Supporting Information

Perrodin et al. 10.1073/pnas.1412817112

SI Materials and Methods

Visual Stimuli. The videos were acquired at 25 frames per second (640 × 480 pixels), 24-bit resolution, and compressed using Indeo video5. The stimuli were filmed while monkeys spontaneously vocalized, seated in a primate chair. All videos were recorded in the same sound-attenuated booth with the same lighting configuration, ensuring each video had similar auditory and visual background. We selected the stimuli to ensure the callers' head position and eye gaze direction were similar across all videos played within one experimental run. Finally, the faces were centered in the images, and the head size was matched for all callers in a given experimental run to occupy similar portions of the visual field. Movie clips were cropped at the beginning of the first mouth movement, with the first frame of each video showing the neutral facial expression. A dynamic mask and uniform black background were placed around the callers' faces to crop all but the moving facial features, so that the entire face was visible while the back of the head and neck were masked. Image contrast and luminance for each channel (RGB) was normalized in all videos using Adobe Photoshop CS2. The vocalization stimuli were split into two experimental runs for presentation. The video clips were 960 and 760 ms in duration, respectively for the two experimental sets (see ref. 1 for more details).

Auditory Stimuli. The audio tracks were acquired at 48 kHz and 16 bits resolution in stereo (PCM format). The vocalization sounds were matched in average RMS energy using MATLAB (MathWorks) scripts. All sounds were stored as WAV files, amplified using a Yamaha amplifier (AX-496), and delivered from two free-field speakers (JBL Professional), which were positioned at ear level 70 cm from the head and 50° to the left and right. Sound presentation was calibrated using a condenser microphone (4188; Brüel & Kjær) and sound level meter (2238 Mediator; Brüel & Kjær) to ensure a linear (±4 dB) transfer function of sound delivery (between 88 and 20 kHz). The intensity of all of the sounds was calibrated at the position of the head to be presented at an average intensity of 65 dB sound pressure level. The duration of the auditory vocalizations was, on average, 402 ± 111 ms (mean ± SD; range: 271–590 ms).

Visual Fixation Task. Recordings were performed in a darkened and sound-insulated booth (Illtec; Illbruck Acoustic GmbH) while the animals sat in a primate restraint chair in front of a 21-inch color monitor. The stimuli and stimulus conditions (such as modality) were randomly selected for presentation. The animals were required to restrict their eye movements to a certain visual fixation window within the video frame around the central spot for the entire duration of the trial. The eye position was measured using an infrared eye-tracking system (iView X RED P/T; SensoMotoric Instruments GmbH). During the stimulation period, a visual stimulus (video sequence only), an auditory stimulus (audio track only, black screen), or an audiovisual stimulus was presented. Successful completion of a trial resulted in a juice reward. A trial began with the appearance of a central fixation spot. Once the animal engaged in the central fixation task, data acquisition started. A trial consisted of an initial 500-ms baseline period, followed by a 1,200-ms stimulation period and a 300-ms poststimulus recording time. Intertrial intervals were at least 1,800 ms. The duration of the stimulation period was chosen to encompass the longest stimuli (960 ms), to ensure that the timing was consistent across different behavioral trials. The visual stimuli (dynamic, vocalizing primate faces) covered a visual field with a 15° diameter.

Monkey 1 performed visual fixation during single trials at a time (2 s), within a 4° diameter fixation window. This subject was scanned anesthetized in the prior fMRI experiment used to localize his voice-sensitive cluster (2). Monkey 2 previously had his anterior voice-area localized with fMRI while conducting a visual fixation task. Because this macaque was accustomed to working on longer fixation trials with more lenient fixation criterion, for this project this subject was allowed to browse the area within which the visual stimuli were presented on the monitor (four to six consecutive trials, 8–12 s, 8°- to 20°-diameter fixation window), aborting the trial if eye movements breached this area.

