primates. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.*, **287A**, 1067–1079.
Sherwood, C.C., Holloway, R.L., Erwin, J.M. & Hof, P.R. (2004a) Cortical orofacial motor representation in old world monkeys, great apes, and humans – II. Stereologic analysis of chemoarchitecture. *Brain Behav. Evol.*, **63**, 82–106.
Sherwood, C.C., Holloway, R.L., Erwin, J.M., Schleicher, A., Zilles, K. & Hof, P.R. (2004b) Cortical orofacial motor representation in old world monkeys, great apes, and humans – I. Quantitative analysis of cytoarchitecture. *Brain Behav. Evol.*, **63**, 61–81.
Sherwood, C.C., Hof, P.R., Holloway, R.L., Semendeferi, K., Gannon, P.J., Frahm, H.D. & Zilles, K. (2005) Evolution of the brainstem orofacial motor system in primates: a comparative study of trigeminal, facial, and hypoglossal nuclei. *J. Human Evol.*, **48**, 45–84.
Smith, Z.M., Delgutte, B. & Oxenham, A.J. (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, **416**, 87–90.
Steckenfinger, S.A. & Ghazanfar, A.A. (2009) Monkey visual behavior falls into the uncanny valley. *Proc. Natl. Acad. Sci. USA*, **106**, 18362–18466.
Sugihara, T., Diltz, M.D., Averbeck, B.B. & Romanski, L.M. (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *J. Neurosci.*, **26**, 11138–11147.
Szucs, A. (1998) Applications of the spike density function in analysis of neuronal firing patterns. *J. Neurosci. Methods*, **81**, 159–167.
Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B. & Tootell, R.B.H. (2003) Faces and objects in macaque cerebral cortex. *Nat. Neurosci.*, **6**, 989–995.
Vitkovitch, M. & Barber, P. (1994) Effect of video frame rate on subjects' ability to shadow one of two competing verbal passages. *J. Speech Hear. Res.*, **37**, 1204–1210.
Vitkovitch, M. & Barber, P. (1996) Visible speech as a function of image quality: effects of display parameters on lipreading ability. *Appl. Cogn. Psychol.*, **10**, 121–140.
Waller, B.M., Parr, L.A., Gothard, K.M., Burrows, A.M. & Fuglevand, A.J. (2008) Mapping the contribution of single muscles to facial movements in the rhesus macaque. *Physiol. Behav.*, **95**, 93–100.
Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J. & McCarthy, G. (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex*, **13**, 1034–1043.