Opinion

Neuronal oscillations and visual amplification of speech

Charles E Schroeder[1], Peter Lakatos[1], Yoshinao Kajikawa[1], Sarah Partan[2] and Aina Puce[3]

[1] Cognitive Neuroscience and Schizophrenia Program, Nathan Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA
[2] School of Cognitive Science, Hampshire College, Adele Simmons Hall, Amherst, MA 01002-5001, USA
[3] Department of Radiology and Center for Advanced Imaging, West Virginia University School of Medicine, P.O. Box 9100, Morgantown, WV 26506-9100, USA

It is widely recognized that viewing a speaker's face enhances vocal communication, although the neural substrates of this phenomenon remain unknown. We propose that the enhancement effect uses the ongoing oscillatory activity of local neuronal ensembles in the primary auditory cortex. Neuronal oscillations reflect rhythmic shifting of neuronal ensembles between high and low excitability states. Our hypothesis holds that oscillations are 'predictively' modulated by visual input, so that related auditory input arrives during a high excitability phase and is thus amplified. We discuss the anatomical substrates and key timing parameters that enable and constrain this effect. Our hypothesis makes testable predictions for future studies and emphasizes the idea that 'background' oscillatory activity is instrumental to cortical sensory processing.

Seeing voices

Over 50 years ago, Sumby and Pollack [1] noted that viewing the face of a speaker increases the intelligibility of vocal communication. Despite the obvious power and ubiquity of this phenomenon, its specific neural underpinnings remain elusive. We hypothesize that visual amplification of speech is operating as early as the first stage of cortical auditory processing in Area A1 (see Glossary), and that the underlying process entails an elegant and efficient modulation or 'shaping' of ongoing neuronal oscillations. This effect could be a model for cortical modulatory processes in general.

Audiovisual integration in the primary auditory cortex?

In the past decade, functional magnetic resonance imaging (fMRI) studies have consistently pointed to the superior temporal sulcus (STS) region as a crucial part of the brain mechanism for audiovisual integration in speech perception [2]. The idea that the human STS is a key substrate for audiovisual speech processing makes sense, as it seems to correspond, at least in part, to the superior temporal polysensory (STP) cortex in the macaque monkey, a classical site of audiovisual convergence [3]. However, although multisensory integration involves the STS, it does not seem to begin there, as shown by the rapidly mounting evidence implicating areas in and near the primary auditory cortex in audiovisual integration during speech processing [4–6]. Furthermore, although a precise physiological interpretation of these findings is limited by the indirect nature of noninvasive imaging measures, the recent increase in new findings on the multisensory properties of the auditory cortex in nonhuman primates (reviewed in Refs [7,8]) suggests that answers to some of the key mechanistic questions could be forthcoming. Given the parallel between human and monkey communication processing (Box 1), certain findings in monkeys are directly relevant to understanding the brain mechanisms of audiovisual communication in humans.

Hypothesis

The traditional understanding of the neural underpinnings of multisensory enhancement is that because of anatomical convergence at the neuronal level [9], excitatory inputs to neurons are able to sum together in some fashion (reviewed in Ref. [10]). Recent research provides a new perspective on this effect [11]. The basic finding is that non-auditory inputs to the primary auditory cortex can modulate ongoing neuronal activity in a way that amplifies appropriately timed auditory inputs. Because maintained activity in the brain is dominated by rhythmic oscillations [12], our observation suggests that

Glossary

A1: primary auditory cortex.
Neuronal oscillation: the periodic shifting of a neuron or neuronal ensemble between high and low excitability states (phases), at some frequency in cycles per second or Hertz (Hz). Neuronal oscillations are often characterized by the frequency range (band) they occupy in the spectrum (see Box 2 for more details).
Nonspecific thalamic systems: systems with a much less orderly representation of the sensory receptor surface (see Box 3).
Oscillatory coupling: a relationship between two oscillations of different frequencies, which can be of several types.
Phase–amplitude coupling (also known as hierarchical coupling): coupling in which the amplitude of a higher-frequency oscillation is systematically related (coupled) to the phase of a lower frequency oscillation (this is the main type of oscillatory coupling relevant here).
Specific thalamic systems: systems in which thalamic neurons carry a dense and orderly representation of the sensory receptor surface (see Box 3).
STP: superior temporal polysensory division of the STS in macaque monkeys
STS: superior temporal sulcus.
V1: primary visual cortex

Corresponding author: Schroeder, C.E. (schrod@nki.rfmh.org).

1364-6613/$ – see front matter doi:10.1016/j.tics.2008.01.002 Available online xxxxxx

Box 1. Is audiovisual communication in monkeys comparable with that in humans?

Monkeys and humans both rely heavily on audiovisual composite signals for mediating complex social behavior [28]. Human audiovisual communications convey a wide range of content, including emotional state, identity of the speaker and semantic content [37]. The combination of monkey vocalizations with facial expression and body posture makes possible a rich repertoire of meaningful signals, with components that can be redundant or nonredundant in meaning [38]. Monkey faces have nearly as many facial muscles as humans do, enabling a similar variety and subtlety of expression [39], and conspecifics can discriminate among others by face alone [40]. The vocalizations of monkeys can convey emotional state and referential information, such as the type of food an individual has found [41]. Furthermore, monkeys can match faces to voices across auditory and visual modalities [42], as can humans [43], an ability that implies that monkey vocalizations convey information about speaker identity. Basic aspects of prosodic expression in monkeys seem to be similar to those in humans [44].