Recording Procedures. A combination of neurological targeting software, fMRI voice vs. nonvoice localizers, stereotactic coordinates of the voice cluster centers, and postmortem histology at the end of the experiments were all used to guide the electrophysiological recording electrodes to the voice-sensitive clusters in each animal or ascertain the position of the recording sites. The coordinates of each electrode along the anteroposterior (AP) and mediolateral (ML) axes were noted, as were the angle of the grid and the depth of the recording sites. Experimental recordings were initiated if at least one electrode had LFPs or neurons that could be driven by any of a large set of search sounds, including tones, frequency modulated sweeps, band-passed noise, clicks, musical samples, and other natural sounds from a large library. No attempt was made to select neurons with a particular response preference and any neuron or LFP site that seemed responsive to sound was recorded. Once a responsive site was isolated, the experiment began. After data collection was completed each electrode was advanced at least 250 μm to a new recording site and until the neuronal activity pattern considerably changed.

Sites in the auditory cortex were distinguished from deeper recording sites in the upper bank of the STS using the depth of the electrodes, the crossing of the lateral sulcus that is devoid of neuronal activity, the occurrence of over 2 mm of white matter between auditory cortex and STS, and the emergence of visual evoked potentials at deeper recording sites.

Electrophysiological Data Analysis. To increase statistical power, for the spiking activity results we combined single and multiunit clusters for analysis. In any case, we confirmed that the main cyclic pattern of results reported in Fig. 2*B*, although underpowered, was also evident in well-isolated single units from the dataset (Fig. S5).

When computing spiking response amplitudes, we selected the 400-ms peak response-centered window to capture the variability of individual neurons' response profiles. The length of the response window was chosen to match with the average duration of the sounds in this experiment. The results were largely comparable to those from using a shorter 200-ms response window (Fig. S4). In addition, the multisensory effects in the spiking response profiles were found to be highly stable over time (Fig. S3). Thus, measuring multisensory enhancement/suppression in the broader 400-ms window seems to satisfactorily capture the effects.

Multisensory Interactions. Nonlinear multisensory units whose response to the audiovisual stimulus significantly differed from the sum of both unimodal responses were identified using

a randomization procedure: A pool of all possible summations (n = #trials * #trials) of trial-based auditory and visual responses for a given stimulus was created. A bootstrapped distribution of trial-averaged, summed unimodal responses was built by averaging n = #trials randomly sampled trial-based values of A+V responses from the pool and repeating this for n = 1,000 iterations. Units for which the trial-averaged audiovisual (AV) response was sufficiently far from the bootstrapped distribution of summed unimodal (A + V) responses (z test, $P < 0.05$) were termed nonadditive (nonlinear) multisensory. False discovery rate correction for multiple comparisons was applied to all P values (3).

The direction and amplitude of the deviation from additivity was quantified using the following index: Additivity = 100*[AV − (A + V)]/(A + V), where A, V, and AV reflect the baseline-corrected response amplitude, averaged in the response window. Positive (negative) values of the additivity index indicate superadditive/enhanced (subadditive/suppressed) multisensory interactions.

Information Associated with Fig. S1: Neuronal Responsiveness and Visual Modulation in Response to Stimuli with Mid- vs. Long-Range VA Delays

In this figure we show that the magnitude of the unisensory response and the prominence of visual modulation were comparable for stimuli with midrange and long VA delays (Fig. S1). At the level of individual units we found that the subset of stimuli with midrange and long VA delays were as effective in eliciting auditory responses (59 units responding to the two stimuli with midrange VA delays and 64 to those with long VA delays). The proportion of visually modulated neurons (25% and 26% for midrange and long VA delays, respectively) was similar for both stimulus types (χ^2 test on the number of nonlinear modulated units, $P > 0.05$). However, the two subsets of stimuli triggered different types of visual influences: Compared with the calls with midrange VA delays, those with long VA delays elicited more audiovisual suppression (χ^2 test on the number of enhanced and suppressed multisensory units, $P = 0.0070$, $\chi^2 = 7.275$; Fig. S1*A*).

Similarly, when comparing the response amplitudes across the population of units for midrange vs. long VA delays ($n = 84$ units, $P > 0.05$), we found no differences in the auditory, visual, or audiovisual responses (paired-sample t test on response amplitudes, all $P > 0.05$) or the magnitude of the visual modulation (paired-sample t test on abs[AV − (A + V)], $P = 0.95$; Fig. S1*B*). However, stimuli with longer VA delays were more likely to elicit audiovisual suppression than the ones with midrange VA delays (paired-sample t test on the amplitude of nonrectified visual modulation AV − (A + V), $P = 0.04$; Fig. S1*C*). These data suggest qualitatively similar auditory and audiovisual processing of calls with different VA delays, with the quantitative difference residing in the type of audiovisual interactions.

