Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex

Catherine Perrodin[a], Christoph Kayser[b], Nikos K. Logothetis[a,c], and Christopher I. Petkov[d,1]

[a]Department of Physiology of Cognitive Processes, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany; [b]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, United Kingdom; [c]Division of Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, United Kingdom; and [d]Institute of Neuroscience, Newcastle University Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom

Edited by Jon H. Kaas, Vanderbilt University, Nashville, TN, and approved December 2, 2014 (received for review July 7, 2014)

When social animals communicate, the onset of informative content in one modality varies considerably relative to the other, such as when visual orofacial movements precede a vocalization. These naturally occurring asynchronies do not disrupt intelligibility or perceptual coherence. However, they occur on time scales where they likely affect integrative neuronal activity in ways that have remained unclear, especially for hierarchically downstream regions in which neurons exhibit temporally imprecise but highly selective responses to communication signals. To address this, we exploited naturally occurring face- and voice-onset asynchronies in primate vocalizations. Using these as stimuli we recorded cortical oscillations and neuronal spiking responses from functional MRI (fMRI)-localized voice-sensitive cortex in the anterior temporal lobe of macaques. We show that the onset of the visual face stimulus resets the phase of low-frequency oscillations, and that the face–voice asynchrony affects the prominence of two key types of neuronal multisensory responses: enhancement or suppression. Our findings show a three-way association between temporal delays in audiovisual communication signals, phase-resetting of ongoing oscillations, and the sign of multisensory responses. The results reveal how natural onset asynchronies in cross-sensory inputs regulate network oscillations and neuronal excitability in the voice-sensitive cortex of macaques, a suggested animal model for human voice areas. These findings also advance predictions on the impact of multisensory input on neuronal processes in face areas and other brain regions.

oscillations | neurons | communication | voice | multisensory

How the brain parses multisensory input despite the variable and often large differences in the onset of sensory signals across different modalities remains unclear. We can maintain a coherent multisensory percept across a considerable range of spatial and temporal discrepancies (1–4): For example, auditory and visual speech signals can be perceived as belonging to the same multisensory "object" over temporal windows of hundreds of milliseconds (5–7). However, such misalignment can drastically affect neuronal responses in ways that may also differ between brain regions (8–10). We asked how natural asynchronies in the onset of face/voice content in communication signals would affect voice-sensitive cortex, a region in the ventral "object" pathway (11) where neurons (*i*) are selective for auditory features in communication sounds (12–14), (*ii*) are influenced by visual "face" content (12), and (*iii*) display relatively slow and temporally variable responses in comparison with neurons in primary auditory cortical or subcortical structures (14–16).

Neurophysiological studies in human and nonhuman animals have provided considerable insights into the role of cortical oscillations during multisensory conditions and for parsing speech. Cortical oscillations entrain to the slow temporal dynamics of natural sounds (17–20) and are thought to reflect the excitability of local networks to sensory inputs (21–24). Moreover, at least in auditory cortex, the onset of sensory input from the nondominant modality can reset the phase of ongoing auditory cortical oscillations (8, 25, 26), modulating the processing of subsequent acoustic input (8, 18, 22, 26–28). Thus, the question arises as to whether and how the phase of cortical oscillations in voice-sensitive cortex is affected by visual input.

There is limited evidence on how asynchronies in multisensory stimuli affect cortical oscillations or neuronal multisensory interactions. Moreover, as we consider in the following, there are some discrepancies in findings between studies, leaving unclear what predictions can be made for regions beyond the first few stages of auditory cortical processing. In general there are two types of multisensory response modulations: Neuronal firing rates can be either suppressed or enhanced in multisensory compared with unisensory conditions (9, 12, 25, 29, 30). In the context of audiovisual communication Ghazanfar et al. (9) showed that these two types of multisensory influences are not fixed. Rather, they reported that the proportion of suppressed and enhanced multisensory responses in auditory cortical local-field potentials varies depending on the natural temporal asynchrony between the onset of visual (face) and auditory (voice) information. They interpret their results as an approximately linear change from enhanced to suppressed responses with increasing asynchrony between face movements and vocalization onset. In contrast, Lakatos et al. (8) found a cyclic, rather than

Significance

Social animals often combine vocal and facial signals into a coherent percept, despite variable misalignment in the onset of informative audiovisual content. However, whether and how natural misalignments in communication signals affect integrative neuronal responses is unclear, especially for neurons in recently identified temporal voice-sensitive cortex in nonhuman primates, which has been suggested as an animal model for human voice areas. We show striking effects on the excitability of voice-sensitive neurons by the variable misalignment in the onset of audiovisual communication signals. Our results allow us to predict the state of neuronal excitability from the cross-sensory asynchrony in natural communication signals and suggest that the general pattern that we observed would generalize to face-sensitive cortex and certain other brain areas.

Author contributions: C.P. and C.I.P. designed research; C.P. performed research; C.K. and N.K.L. contributed new reagents/analytic tools; C.P. analyzed data; and C.P. and C.I.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: chris.petkov@ncl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412817112/-/DCSupplemental.