Despite these parallels, there are limitations on the degree to which monkey vocalizations can serve as a model for human language. The most obvious, widely acknowledged limitation is in the amount of referential and semantic content that can be conveyed in monkey vocalizations. There is also some controversy over the main functions that audiovisual integration serves for the listening monkey. One perspective, for example, is that its main function would be to reinforce detection and spatial localization of the 'speaker,' a function that would be important for monkeys and for humans. Taking the various perspectives into account, it seems safe to regard monkey audiovisual communication as an adequate model for rudimentary aspects of human communication, including speaker identification, speaker localization, limited referential content, emotional content and prosody. The largest limitation is in terms of semantic content.

ongoing neuronal oscillations in the cortex might be a crucial mechanistic component of cortical processing. The idea that oscillations and oscillatory synchrony are crucial to brain operations [13] has been debated extensively over the past decade (see, for example, Ref. [14]); At one extreme, for example, the widely practiced technique of signal averaging treats oscillations and other activity components that are not strictly phase-locked to a stimulus as 'noise.' Recent evidence, however, lends weight to the hypothesis that we advance here: visual cues amplify the cortical processing of accompanying vocalizations by shifting the phase of ongoing neuronal oscillations so that the auditory inputs tend to arrive during a 'high excitability state.' The physiological significance of neuronal oscillations are described in Box 2 and Figure 1.

Four rules about neuronal oscillations

Four rules are key to our hypothesis. First, as illustrated in Figures 1 and 2a, neuronal oscillations reflect synchronized fluctuation of a local neuronal ensemble between high and low excitability states [15–17]. Thus, each oscillation has 'ideal' and 'worst' phases for stimulus processing [16]. This cycling of excitability enables the oscillation to have a role in processing. Inputs that arrive during the ideal phase are 'amplified' (i.e. they generate relatively large responses), whereas inputs that arrive during the worst phase are 'suppressed' (i.e. they generate relatively small responses). In the absence of phase control, the impact of ambient oscillatory activity on sensory inputs to cortex seems almost random [11,16,18].

Box 2. Physiological significance of neuronal oscillations

Over 75 years ago, Bishop [15] proposed that the neuroelectric oscillations comprising the electroencephalogram represent rhythmic shifting of neuronal ensembles between high and low excitability states. Although neuronal oscillations are usually grouped into bands: slow oscillations (below 1 Hz), delta-band (1–4 Hz), theta-band (5–7 Hz), alpha-band (8–12 Hz), beta-band (13–25 Hz) and gamma-band (over 25 Hz) (see, for example, Ref. [12]), this fundamental relationship between oscillatory phase and excitability is independent of the frequency of oscillation [16]. Figure 1, which is based on extensive sampling (25 experiments) from the primary auditory cortex in four macaque monkeys, illustrates this relationship [16]. Neuronal oscillations are often indexed using macroscopic field potentials recorded in the extracellular medium within the brain or even by their volume-conducted electrical signature in the scalp EEG. Field potentials arise mainly from transmembrane currents [45], as do the metabolic demands that drive fMRI signals [46]. They thus signal net local neuronal activation or deactivation (or rhythmic cycling between the two in the case of neuronal oscillations), whether or not the activity leads to obvious phasic changes in action potentials. Although the sensitivity of field potentials is a strength, it also triggers the widely held reservation that in lieu of obvious action potential concomitants, field potential oscillations could be 'epiphenomena,' whose effects would not be transmitted to other brain regions, or even to nearby neurons. Figure 1 addresses this concern, showing that even under baseline conditions when neurons fire spontaneously, the level of firing in a local ensemble is systematically related to the phase of ongoing transmembrane current flow oscillations in the ensemble. This shows that when appropriately sensitive measures are applied, changes in firing generally do indeed attend neuronal oscillations. Also, it underscores the idea that oscillations index change in excitability and, thus, the oscillatory phase influences the probability that action potentials will be generated when appropriate inputs are applied. The fact that significant variations in neuronal excitability are tied to the phase of neuronal oscillations means that oscillations themselves are likely to have a strong impact on the operations that neuronal ensembles perform. There are then two generic possibilities: either oscillations are used as instruments of brain operations, or they constitute a large 'noise' source that degrades the brain's functioning. Most of the data are consistent with the former (instrumental) position.

However, the second rule (Figure 2b) is that the oscillatory phase can be reset by stimulus inputs [11,16]. Phase resetting is necessary for the phenomenon of EEG entrainment to a rhythmic sensory stimulus train [16,19]. More importantly in the present context, phase reset allows the ideal phase of an oscillation to be aligned with an input pattern and thus is key to the use of an oscillation as an input amplifier. If two stimuli occur with a reasonably predictable lag, the first stimulus can 'predictively' reset an oscillation to its ideal phase and thus enhance the response to the second stimulus. As explained below, a predictable visual–auditory lag is integral to cross modal enhancement of vocalization processing.