Information Associated with Fig. S2: Topographic Organization of Multisensory Responses

To assess whether units showing more enhancement/suppression cluster in different anatomical locations within voice-sensitive cortex, we plotted the spatial distribution of multisensory response types for each of the two animals studied (Fig. S2). Sensory responsive units were distributed across voice-sensitive cortex. Multisensory units were more scattered throughout. Moreover, both enhanced and suppressed units were found in relatively uniform distributions at various electrode penetration sites in both animals, and there was no obvious topographic pattern.

Information Associated with Fig. S3: Time Stability of Multisensory Spiking Responses

To quantify the extent to which multisensory neurons show consistent enhancement, suppression, or a dynamically varying mixture of both types of effects over time, we computed the proportion of time bins during sound presentation in which the direction of the multisensory effect was consistent with the global direction captured by our measure in a 400-ms response window. Across the population of visually modulated units ($n = 81$ units), the multisensory effect direction was consistent with our global measure in at least 68% of time points (average 94%) across different time windows ranging from 5 to 200 ms (Fig. S3). This time-resolved analysis suggests that the direction of the multisensory modulation in the spiking response profiles is stable over time for each unit and is reliably captured by the measure of multisensory effect direction in a 400-ms window centered on each neuron's peak sensory response.

Information Associated with Fig. S8: No Consistent Stimulus Specificity of Phase-Resetting and Multisensory Effects

We first asked whether the phase-resetting effect is a general (stimulus-nonspecific) process or shows evidence for being stimulus-specific—for instance, whether conspecific monkey faces would elicit a stronger increase in phase coherence than heterospecific human faces. We found very similar response patterns elicited by human and monkey faces (Fig. S8*A*) that did not differ significantly (n = 52 sites, 100-ms successive time bins, t test, $P > 0.05$ uncorrected). This suggests that the phase reset of ongoing oscillations is comparable for faces from different primate species.

Next we looked more generally into the voice specificity of the multisensory effect, by comparing the direction of multisensory interactions in response to faces paired with intact vocalizations, and to the same faces paired with phase-scrambled versions of the original vocalizations (i.e., an acoustically degraded version of the vocalizations that preserves the overall frequency spectrum but eliminates all of the temporal envelope information). Because the face is identical in both cases and the VA delay remains constant, specificity to an intact vocalization would be indicated by deviations from the proportions of enhanced vs. suppressed units as predicted by the original VA-delay pattern in the main text (Fig. 2*B*). We found that for one voice–face pair the direction of multisensory interactions was similar across the intact and phase-scrambled pairs (coo, Fig. S8*B*, first two columns, χ^2 test, $P > 0.05$), whereas for the other stimulus the proportion of multisensory enhancement vs. suppression significantly differed across pairs with the intact and the manipulated vocalization (grunt, Fig. S8*B*, last two columns, χ^2 test, $P = 1.99 \times 10^{-5}$, X = 18.20). This suggests little, or at least inconsistent, specificity of the multisensory effect for intact vocalizations vs. other types of sounds. This is in agreement with previous data (1) indicating that multisensory responses in the voice area occur in response to different stimuli, including mismatched voice–face pairs.

Together, both findings indicate that the visually evoked phase reset and the subsequent multisensory modulation of spiking responses are a general rather than voice- or face-specific mechanism.

1. Perrodin C, Kayser C, Logothetis NK, Petkov CI (2014) Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *J Neurosci* 34(7): 2524–2537.
2. Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. *Curr Biol* 21(16):1408–1415.
3. Benjamini YHY (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300.