NEUROSCIENCE

linear, pattern of multisensory enhancement and suppression in auditory cortical neuronal responses as a function of increasing auditory–somatosensory stimulus onset asynchrony. This latter result suggests that the proportion of suppressed/enhanced multisensory responses varies nonlinearly (i.e., cyclically) with the relative onset timing of cross-modal stimuli. Although such results highlight the importance of multisensory asynchronies in regulating neural excitability, the differences between the studies prohibit generalizing predictions to other brain areas and thus leave the general principles unclear.

In this study we aimed to address how naturally occurring temporal asynchronies in primate audiovisual communication signals affect both cortical oscillations and neuronal spiking activity in a voice-sensitive region. Using a set of human and monkey dynamic faces and vocalizations exhibiting a broad range of audiovisual onset asynchronies (Fig. 1), we demonstrate a three-way association between face–voice onset asynchrony, cross-modal phase resetting of cortical oscillations, and a cyclic pattern of dynamically changing proportions of suppressed and enhanced neuronal multisensory responses.

Results

We targeted neurons for extracellular recordings in a right-hemisphere voice-sensitive area on the anterior supratemporal plane of the rhesus macaque auditory cortex. This area resides anterior to tonotopically organized auditory fields (13). Recent work has characterized the prominence and specificity of multisensory influences on neurons in this area: Visual influences on these neurons are typically characterized by nonlinear multisensory interactions (12), with audiovisual responses being either superadditive or subadditive in relation to the sum of the responses to the unimodal stimuli (Fig. 2*A*). The terms "enhanced" and "suppressed" are often used to refer to a multisensory difference relative to the maximally responding unisensory condition. These types of multisensory responses can be comparable to superadditive/subadditive effects (i.e., relative to the summed unimodal responses) if there is a weak or no response to one of the stimulus conditions (e.g., visual stimuli in auditory cortex). In our study the effects are always measured in relation to the summed unimodal response, yet we use the terms enhanced/suppressed simply for readability throughout.

We investigated whether the proportions of suppressed/enhanced neuronal spiking responses covary with the asynchrony between the visual and auditory stimulus onset, or other sound features. The visual–auditory delays (VA delays) ranged from 77 to 219 ms (time between the onset of the dynamic face video and the onset of the vocalization; Fig. 1). When the vocalization stimuli were sorted according to their VA delays, we found that the relative proportion of units showing either multisensory enhancement or suppression of their firing rates strongly depended on VA delay. Multisensory enhancement was most likely for midrange VA delays between 109 and 177 ms, whereas suppression was more likely for very short VA delays (77–89 ms) and long VA delays (183–219 ms; Fig. 2*B*).

We first ruled out trivial explanations for the association between VA delays and the proportions of the two forms of multisensory influences. The magnitude of the unisensory responses and the prominence of visual modulation were found to be comparable for both midrange and long VA delays (Fig. S1). Moreover, we found little support for any topographic differences in the multisensory response types (Fig. S2), and no other feature of the audiovisual communication signals, such as call type, caller body size, or caller identity, was as consistently associated with the direction of visual influences (12). Together, these observations underscore the association between VA delays and the form of multisensory influences. Interestingly, the relationship between the type of multisensory interaction and the VA delay seemed to follow a cyclic pattern and was well fit by an 8.4-Hz sinusoidal function (red curve in Fig. 2*B*; adjusted $R^2 = 0.583$). Fitting cyclic functions with other time scales explained much smaller amounts of the data variance (e.g., 4 Hz: adjusted $R^2 = 0.151$; 12 Hz: adjusted $R^2 = -0.083$), suggesting that a specific time scale around 8 Hz underlies the multisensory modulation pattern.

We confirmed that this result was robust to the specific way of quantification. First, the direction of multisensory influences was stable throughout the 400-ms response window used for analysis (Fig. S3), and the cyclic modulation pattern was evident when using a shorter response window (Fig. S4). Second, this cyclic pattern was also evident in well-isolated single units from the dataset (Fig. S5), and a similar pattern was observed using a nonbinary metric of multisensory modulation (additivity index; Fig. S6).

Given the cyclic nature of the multisensory interaction and VA delay association, we next asked whether and how this relates to cortical oscillations in the local-field potential (LFP). To assess the oscillatory context of the spiking activity for midrange vs. long VA delays, we computed the stimulus-evoked broadband LFP response to the auditory, visual, and audiovisual stimulation. The grand average evoked potential across sites and stimuli revealed strong auditory- and audiovisually evoked LFPs, including a visually-evoked LFP (Fig. S7*A*). We observed that purely visual stimulation elicited a significant power increase in the low frequencies (5–20 Hz; bootstrapped significance test, $P < 0.05$, Bonferroni corrected; Fig. 3*A*). This visual response was