The third rule is that oscillatory phase modulates subsequent stimulus processing (Figure 2c). The advantage is that after a reset, inputs that arrive within the ideal (high-excitability) phase evoke amplified responses, whereas the responses to inputs that arrive slightly later during the worst phase are suppressed [11,16,20]. The additional benefit stemming from predictive visual reset of auditory cortical oscillations is efficiency, in that it promotes amplification of the auditory input at its onset. It is also likely that the effect of the initial (visual) reset is compounded by the phase-resetting influence of the subsequent auditory

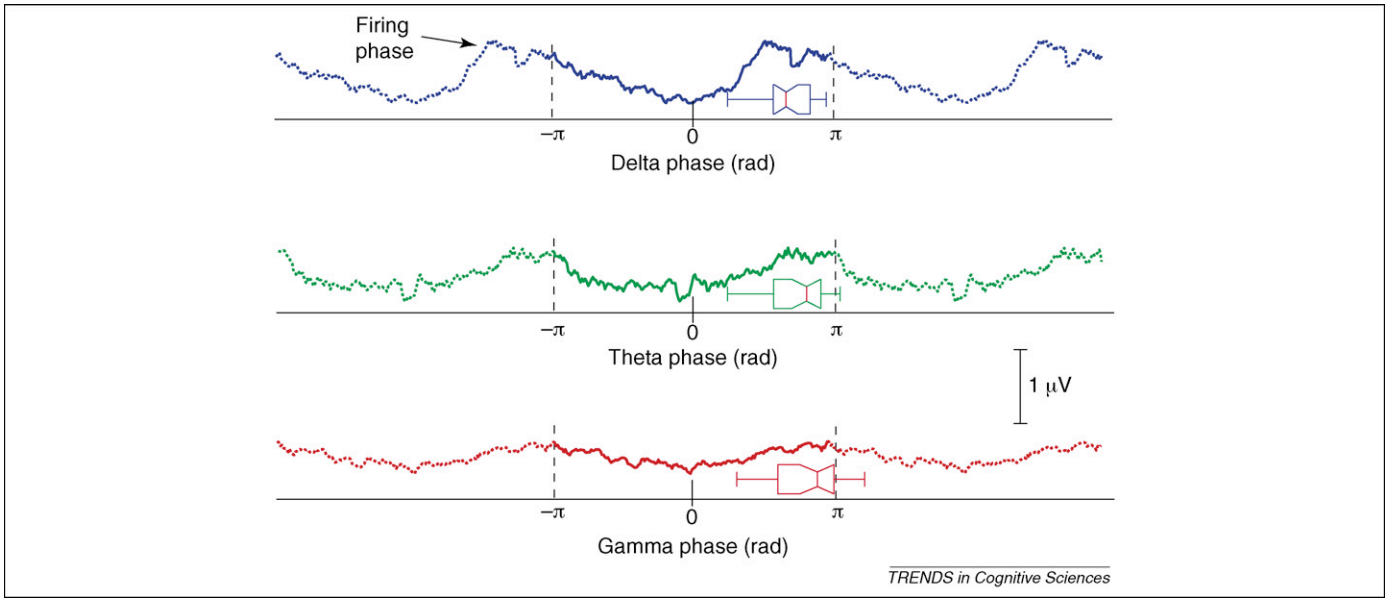

Figure 1. The relationship of oscillation phase to neuronal excitability. The three waveforms depict multiunit activity (MUA) amplitude as a function of the phase of spontaneous delta (1–4 Hz, black), theta (5–7 Hz, green) and gamma (25–50 Hz, red) oscillations, measured at a supragranular layer recording site in an individual experiment that sampled from the primary auditory cortex (A1) in an awake macaque monkey. MUA reflects the net action potential activity from neurons surrounding the recording site. In the absence of sensory input, MUA variations thus reflect net increases and decreases in the excitability of the local neuronal ensemble. MUA amplitude variations over three oscillation cycles are shown in each case. 'Firing phase' is the phase of the spontaneous oscillation currents during which neurons are most excitable, and therefore, most likely to generate action potentials (this corresponds to the largest MUA signals). Overlaid box and whisker plots show firing phase data pooled across all experiments (lines depict lower quartile, median and upper quartile values; whiskers depict the range of the observations). There is a clear phase-related modulation of the MUA amplitude in all the layers (the difference in MUA between the phase with maximal MUA ('firing phase') and the opposite phase was significant for all frequencies, in all cortical layers (Wilcoxon signed rank test, $p < 0.01$). Adapted, with permission, from Ref. [16].

input. Crucially, for near-threshold inputs, or for inputs occurring in a noisy auditory scene, these effects would determine whether or not inputs generate reliable post-synaptic responses.

Our recent study [11] demonstrates how somatosensory input, using the above rules, can enhance or suppress the processing of auditory inputs to A1. Note, however, that the stimuli used in Lakatos *et al.* [11] were brief and simple, whereas vocalizations tend to be extended and complex. That is where the fourth rule comes in (Figure 2d); oscillations at different frequencies tend to be phase–amplitude coupled in a hierarchical fashion [16]. The typical coupling we observe in the macaque A1 is that gamma frequency (25–50 Hz) amplitude varies systematically with the phase of an underlying theta (5–9 Hz) oscillation, and theta amplitude in turn is coupled to the underlying delta (1–2 Hz) phase. This coupling or 'nesting' of oscillation frequencies might reflect a general organizational principle, as evidence of coupling (mainly theta–gamma) has also been observed in humans [21,22], cats [23,24] and rats [25]. In any case, we think that oscillatory coupling makes it possible for non-auditory inputs to facilitate processing of the complex sound patterns in primate (including human) vocalizations, because the sounds are rhythmic and predictable, and their energy content is very similar to that of the oscillations in the auditory cortex.