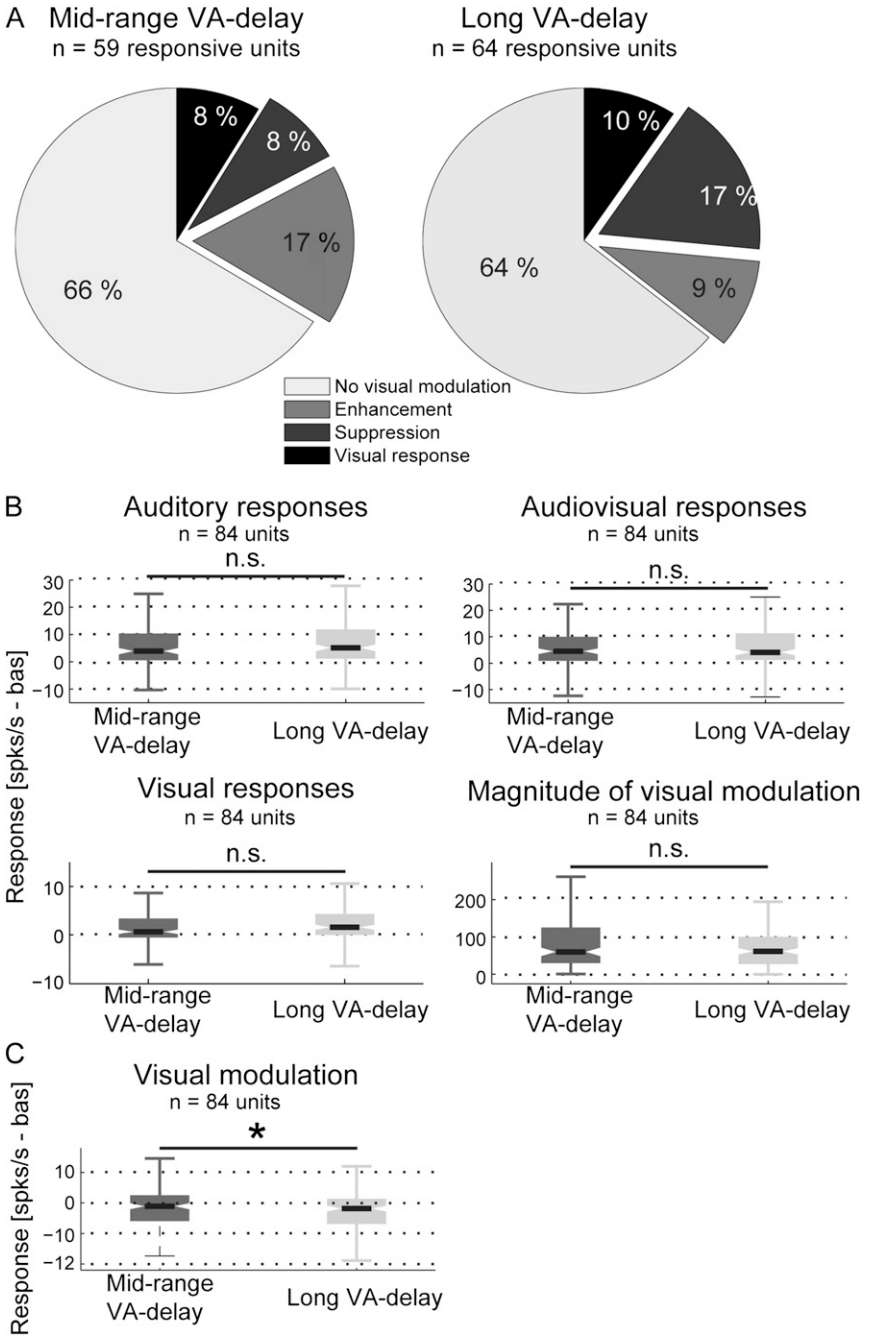

Fig. S1. Responsiveness and visual modulation in response to stimuli with midrange vs. long VA delays. (*A*) Summary of the type of sensory spiking responses in the anterior voice-sensitive supratemporal plane, for each subset of stimuli ($n = 59$ sensory responsive units for midrange VA delays and $n = 64$ sensory responsive units for long VA delays, respectively). (*B*) Auditory, visual, and audiovisual response amplitudes, as well as the magnitude of the visual modulation (abs[AV − (A + V)]) were comparable in response to stimuli with midrange vs. long VA delays. (*C*) The nonrectified visual modulation values differed in response to vocalizations with midrange vs. long VA delays, with more negative values for long VA delays. The box plots represent the median, upper, and lower quartiles of the spiking response amplitudes across the population of auditory responsive units ($n = 84$ units). *$P < 0.05$; n.s., not significant.