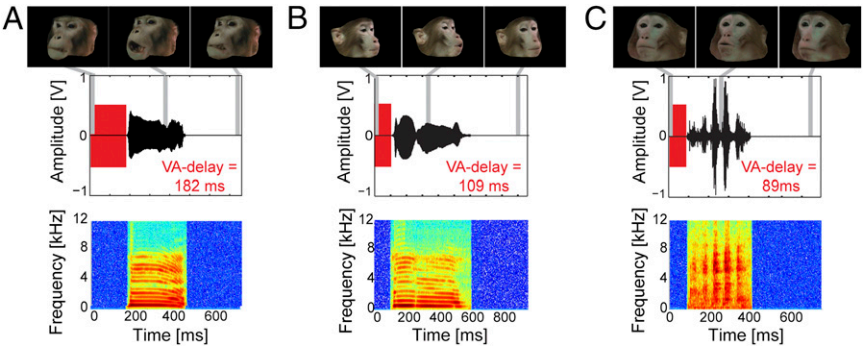

Fig. 1. Audiovisual primate vocalizations and visual–auditory delays. (*A*–*C*) Examples of audiovisual rhesus macaque coo (*A* and *B*) and grunt (*C*) vocalizations used for stimulation and their respective VA delays (time interval between the onset of mouth movement and the onset of the vocalization; red bars). The video starts at the onset of mouth movement, with the first frame showing a neutral facial expression, followed by mouth movements associated with the vocalization. Gray lines indicate the temporal position of the representative video frames (top row). The amplitude waveforms (middle row) and the spectrograms (bottom row) of the corresponding auditory component of the vocalization are displayed below.

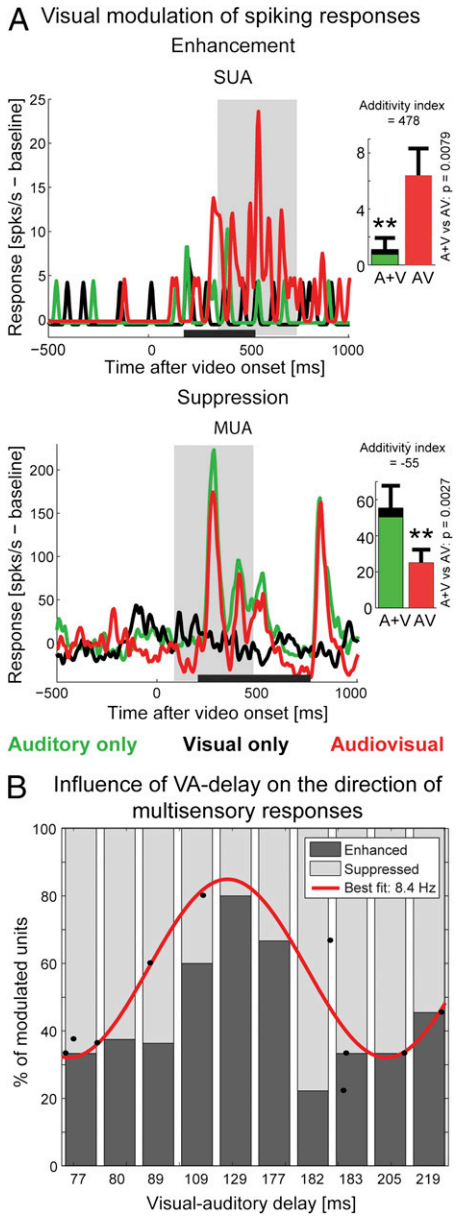

Fig. 2. VA delay and the direction (sign) of multisensory interactions. (*A*) Example spiking responses with nonlinear visual modulation of auditory activity: enhancement (superadditive multisensory effect, *Top*) and suppression (subadditive multisensory, *Bottom*). The horizontal gray line indicates the duration of the auditory stimulus, and the light gray box represents the 400-ms peak-centered response window. Bar plots indicate the response amplitudes in the 400-ms response window (shown is mean ± SEM). The *P* values refer to significantly nonlinear audiovisual interactions, defined by comparing the audiovisual response with all possible summations of the unimodal responses [A vs. (A + V), *z* test, $*P < 0.01$]. (*B*) Proportions of enhanced and suppressed multisensory units by stimulus, arranged as a function of increasing VA delays ($n = 81$ units). Note that the bars are spaced at equidistant intervals for display purposes. Black dots indicate the proportion of enhanced units 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 (8.4 Hz, adjusted $R^2 = 0.58$).

accompanied by a significant increase in intertrial phase coherence, restricted to the 5- to 10-Hz frequency band, between 130 and 350 ms after video onset (phase coherence values significantly larger than a bootstrapped null distribution of time-frequency values; $P < 0.05$, Bonferroni corrected; Fig. 3*B*). In contrast, auditory and audiovisual stimulation yielded broadband increases in LFP power (Fig. S7*B*) and an increased phase coherence spanning a wider band (5–45 Hz; Fig. S7*C*). Thereby, the response induced by the purely visual stimuli suggests that dynamic faces may influence the state of slow rhythmic activity in this temporal voice area via phase resetting of ongoing low-frequency oscillations. Noteworthy, the time scale of the relevant brain oscillations (5–10 Hz) and the time at which the phase coherence increases (~100–350 ms; Fig. 3*B*) match the time scale (8 Hz) and range of VA delays at which the cyclic pattern of multisensory influences on the spiking activity emerged (Fig. 2*B*).