Hierarchical coupling of neuronal oscillations

Coupling facilitates the processing of communication sounds for two reasons. First, the temporal amplitude envelope of vocalizations has a remarkable *a priori* match with intrinsic brain rhythms, both in humans and in monkeys. Perceptually salient envelope frequencies in normal human speech are focused below 16 Hz, particularly 4–8 Hz, and many transitions within the envelope occur over 20–30 ms periods [26,27]. These correspond to the ranges of theta and gamma oscillations, respectively, which are two of the three most prominent rhythms in the primary auditory cortex [16]. Also, because of the temporal structure of vocalizations, high-frequency events such as formant transitions (e.g. a consonant sound between two vowel sounds) are 'nested' within the lower frequency envelope. Most behaviorally-relevant temporal envelope features of monkey vocalizations are in these ranges, and this generalization holds across diverse primate species ranging from marmosets and squirrel monkeys to macaques [28,29].

Given the influence of delta phase on theta and gamma amplitudes, it is at first paradoxical that little of the energy in vocalizations is found in the delta range, but prosody (intonation and rhythm) is important in speech perception and it is conveyed at rates of 1–3 Hz [30], which corresponds to the lower delta oscillation band. Facial gestures and head movements often coincide with prosodic inflections in speech and increase their salience. Considering all these factors, the match between the temporal structure of ambient activity and the temporal structure of vocalizations is reasonably good. Because of this, there is a chance for the natural EEG rhythm to synchronize with the input pattern and amplify the cortical response, provided the phase and frequency of the EEG rhythm can be 'tuned' to the input pattern (the second reason why coupling facilitates the processing of communication sounds).

Regarding phase and frequency flexibility of the delta rhythm, earlier findings from our group indicate that both are indeed shaped by auditory stimuli [16]. Moreover,