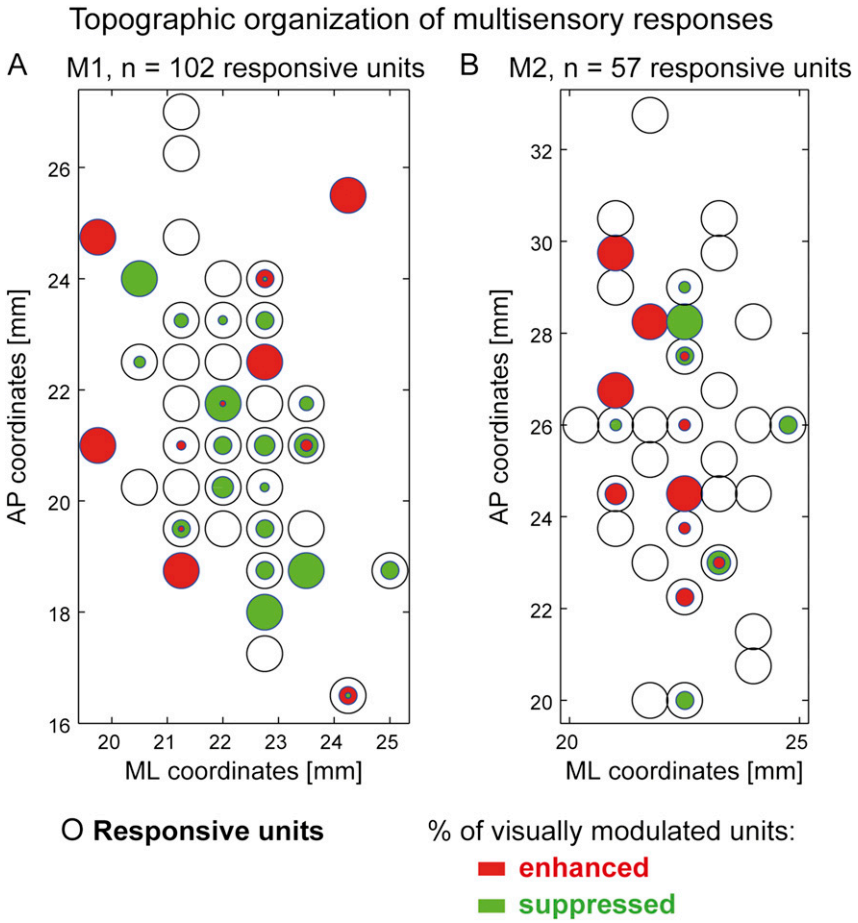

Fig. S2. Topographic analysis of multisensory responses. (*A* and *B*) Spatial organization of multisensory spiking responses, displayed using the AP and ML coordinates of the electrophysiological recording sites spanning the anterior voice-sensitive area in both animals. The stereotaxic coordinates used the Frankfurt-zero standard, where the origin is defined as the midpoint of the interaural line and the infraorbital plane. Black circles indicate the total number of responsive units encountered along electrode penetrations in a given location. The colored areas represent the percentage of units with significant audiovisual interactions (red, multisensory enhanced units; green, multisensory suppressed units).