We found little evidence that the phase resetting was species-specific, because both human and monkey dynamic face stimuli elicited a comparable increase in intertrial phase coherence (Fig. S8*A*). Similarly, the relative proportion of enhanced/suppressed units was not much affected when a coo or grunt call was replaced with a phase-scrambled version (that preserves the overall frequency content but removes the temporal envelope; Fig. S8*B* and ref. 12). Both observations suggest that the underlying

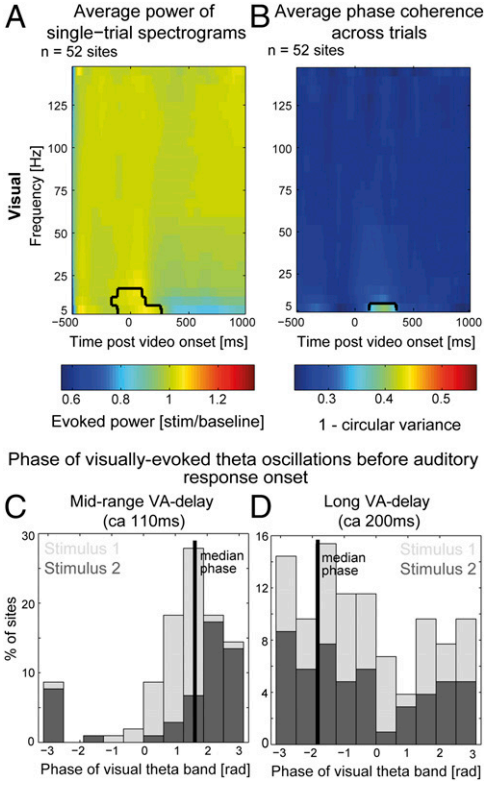

Fig. 3. Visually evoked oscillatory context surrounding the spiking activity at different audiovisual asynchronies in voice-sensitive cortex. (*A*) Time–frequency plot of averaged single-trial spectrograms in response to visual face stimulation. The population-averaged spectrogram has been baseline-normalized for display purposes. (*B*) Time–frequency plot of average phase coherence values across trials. The color code reflects the strength of phase alignment evoked by the visual stimuli. The range of values in *A* and *B* are the same as in Fig. S7, to allow for closer comparisons. 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). (*C*) Distribution of the theta/low-alpha (5- to 10-Hz band) phase values at the time of auditory response, for vocalizations with midrange VA delays ($n = 52$ sites). (*D*) Distribution of theta/low-alpha band phase values at sound arrival, for the stimuli with long VA delays. The vertical black bar indicates the value of the circular median phase angle.

processes are not stimulus-specific but reflect more generic visual modulation of voice area excitability.

The frequency band exhibiting phase reset included the ∼8-Hz frequency at which we found the cyclic variation of multisensory enhancement and suppression in the firing rates (Fig. 2*B*). Thus, we next asked whether the oscillatory context in this band could predict the differential multisensory influence on neuronal firing responses at different VA delays. We computed the value of the visually evoked phase in the relevant 5- to 10-Hz theta band for each recording site at the time at which the vocalization sound first affects these auditory neurons' responses. This time point was computed as the sum of the VA delay for each vocalization stimulus and the sensory processing time, which we estimated using the mean auditory latency of neurons in the voice area (110 ms; see ref. 12). The phase distributions of the theta band oscillations at midrange and long VA delays is shown in Fig. 3 *C* and *D*. Both distributions deviated significantly from uniformity (Rayleigh test, $P = 1.2 \times 10^{-25}$, Z = 41.1 for midrange VA delays; $P = 0.035$, Z = 3.4 for long VA delays). For midrange VA delays the phase distributions were centered around a median phase angle of 1.6 rad (92°), and for long VA delays at −1.9 rad (−109°). The phase distributions significantly differed across midrange and long VA delays (Kuiper two-sample test: $P = 0.001$, K = 5,928). These results show that the auditory stream of the multisensory signal reaches voice-sensitive neurons in a different oscillatory context for the two VA-delay durations. In particular, at midrange VA delays the preferred phase angle (*ca.* π/2) corresponds to the descending slope of ongoing oscillations and is typically considered the "ideal" excitatory phase: The spiking response to a stimulus arriving at that phase is enhanced (8, 25, 28). In contrast, at long VA delays the preferred phase value corresponds to a phase of less optimal neuronal excitability.