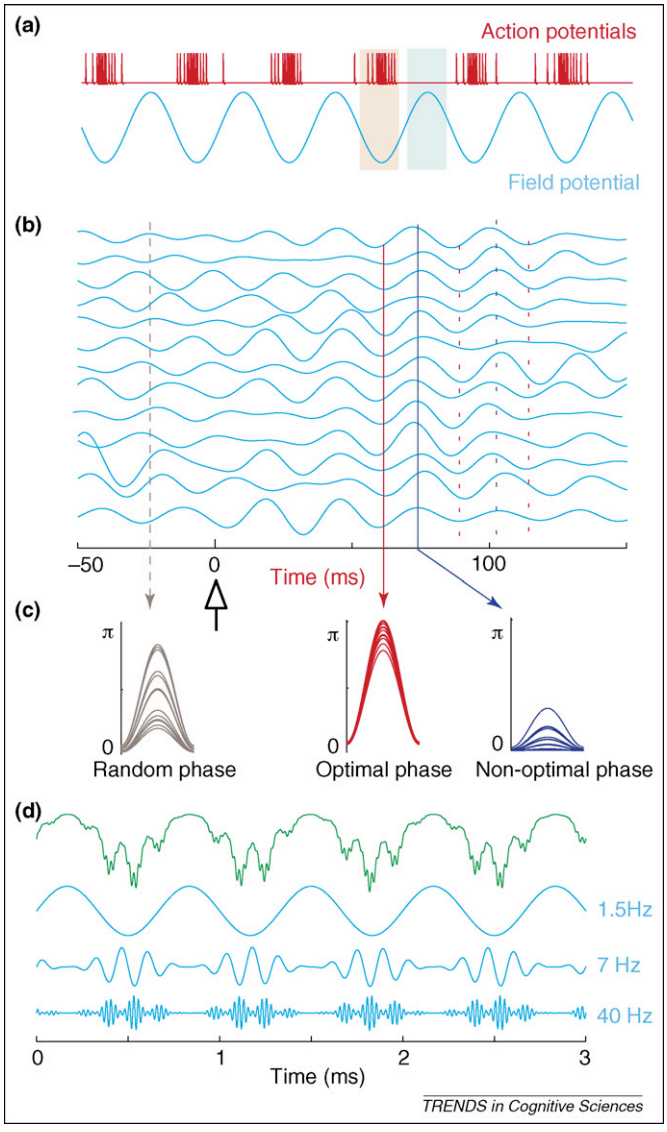

Figure 2. Functional consequences of oscillation phase. **(a)** The relationship between excitability, as indexed by the action potential firing rate (red), and the phase of oscillation in the local neuronal ensemble, as indexed by a local field potential (blue). From this type of experimental observation [11,16] we have proposed that ongoing neuronal oscillations have optimal (high excitability) and non-optimal (low excitability) phases. **(b)** A series of simulated single-trial responses representing activity in A1 as affected by visual inputs. When the system is at rest and unengaged (baseline pre-stimulus period to the left of zero) oscillations within a given frequency have a high degree of phase-variability across trials (gray dashed line). Presentation of a visual stimulus at time zero (arrow) can cause a phase reset of the ongoing oscillations, such that the oscillation develops strong phase coherence between trials; under these conditions, the optimal phases (red lines) and non-optimal phases (blue lines) align separately. **(c)** Sensory inputs arriving during the baseline (gray) generate highly variable response amplitudes. Inputs arriving during the optimal phase (red) are amplified, whereas those arriving during the non-optimal phase (blue) are suppressed. Over time, the cross-trial coherence dissipates, and the system goes back to its resting (random phase) state. **(d)** The top (green) trace illustrates the typical observation: oscillations recorded in the brain are normally complex mixtures of components at different frequencies. The traces below illustrate the individual oscillatory components in the delta (1.5 Hz), theta (7 Hz) and gamma (40 Hz) band that comprise the composite waveform. We and others have noted (see text) that in normal systems, there is strong phase–amplitude coupling between frequencies, and it has a hierarchical organization. Gamma oscillatory amplitude varies with the phase of the underlying theta oscillation, and theta oscillatory amplitude varies with the phase of the underlying delta oscillation. As explained in the text, we propose that this 'nesting' of higher-in-lower frequencies might optimize the processing of conspecific vocalizations, which have similar temporal structure.

there is independent evidence that the onset of a speech sound resets the phase of ongoing cortical rhythms in the auditory cortex [31], and that the frequency of cortical rhythms readily adapts to the rate of stimulation [16]. Finally, recent findings show that the phase of the auditory cortical theta oscillation not only tracks speech sound input patterns, but also predicts their intelligibility [32].

Obstacles to audiovisual integration in A1

If audiovisual integration is to operate as we propose, visual inputs have to reach the primary auditory cortex. There are at least three anatomical pathways by which this can occur, including a direct ascending (i.e. a nonspecific thalamic) input, a direct lateral connection from the visual cortex, and an indirect feedback input from the multisensory areas of the superior temporal sulcus (Box 3). A second issue is that for visual inputs to be effective, they must arrive in A1 slightly before auditory inputs. This is a challenge, because for nearby events in which auditory and visual components are perfectly synchronized, the auditory cortex is activated much faster than the visual cortex (Box 4). However, when producing a speech sound, orofacial movements usually occur well before any vocalization occurs (Figure 3). Also significant is that much of the prosodic (1–3 Hz) information available comes in the form of visual inputs generated by the vocalizer's head and hand movements [30], which, like articulatory gestures, generally precede audible vocalizations. Consideration of the timing factors operating in natural audiovisual communication (e.g. visual–auditory offset in Figure 3), along with the measured timing of neural responses at crucial locations in the brain (Box 4), suggests that several of the proposed pathways could convey visual inputs to the auditory cortex before the arrival of the associated auditory input.

Key questions and predictions

Our hypothesis raises several questions and predictions.

What determines the temporal integration window?

Because neuronal oscillations cover a wide frequency spectrum, from well below 1 Hz to well over 200 Hz [12], they enable the integration of inputs on many biologically relevant time scales. Consistent with this idea, we have shown that with very brief (100 ms) stimuli, auditory cortical ensembles integrate over a range of intervals that corresponds to half cycles of several low, middle and high frequency oscillations. Thus, 40 Hz oscillations would integrate inputs that arrive within the duration of their ideal phase (half the period of a 40 Hz oscillation is 12.5 ms), whereas theta and delta oscillations would integrate over correspondingly longer intervals (70–100 ms for theta and 125–250 ms for delta).

If very brief stimuli can be integrated over multiple windows, why then does audiovisual integration in speech sound processing have a much longer window [33]? For circumscribed events embedded in a stream (e.g. syllables and transitions), the duration of the event will determine the oscillation frequency that is most relevant, and its half cycle period will correspond to the integration window. This explanation is consistent with the idea that speech

Box 3. Underlying anatomical circuits for audiovisual integration in the auditory cortex

Because there is no evidence of strong feedforward inputs to the auditory cortex from the central visual pathways, investigators have generally speculated that visual modulation of auditory cortical processing, inferred from human neuroimaging studies, is accomplished by feedback from higher-order cortical structures [47]. Anatomical studies in monkeys show that, in addition to cortical feedback [48–50], there are two other equally plausible anatomical routes that visual inputs can use to access the auditory cortex: (i) feedforward projections from 'nonspecific' thalamic afferents, and (ii) direct lateral projections from the visual cortex [48,50–52] (Figure I).

The distinction between 'specific' and 'nonspecific' thalamic systems derives from the fact that in the specific system, thalamic neurons carry a dense and orderly representation of the sensory receptor surface, with sharp tuning for the dimension(s) coded spatially in the receptor surface (e.g. space for the retina and hand surface, and sound frequency for the cochlea). By contrast, those in the nonspecific system have a much less orderly representation of the sensory receptor surface, have large spatial or spectral receptive fields and are usually poorly tuned to the features encoded in the receptor surface. The sensory systems that project through the brainstem entail separate pathways (lemniscal and extralemniscal) for specific and nonspecific afferents, whereas at the thalamic level, the neurons of the two systems are differentially labeled by the calcium binding proteins parvalbumin (specific) and calbindin (nonspecific). The names of the systems are also descriptive of their central projections, with the specific system projecting through the hierarchical stages described in widely accepted models of sensory processing (e.g. V1, V2 and V4), and the non-specific system projecting widely to the neocortex, ignoring hierarchical progressions and often projecting outside of their systems of origin.