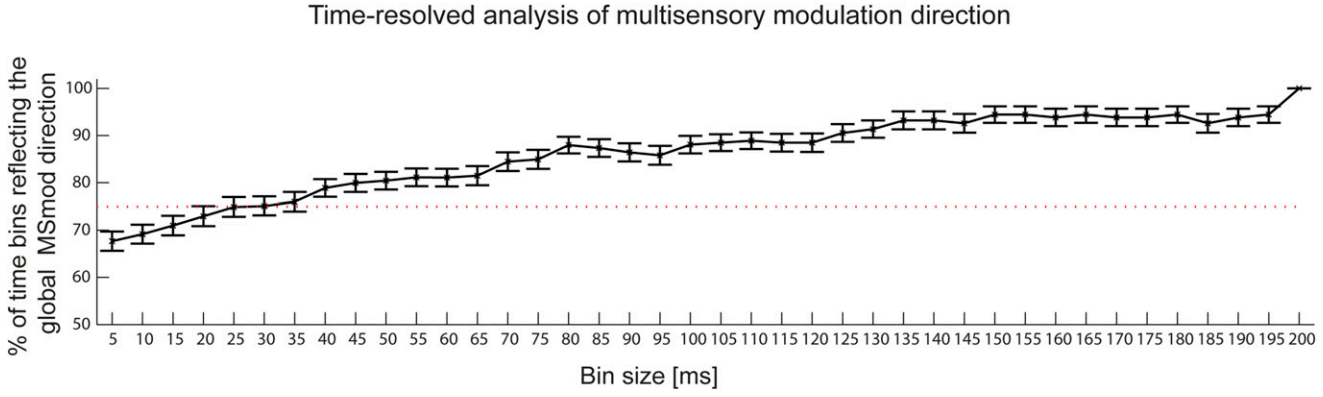

Fig. S3. Time stability of multisensory effect direction. Proportion of bins during which the multisensory (enhanced/suppressed) effect direction was consistent with the direction calculated in a 400-ms window. Shown is the mean ± SEM ($n = 81$ visually modulated units). The red dotted line marks 75% of bins reflecting the global effect direction.

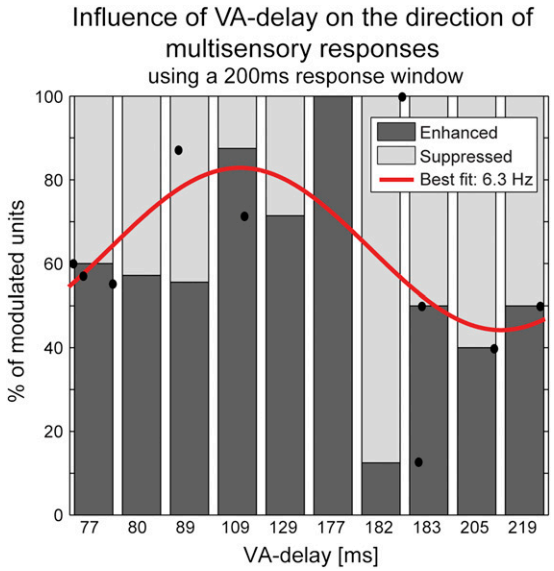

Fig. S4. Pattern of multisensory responses as a function of VA delay, calculated using a 200-ms neuronal response window. Proportions of enhanced and suppressed multisensory units by stimulus, arranged with increasing VA delays (n = 66 visually modulated units). Note that the bars are spaced at equidistant intervals for display purposes, thus forming a discrete subsampling of the actual VA delay values (dots). Black dots indicate the proportion of enhanced responses for each VA delay value, while respecting the real relative positions of VA delay values. The red line represents the sinusoid with the best-fitting frequency (6.3 Hz, adjusted R^2 = −0.094).

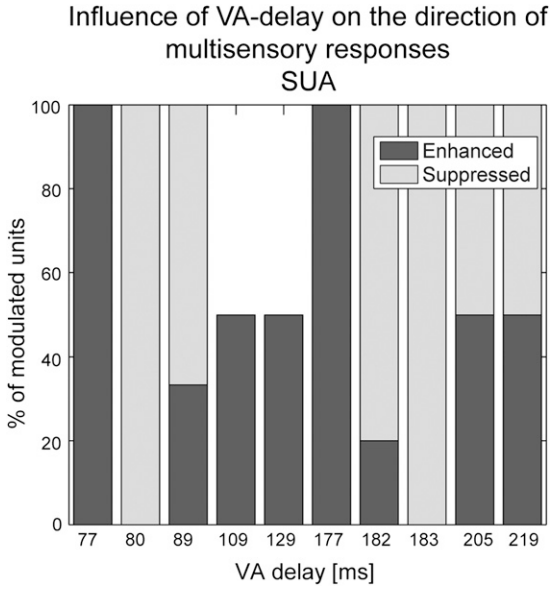

Fig. S5. Pattern of multisensory interactions as a function of VA delay in well-isolated single units. Proportions of enhanced and suppressed multisensory units by stimulus, arranged with increasing VA delays (n = 29 single unit activity).