Finally, we directly probed the association between the cross-modal phase at the time of auditory response onset and the direction of subsequent multisensory spiking responses. We labeled each unit that displayed significant audiovisual interactions with the corresponding visually evoked theta phase angle immediately before the onset of the vocalization response. This revealed that the proportions of enhanced and suppressed spiking responses significantly differed between negative and positive phase values [χ^2 test, $P = 0.0081$, $\chi^2(1, n = 41) = 7.02$]. Multisensory enhancement was more frequently associated with positive phase angles (10/27 = 37% of units) compared with negative phase angles (3/14 = 21% of units; Fig. S9). In summary, the findings show three-way relationships between the visual–auditory delays in communication signals, the phase of theta oscillations, and a cyclically varying proportion of suppressed vs. enhanced multisensory neuronal responses in voice-sensitive cortex.

Discussion

This study probed neurons in a primate voice-sensitive region, which forms a part of the ventral object processing stream (11) and has links to human functional MRI (fMRI)-identified temporal voice areas (31, 32). The results show considerable impact on the state of global and local neuronal excitability in this region by naturally occurring audiovisual asynchronies in the onset of informative content. The main findings show a three-way association between (*i*) temporal asynchronies in the onset of visual dynamic face content and the onset of vocalizations, (*ii*) the phase of low-frequency neuronal oscillations, and (*iii*) cyclically varying proportions of enhanced vs. suppressed multisensory neuronal responses.

Prior studies do not provide a consistent picture of how cross-sensory stimulus asynchronies affect neuronal multisensory responses. One study evaluated the impact on audiovisual modulations in LFPs around core and belt auditory cortex using natural face–voice asynchronies in videos of vocalizing monkeys (9). The authors reported a gradual increase in the prominence of multisensory suppression with increasing visual–auditory onset delays. However, another study recording spiking activity from auditory cortex found that shifting somatosensory nerve stimulation relative to sound stimulation with tones resulted in a cyclic, rather than linear, pattern of alternating proportions of enhanced and suppressed spiking responses (8). A third study mapped the neural window of multisensory interaction in A1 using transient audiovisual stimuli with a range of onsets (25), identifying a fixed time window (20–80 ms) in which sounds interact in a mostly suppressive manner. Finally, a fourth study recorded LFPs in the superior-temporal sulcus (STS) and found that different frequency bands process audiovisual input streams differently (10). The study also showed enhancement for short visual–auditory asynchronies in the alpha band and weak to no dependancy on visual–auditory asynchrony in the other LFP frequency bands, including theta (10). Given the variability in results, the most parsimonious interpretation was that multisensory asynchrony effects on neuronal excitability are stimulus-, neuronal response measure-, and/or brain area-specific.

Comparing our findings to these previous results suggests that the multisensory effects are not necessarily stimulus-specific and the differences across brain areas might be more quantitative than qualitative. Specifically, our data from voice-sensitive cortex show that the direction of audiovisual influences on spiking activity varies cyclically as a function of VA delay. This finding is most consistent with the data from auditory cortical neurons showing cyclic patterns of suppressed/enhanced responses to somatosensory–auditory stimulus asynchronies (8). Together these results suggest a comparable impact on cortical oscillations and neuronal multisensory modulations by asynchronies in different types of multisensory stimuli (8, 25). Interestingly, when looked at in detail some of the other noted studies (9, 25) show at least partial evidence for a cyclic pattern of multisensory interactions.

Some level of regional specificity is expected, given that, for example, relatively simple sounds are not a very effective drive for neurons in voice-sensitive cortex (13, 14). However, we did not find strong evidence for any visual or auditory stimulus specificity in the degree of phase resetting or the proportions of multisensory responses. Hence, it may well be that some oscillatory processes underlying multisensory interactions reflect general mechanisms of cross-modal visual influences, which are shared between voice-sensitive and earlier auditory cortex. It is an open question whether regions further downstream, such as the frontal cortex or STS (33, 34), might integrate cross-sensory input differently. In any case, our results emphasize the rhythmicity underlying multisensory interactions and hence generate specific predictions for other sensory areas such as face-sensitive cortex (35).

The present results predict qualitatively similar effects for face-sensitive areas in the ventral temporal lobe, with some key quantitative differences in the timing of neuronal interactions, as follows. The dominant input from the visual modality into face-sensitive neurons would drive face-selective spiking responses with a latency of ∼100 ms after the onset of mouth movement (35–37). Nearly coincident cross-modal input into face-sensitive areas from the nondominant auditory modality would affect the phase of the ongoing low-frequency oscillations and likely affect face areas at about the same time as the face-selective spiking responses (38) or later for any VA delay. Based on our results, we predict a comparable cyclic pattern of auditory modulation of the visual spiking activity, as a function of VA delay. However, because in this case the dominant modality for face-area neurons is visual, and in natural conditions visual onset often precedes vocalization onset, the pattern of excitability is predicted to be somewhat phase-shifted in relation to those from the voice area. For example, shortly after the onset of face-selective neuronal

responses, perfectly synchronous stimulation (0 ms), or those with relatively short VA delays (~75 ms), would be associated predominantly with multisensory suppression. Interestingly, some evidence for this prediction can already be seen in the STS results of a previous study (39) using synchronous face–voice stimulation.