Examination of the timing and laminar profile of auditory cortical activation provides direct physiological evidence for both feedforward [53] and feedback [54,55] mechanisms of non-auditory influences on auditory processing. Phase resetting by non-auditory inputs seems to be strongly biased toward the supragranular laminae [11], which implicates feedforward input by nonspecific thalamic afferents as a potential causal element in the effect [56,57]. This possibility is particularly intriguing in light of the proposition that nonspecific (also variously known as 'extralemniscal,' 'matrix,' and 'Koniocellular') thalamic afferents could be uniquely important in promoting cortical synchrony [56]. However, it is also possible that feedforward, feedback and lateral circuits participate in interlocking aspects of visual–auditory integration in the auditory cortex. Although there are several credible alternatives, the specific anatomical substrate of these effects remains an open question.

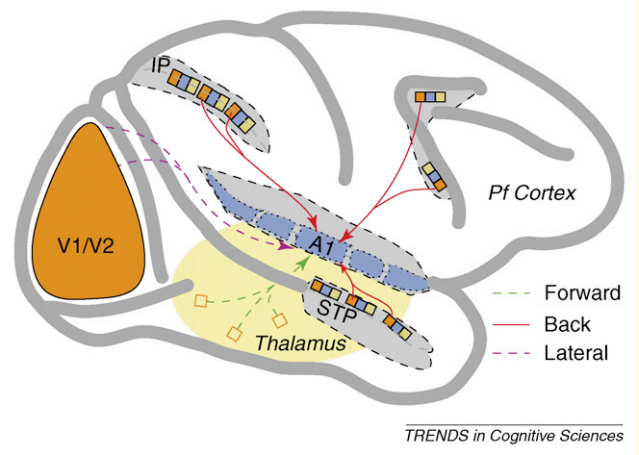

Figure I. A schematic summary of visual projections that terminate in and near A1, including feedback (red solid lines) from superior temporal polysensory (STP) area, prefrontal (Pf) cortex and Intraparietal (IP) areas, lateral projections (green dashed lines) from primary and secondary visual cortices (V1/V2) and feedforward (purple dashed line) inputs from nonspecific and higher order thalamic regions (yellow shading) such as supragenicular, posterior, anterior dorsal and magnocellular divisions of the medial geniculate complex, as well as portions of the pulvinar complex.

perception is a 'multi-time resolution process,' with integration windows that correspond to the time frame of delta, theta and gamma oscillations, [34,35], but the predictions of this explanation remain to be tested.

How would phase-reset amplification operate for extended vocalizations?

Phase-reset amplification should, under our hypothesis, be especially advantageous for more extended vocalizations (such as sentences), because of the remarkable match between the temporal organization pattern of speech and that of the hierarchically coupled, rhythmic oscillatory complex in A1. Consider a 'cocktail party conversation,' in which 'A' is talking to 'B' in the presence of numerous others, all conversing loudly. Here, B's ability to view A's visual gestures should be crucial for intelligibility. On the basis of the above considerations, with emphasis on those discussed in Refs [1,11,16,30,32], we would predict that in

Box 4. Key timing constraints

One complication that must be considered in audiovisual interactions is that the timing relationship of the cues changes as a function of distance, due to the relatively slow (343 m s^{-1}) speed of sound through air [38,58]. An additional complication inherent to biological phenomena such as audiovisual speech is that the visual cues often precede generation of an auditory output. For example, a swinging hammer provides visual movement cues well before it strikes a nail and produces a sound. Significant audiovisual lag is common in both human speech [6] and monkey vocalizations [59]; that is, a facial gesture (e.g. facial posture shift and/or mouth opening) generally precedes the generation of a vocalization. Field potential recordings in macaque A1 [36] indicate that the auditory cortex is tuned for a relatively specific visual–auditory lag; if the lag exceeds a certain value, multisensory enhancement shifts to suppression. From studies of response timing in the monkey auditory cortex [60] it is clear that auditory cortical responses to complex auditory transients (vocalizations) originating at conversational distances and intensities begin, on average at ~8.5 ms post-stimulus; at a nominal distance of 1.5 m, the conduction time of sound through air would take just under 5 ms (total of 13.5 ms). For relatively high contrast stimuli, mean visual latencies in higher-order visual areas that are known to send feedback projections to the auditory cortex (the intraparietal cortex, the medial temporal and medial superior temporal complex and the superior temporal polysensory cortex) are in the range 27–34 ms [45,55,61]. Extrapolating these to human values using a three-fifths rule (i.e. given that monkey latencies are generally about three-fifths of corresponding human values [62]) yields ~22 ms auditory cortical latency and ~45–57 ms visual cortical latency in humans. Even if the visual latency is doubled to allow for the feedback loop to the auditory cortex (to give ~90–114 ms), there is still sufficient time before auditory activation of the auditory cortex (i.e. a nominal 150 ms visual–auditory lag in speech plus 22 ms auditory cortical response latency) for visually driven feedback to precede and influence the processing of the auditory concomitants in the auditory cortex. Thus, for normal face-to-face audiovisual communication, feedback-mediated visual input will tend to arrive in the auditory cortex at or slightly before the arrival time of the input from the associated vocalization. This enables visual inputs related to a vocal communication to modulate its cortical representation.