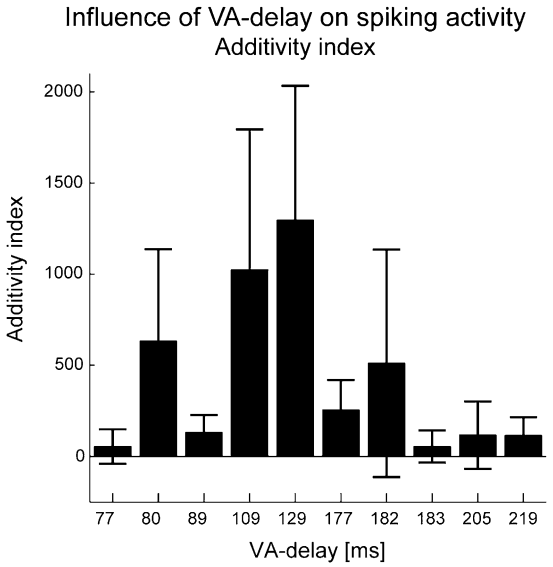

Fig. S6. Pattern of multisensory interactions as a function of VA delay, calculated using a nonbinary multisensory metric. Additivity index values averaged across nonlinear multisensory units by stimulus, arranged with increasing VA delays (n = 81 units). Shown is the mean ± SEM.