The general mechanism of cross-modal phase resetting of cortical oscillations and its impact on neuronal response modulation has been described in the primary auditory cortex of nonhuman primates (8, 25) and in auditory and visual cortices in humans (27, 40). Prior work has also highlighted low-frequency (e.g., theta) oscillations and has hypothesized that one key role of phase-resetting mechanisms is to align cortical excitability to important events in the stimulus stream (8, 22, 24, 26). Our study extends these observations to voice-sensitive cortex: We observed that visual stimulation resets the phase of theta/low-alpha oscillations and that the resulting multisensory modulation of firing rates depends on the audiovisual onset asynchrony. We also show how cross-sensory asynchronies in communication signals affect the phase of low-frequency cortical oscillations and regulate periods of neuronal excitability.

Cross-modal perception can accommodate considerable temporal asynchronies between the individual modalities before the coherence of the multimodal percept breaks down (5–7), in contrast to the high perceptual sensitivity to unisensory input alignment (41). For example, observers easily accommodate the asynchrony between the onset of mouth movement and the onset of a vocalization sound, which can be up to 300 ms in monkey vocalizations (9) or human speech (42). More generally, a large body of behavioral literature shows that multisensory perceptual fusion can be robust over extended periods of cross-sensory asynchrony, without any apparent evidence for "cyclic" fluctuations in the coherence of the multisensory percept (5–7). Given the variety of multisensory response types elicited during periods in which stable perceptual fusion should occur, our results underscore the functional role of both enhanced and suppressed spiking responses (43, 44). However, this perceptual robustness is in apparent contrast to the observed rhythmicity of neuronal integrative processes (8, 9, 25, 30).

It could be that audiovisual asynchronies and their cyclic effects on neuronal excitability are associated with subtle fluctuations in perceptual sensitivity that are not detected with suprathreshold stimuli. Evidence supporting this possibility in human studies shows that the phase of entrained or cross-modally reset cortical oscillations can have subtle effects on auditory perception (24, 45, 46), behavioral response times (26), and visual detection thresholds (23, 40, 47, 48). Previous work has also shown both that the degree of multisensory perceptual binding is modulated by stimulus type (5), task (49), and prior experience (50), and that oscillatory entrainment adjusts as a function of selective attention to visual or auditory stimuli (51, 52). Given this it seemed important to first investigate multisensory interactions in voice-sensitive cortex during a visual fixation task irrelevant to the specific face/voice stimuli, so as to minimize task-dependent influences. This task-neutral approach is also relevant given that the contribution of individual cortical areas to multisensory voice perception remains unclear. Future work needs to compare the patterns of multisensory interactions across temporal lobe regions and to identify their specific impact on perception.

By design, the start of the videos in our experiment is indicative of the onset of a number of different sources of visually informative content. Although articulatory mouth movements seem to dominantly attract the gaze of primates (53, 54), a continuous visual stream might offer a number of time points at which visual input can influence the phase of the ongoing auditory cortical oscillations by capturing the animal's attention and gaze direction (55). Starting from our results, future work can specify whether and how subsequent audiovisual fluctuations in the onset of informative content alter or further affect the described multisensory processes.

In summary, our findings show that temporal asynchronies in audiovisual face/voice communication signals seem to reset the phase of theta-range cortical oscillations and regulate the two key types of multisensory neuronal interactions in primate voice-sensitive cortex. This allows predicting the form of local and global neuronal multisensory responses by calculating the naturally occurring asynchrony in the audiovisual input signal. This study can serve as a link between neuron-level work in nonhuman animal models and work using noninvasive approaches in humans to study the neurobiology of multisensory processes.

Materials and Methods

Full methodological details are provided in *SI Materials and Methods* and are summarized here. Two adult male Rhesus macaques (*Macaca mulatta*) participated in these experiments. All procedures were approved by the local authorities (Regierungspräsidium Tübingen, Germany) and were in full compliance with the guidelines of the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals.

Audiovisual Stimuli. Naturalistic audiovisual stimuli consisted of digital video clips (recorded with a Panasonic NV-GS17 digital camera) of a set of "coo" and "grunt" vocalizations by rhesus monkeys and recordings of humans imitating monkey coo vocalizations. The stimulus set included $n = 10$ vocalizations. For details see ref. 12 and *SI Materials and Methods*.

Electrophysiological Recordings. Electrophysiological recordings were obtained while the animals performed a visual fixation task. Only data from successfully completed trials were analyzed further (*SI Materials and Methods*). The two macaques had previously participated in fMRI experiments to localize their voice-preferring regions, including the anterior voice-identity sensitive clusters (see refs. 13 and 31). A custom-made multielectrode system was used to independently advance up to five epoxy-coated tungsten microelectrodes (0.8–2 MOhm impedance; FHC Inc.). The electrodes were advanced to the MRI-calculated depth of the anterior auditory cortex on the supratemporal plane (STP) through an angled grid placed on the recording chamber. Electrophysiological signals were amplified using an amplifier system (Alpha Omega GmbH), filtered between 4 Hz and 10 kHz (four-point Butterworth filter) and digitized at a 20.83-kHz sampling rate. For further details see *SI Materials and Methods* and ref. 13.