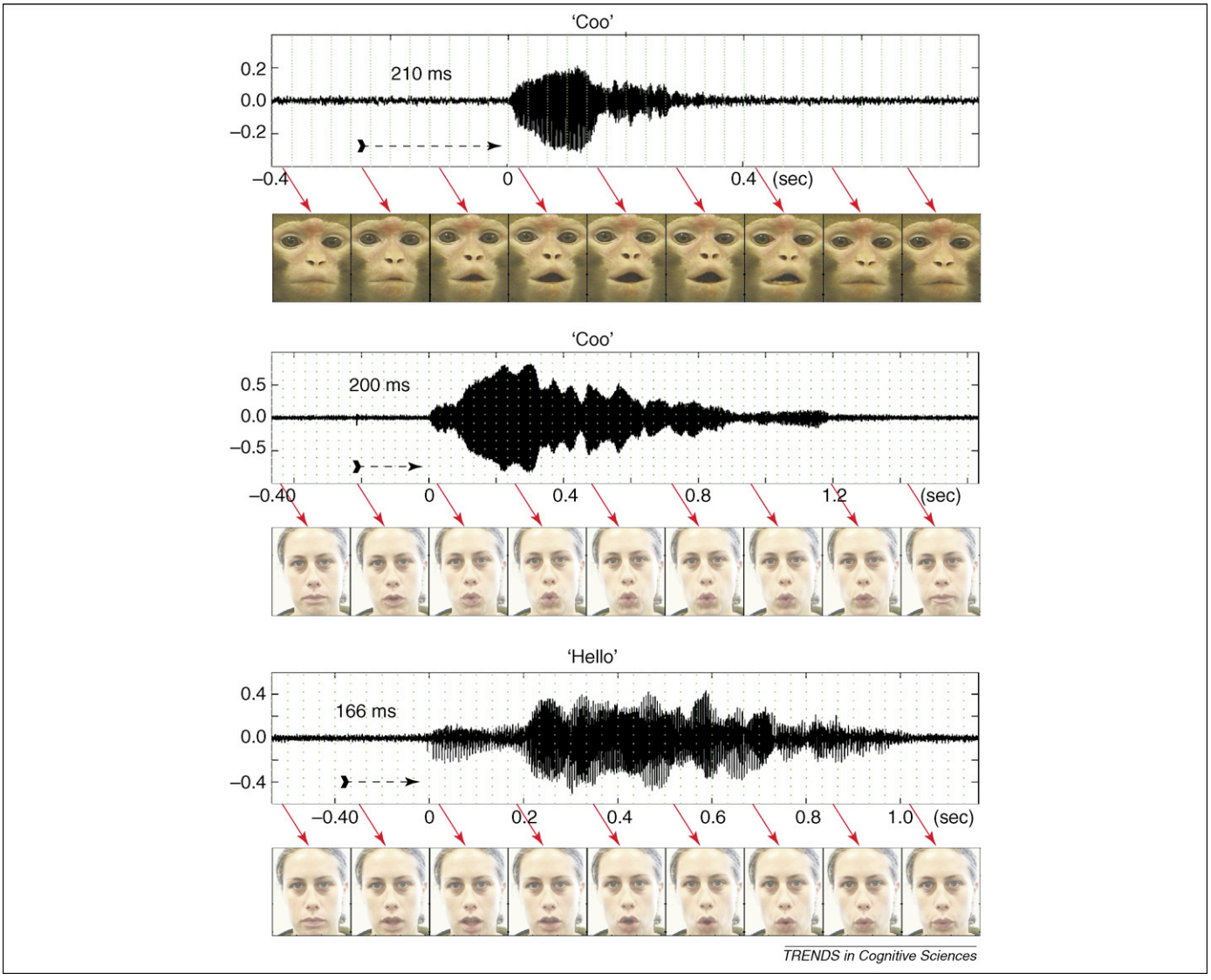

Figure 3. In order for visual inputs to modulate primary auditory processing, the ideal arrangement would be for visual inputs to arrive there before the time of auditory response onset. One factor that allows this to occur is the typical delay between visual articulatory gestures and the accompanying vocalizations. Examples of this visual–auditory offset are illustrated here, using a monkey making a 'coo' call (top), a human imitating this monkey coo (middle), and a human making a similar human vocalization ('hello', bottom). In each case, the auditory amplitude envelope of the call (sampled at 44.1 kHz) is displayed above a series of simultaneous video frames; these were acquired at 30 Hz (33.3 ms per frame), but only a key subset of the frames are shown, linked by arrows to the appropriate point in the auditory time line. The lag between the first detectable opening of the mouth and the onset of the auditory envelope function is displayed for each case (arrows and corresponding values in ms).

this situation dynamic resetting and driving of B's auditory delta oscillation by the pattern of prosodic expression in A's movements – particularly manual gestures, eye fixations and head inclinations – would be the key to visual–auditory facilitation in B's auditory cortex. In this sense, A controls B's brain rhythms, for a period of time. Because A's vocalizations (e.g. syllable strings) have theta band envelope frequencies and are coupled to A's prosodic expression, and because B's cortical theta oscillations are coupled to B's underlying delta oscillations (which are driven by A's prosodic expression), A's vocalizations should tend to align with the ideal excitability phase of B's theta oscillation, with a resulting response amplification. The coupling of A's vocal transitions to A's vocalization envelope, and of B's gamma oscillations to B's theta oscillations, predicts a similar tendency toward ideal phase alignment. Given the dynamics of this situation, the 'entrainment' of B's cortical oscillations to A's communicative expressions, the alignment can be only approximate. However, in the best case scenario, all the other vocalizations within earshot will have a random phase relationship to B's cortical oscillations (about half amplified in the 'optimal phase' and about half suppressed in the 'nonoptimal phase'), and their processing will suffer by comparison.