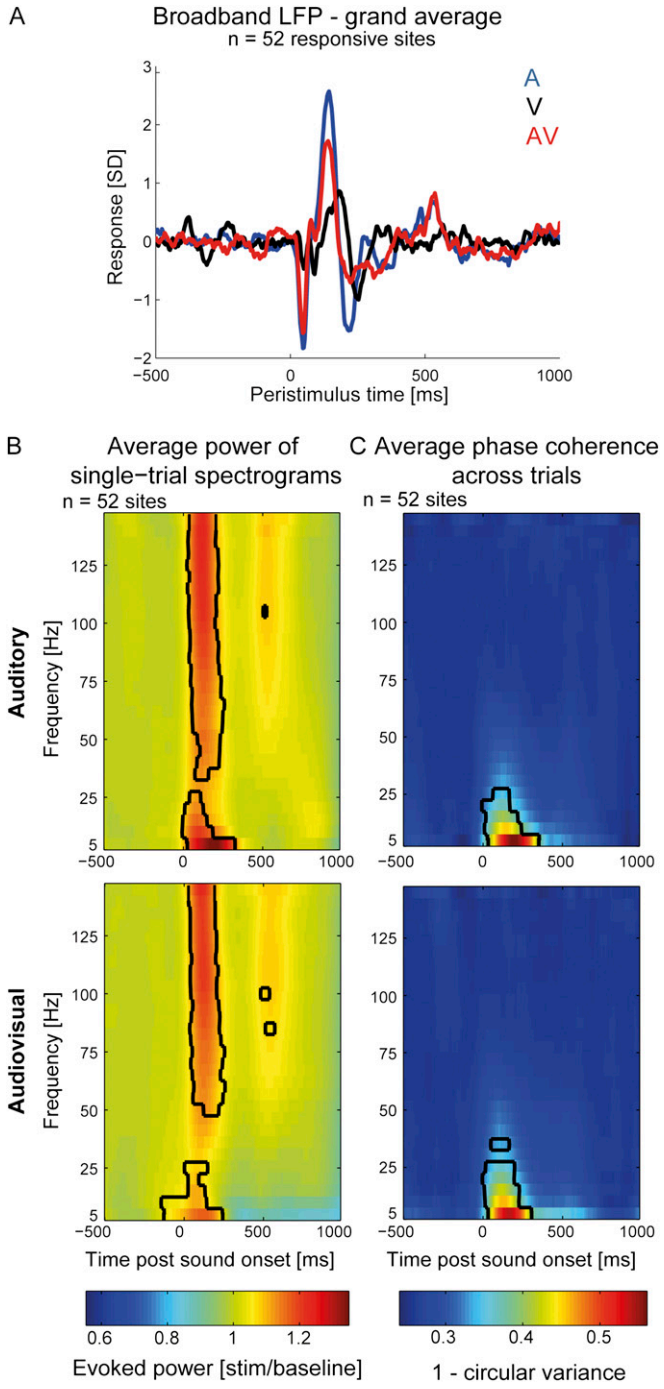

Fig. S7. Grand average broadband evoked potential and oscillatory context evoked by auditory and audiovisual stimulation. (*A*) Time course of broadband-evoked LFP in response to stimulation in the primary (auditory, A), nondominant (visual, V), and combined (audiovisual, AV) sensory conditions. The traces are sensory responses averaged across the four stimuli studied (two with long VA delays, two with midrange VA delays), and all recording sites that contained responsive units. Time t = 0 indicates the onset of the stimulus in the relevant modality (sound onset for auditory and audiovisual modalities, and video onset for visual). Note: The different VA delays of individual vocalizations were compensated for when averaging auditory and audiovisual responses, so that vocalization onsets are aligned across stimuli. In reality, the sound onset is shifted by t = 0 + VA delay for each vocalization. (*B*) Time-frequency plot of averaged single-trial spectrograms in response to auditory and audiovisual stimulation. The population-averaged spectrogram has been baseline-normalized for display purposes. (*C*) Time-frequency plot of average phase coherence values across trials. The color code reflects the strength of phase alignment evoked by the auditory and audiovisual stimuli. Black contours indicate the pixels with significant power or phase coherence increase, identified using a bootstrapping procedure (right-tailed *z* test, $P < 0.05$, Bonferroni corrected).

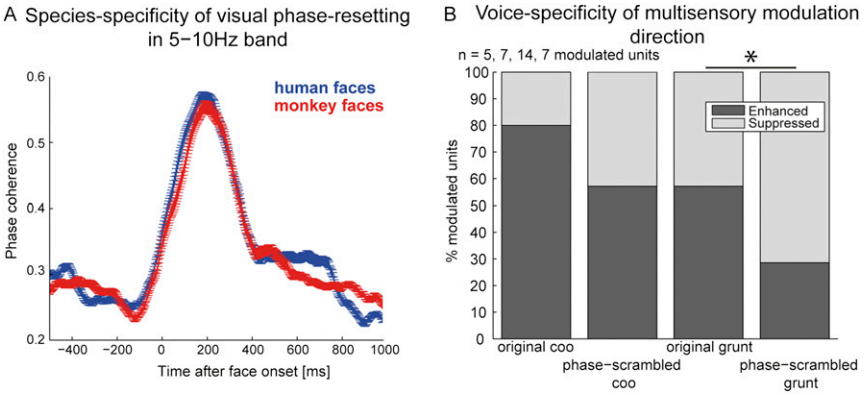

Fig. S8. Voice- and face-specificity of the cross-modal phase-resetting and multisensory effects. (*A*) Time course of the phase-coherence increase in the 5- to 10-Hz frequency band elicited by human and monkey faces ($n = 52$ sites). Shown is mean ± SEM. (*B*) Proportion of enhanced/suppressed units in response to audiovisual pairs combining a face with an intact vocalization (coo and grunt) and to incongruent audiovisual pairs combining the face with a phase-scrambled version of the original vocalizations. $*P < 0.01$, χ^2 test.

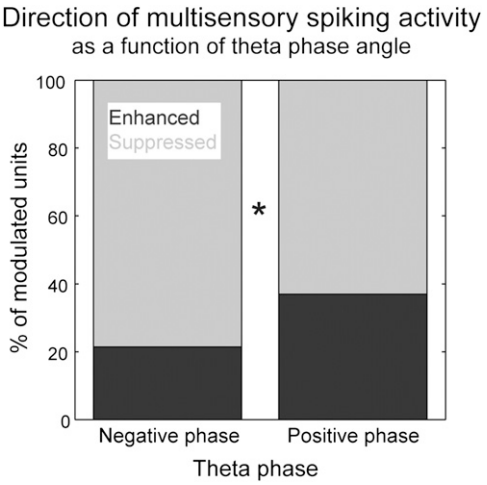

Fig. S9. Positive values of cross-modal theta phase are linked to larger proportions of enhanced multisensory spiking responses. Proportion of enhanced and suppressed multisensory units, ordered according to the value of the visually evoked theta (5–10 Hz) phase angle immediately before the sound onset ($n = 14$ and $n = 27$ multisensory units for negative and positive phase angles, respectively). $*P < 0.05$, χ^2 test.