The data were analyzed in MATLAB (MathWorks). The spiking activity was obtained by first high-pass filtering the recorded broadband signal at 500 Hz (third-order Butterworth filter) then extracted offline using commercial spike-sorting software (Plexon Offline Sorter; Plexon Inc.). Spike times were saved at a resolution of 1 ms. Peristimulus time histograms were obtained using 5-ms bins and 10-ms Gaussian smoothing (FWHM). LFPs were obtained by low-pass filtering the recorded broadband signal at 150 Hz (third-order Butterworth filter). The broadband evoked potentials were full-wave rectified. For time-frequency analysis, trial-based activity between 5 and 150 Hz was filtered into 5-Hz-wide bands using a fourth-order Butterworth filter. Instantaneous power and phase were extracted using the Hilbert transform on each frequency band.

Data Analysis. A unit was considered auditory-responsive if its average response amplitude exceeded 2 SD units from its baseline activity during a continuous period of at least 50 ms, for any of the experimental sounds in the set of auditory or audiovisual stimuli. A recording site was included in the LFP analysis if at least one unit recorded at this site showed a significant auditory response. For each unit and each stimulus, the mean of the baseline response was subtracted to compensate for fluctuations in spontaneous activity. Response amplitudes were defined as the average response in a 400-ms window centered on the peak of the trial-averaged stimulus response. The same window was used to compute the auditory, visual, and audiovisual response amplitudes for each stimulus.

Multisensory interactions were assessed individually for each unit with a significant response to sensory stimulation (A, V, or AV). A sensory-responsive unit was termed "nonlinear multisensory" if its response to the audiovisual stimulus was significantly different from a linear (additive) sum of the two unimodal responses [$AV \sim (A + V)$]. This was computed for each unit and for each stimulus that elicited a significant sensory response, by implementing a randomization procedure (25, 56) described in more details in *SI Materials and Methods*.

The parameters and goodness of fit of a sinusoid of the form $F(x) = a_0 + a_1*\cos(\omega x) + b_1*\sin(\omega x)$ were estimated using the MATLAB curve-fitting toolbox. To compare the differential impact of midrange and long VA delays on neuronal activity, the analysis focused on vocalizations representative of midrange (109 and 129 ms) and long (205 and 219 ms) VA delays. The significance of stimulus-evoked increase in phase coherence was assessed using a randomization procedure. For each frequency band, a bootstrapped distribution of mean phase coherence was created by randomly sampling $n = 1{,}000$ phase coherence values across time bins. Time-frequency bins were deemed significant if their phase coherence value was sufficiently larger than the bootstrapped distribution (right-tailed z test, $P < 0.05$, Bonferroni corrected). Statistical testing of single-trial phase data was performed using the CircStat MATLAB toolbox (57).

ACKNOWLEDGMENTS. We thank J. Obleser and A. Ghazanfar for comments on previous versions of the manuscript. This work was supported by the Max-Planck Society (C.P., C.K., and N.K.L.), Swiss National Science Foundation Grant PBSKP3_140120 (to C.P.), Wellcome Trust Grants WT092606/Z/10/Z and WT102961MA (to C.I.P.), and Biotechnology and Biological Sciences Research Council Grant BB/L027534/1 (to C.K.).

1. Shams L, Kamitani Y, Shimojo S (2000) Illusions. What you see is what you hear. *Nature* 408(6814):788.
2. Howard IP, Templeton WB (1966) *Human Spatial Orientation* (Wiley, London), p 533.
3. Slutsky DA, Recanzone GH (2001) Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12(1):7–10.
4. McGrath M, Summerfield Q (1985) Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J Acoust Soc Am* 77(2):678–685.
5. Stevenson RA, Wallace MT (2013) Multisensory temporal integration: Task and stimulus dependencies. *Exp Brain Res* 227(2):249–261.
6. van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45(3):598–607.
7. Vatakis A, Spence C (2006) Audiovisual synchrony perception for music, speech, and object actions. *Brain Res* 1111(1):134–142.
8. Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53(2):279–292.
9. Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J Neurosci* 25(20): 5004–5012.
10. Chandrasekaran C, Ghazanfar AA (2009) Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *J Neurophysiol* 101(2): 773–788.
11. Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat Neurosci* 12(6):718–724.
12. 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.
13. Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. *Curr Biol* 21(16):1408–1415.
14. Kikuchi Y, Horwitz B, Mishkin M (2010) Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J Neurosci* 30(39):13021–13030.
15. Bendor D, Wang X (2007) Differential neural coding of acoustic flutter within primate auditory cortex. *Nat Neurosci* 10(6):763–771.
16. Creutzfeldt O, Hellweg FC, Schreiner C (1980) Thalamocortical transformation of responses to complex auditory stimuli. *Exp Brain Res* 39(1):87–104.
17. Ghitza O (2011) Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol* 2:130.
18. Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat Neurosci* 15(4):511–517.
19. Ding N, Chatterjee M, Simon JZ (2013) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88C:41–46.
20. Ng BS, Logothetis NK, Kayser C (2013) EEG phase patterns reflect the selectivity of neural firing. *Cereb Cortex* 23(2):389–398.
21. Engel AK, Senkowski D, Schneider TR (2012) Multisensory integration through neural coherence. *The Neural Bases of Multisensory Processes*, Frontiers in Neuroscience, eds Murray MM, Wallace MT (CRC, Boca Raton, FL).
22. Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* 12(3):106–113.
23. Thut G, Miniussi C, Gross J (2012) The functional importance of rhythmic activity in the brain. *Curr Biol* 22(16):R658–R663.
24. Ng BS, Schroeder T, Kayser C (2012) A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *J Neurosci* 32(35):12268–12276.
25. Kayser C, Petkov CI, Logothetis NK (2008) Visual modulation of neurons in auditory cortex. *Cereb Cortex* 18(7):1560–1574.
26. Thorne JD, De Vos M, Viola FC, Debener S (2011) Cross-modal phase reset predicts auditory task performance in humans. *J Neurosci* 31(10):3853–3861.
27. van Atteveldt N, Murray MM, Thut G, Schroeder CE (2014) Multisensory integration: Flexible use of general operations. *Neuron* 81(6):1240–1253.
28. Lakatos P, et al. (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94(3):1904–1911.
29. Sugihara T, Diltz MD, Averbeck BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *J Neurosci* 26(43):11138–11147.
30. Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cereb Cortex* 17(9): 2172–2189.
31. Petkov CI, et al. (2008) A voice region in the monkey brain. *Nat Neurosci* 11(3): 367–374.
32. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403(6767):309–312.
33. Werner S, Noppeney U (2010) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J Neurosci* 30(7): 2662–2675.
34. Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? *Trends Cogn Sci* 10(6):278–285.
35. Tsao DY, Freiwald WA, Tootell RB, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. *Science* 311(5761):670–674.
36. Leopold DA, Bondar IV, Giese MA (2006) Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442(7102):572–575.
37. Perrett DI, Rolls ET, Caan W (1982) Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res* 47(3):329–342.
38. Schall S, Kiebel SJ, Maess B, von Kriegstein K (2013) Early auditory sensory processing of voices is facilitated by visual mechanisms. *Neuroimage* 77:237–245.
39. Dahl CD, Logothetis NK, Kayser C (2010) Modulation of visual responses in the superior temporal sulcus by audio-visual congruency. *Front Integr Neurosci* 4:10.
40. Romei V, Gross J, Thut G (2012) Sounds reset rhythms of visual cortex and corresponding human visual perception. *Curr Biol* 22(9):807–813.
41. Zampini M, Guest S, Shore DI, Spence C (2005) Audio-visual simultaneity judgments. *Percept Psychophys* 67(3):531–544.
42. Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. *PLOS Comput Biol* 5(7):e1000436.
43. Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. *Curr Biol* 20(1):19–24.
44. Ohshiro T, Angelaki DE, DeAngelis GC (2011) A normalization model of multisensory integration. *Nat Neurosci* 14(6):775–782.
45. Henry MJ, Obleser J (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc Natl Acad Sci USA* 109(49): 20095–20100.
46. Henry MJ, Herrmann B, Obleser J (2014) Entrained neural oscillations in multiple frequency bands comodulate behavior. *Proc Natl Acad Sci USA* 111(41):14935–14940.
47. Busch NA, Dubois J, VanRullen R (2009) The phase of ongoing EEG oscillations predicts visual perception. *J Neurosci* 29(24):7869–7876.
48. Fiebelkorn IC, et al. (2011) Ready, set, reset: Stimulus-locked periodicity in behavioral performance demonstrates the consequences of cross-sensory phase reset. *J Neurosci* 31(27):9971–9981.
49. Gleiss S, Kayser C (2013) Eccentricity dependent auditory enhancement of visual stimulus detection but not discrimination. *Front Integr Neurosci* 7:52.
50. Powers AR, 3rd, Hillock AR, Wallace MT (2009) Perceptual training narrows the temporal window of multisensory binding. *J Neurosci* 29(39):12265–12274.
51. Landau AN, Fries P (2012) Attention samples stimuli rhythmically. *Curr Biol* 22(11): 1000–1004.
52. Lakatos P, et al. (2013) The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77(4):750–761.
53. Ghazanfar AA, Nielsen K, Logothetis NK (2006) Eye movements of monkey observers viewing vocalizing conspecifics. *Cognition* 101(3):515–529.
54. Lansing CR, McConkie GW (2003) Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Percept Psychophys* 65(4):536–552.
55. Lakatos P, et al. (2009) The leading sense: Supramodal control of neurophysiological context by attention. *Neuron* 64(3):419–430.
56. Stanford TR, Quessy S, Stein BE (2005) Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J Neurosci* 25(28):6499–6508.
57. Berens P (2009) CircStat: A MATLAB toolbox for circular statistics. *J Stat Softw* 31(10): 1–21.