How important is attention?

In the above example, B's attention to A's communication is assumed. Overall, we would predict that the phase-reset amplification effects we describe would be highly sensitive to attention. One might then wonder whether the effects on vocalization processing (see, for example, Ref. [36]) of viewing the speaker are due solely to the arousing or attention-eliciting effects of the visual stimulus. It seems not, however, because in this case, as in many of the studies reviewed above, multisensory enhancement is extremely

sensitive to the introduction of small asynchrony in the stimuli. In fact, at certain stimulus onset asynchrony values, multisensory enhancement inverts into a suppressive effect [11,36]. The temporal sensitivity of multisensory effects on both perception and behavior does not suggest that attention and arousal are unimportant, but it discounts these variables as sole explanations for multisensory enhancement.

Concluding remarks

We propose that non-auditory inputs modulate primary auditory cortical processing by 'predictively' resetting the phase of the ongoing oscillatory cycles of the local neuronal ensembles. Visually induced phase reset places the oscillations in an ideal excitability phase, result in an amplified cortical response to associated vocalizations. This hypothesis makes reasonable empirical predictions but leaves open some important questions about primate (human and monkey) communication. These include the relative primacy of different visual cues (e.g. mouth versus head movements) in controlling auditory cortical oscillations and the degree to which the human–simian analogy extends into higher order semantic components of communication.

We expect that the phase-reset amplification mechanism we describe here will generalize beyond audiovisual communication. Across a wide range of real-world events, generally recognized as 'biological motion,' prominent non-auditory stimuli are generated before auditory stimulus onset because some visible action is required to produce a sound. For example, when we observe someone striking a nail with a hammer or running past us, the rhythmic temporal pattern of arm swinging or legs moving precedes and predicts the temporal pattern of hammer strike and footfall noises, particularly as the visual–auditory lag increases with distance. Visual cues often predict auditory events and are thus in a position to modulate auditory perception.

It is of fundamental importance that the rhythms of the natural environment have a striking parallel in the rhythms of neuronal oscillation in the brain. The fact that the internal oscillations can be driven by external events, and can influence neuronal processing of the same events, reinforces the view that they are instrumental rather than incidental to sensory processing.

Acknowledgements

We thank our colleagues T. McGinnis, M.N. O'Connell, A. Mills and A. Falchier for their assistance. We thank Asif Ghazanfar, Virginie van Wassenhove and Kevin Munhall for helpful comments and suggestions. Finally, we thank the editors and referees whose thoughtful advice and criticism lead to significant improvements in the paper. This work is supported the National Institutes of Health (MH 061989 and NS 049436).

References

1 Sumby, W.H. and Polack, I. (1954) Perceptual amplification of speech sounds by visual cues. *J. Acoust. Soc. Am.* 26, 212–215
2 Calvert, G.A. *et al.* (1997) Activation of auditory cortex during silent lipreading. *Science* 276, 593–596
3 Bruce, C. *et al.* (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* 46, 369–384
4 Besle, J. *et al.* (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234
5 Pekkola, J. *et al.* (2005) Primary auditory cortex driven by visual speech: an fMRI study at 3T. *Neuroreport* 16, 125–128
6 van Wassenhove, V. *et al.* (2005) Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1181–1186
7 Ghazanfar, A.A. and Schroeder, C.E. (2006) Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285
8 Schroeder, C.E. and Foxe, J. (2005) Multisensory contributions to low-level, 'unisensory' processing. *Curr. Opin. Neurobiol.* 15, 454–458
9 Meredith, M.A. (2002) On the neuronal basis for multisensory convergence: a brief overview. *Brain Res. Cogn. Brain Res.* 14, 31–40
10 Stein, B.E. and Meredith, M.A. (1993) *The merging of the senses.* MIT Press
11 Lakatos, P. *et al.* (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53, 279–292
12 Buzsaki, G. and Draguhn, A. (2004) Neuronal oscillations in cortical networks. *Science* 304, 1926–1929
13 Singer, W. and Gray, C.M. (1995) Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586
14 Shadlen, M.N. and Movshon, J.A. (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* 24, 67–77
15 Bishop, G. (1933) Cyclical changes in excitability of the optic pathway of the rabbit. *Am. J. Physiol.* 103, 213–224
16 Lakatos, P. *et al.* (2005) An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* 94, 1904–1911
17 Steriade, M. *et al.* (1993) Thalamocortical oscillations in the sleeping and aroused brains. *Science* 262, 679–685
18 Fiser, J. *et al.* (2004) Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature* 431, 573–578
19 Galambos, R. *et al.* (1981) A 40 Hz auditory potential recorded from the human scalp. *Proc. Natl. Acad. Sci. U. S. A.* 78, 2643–2647
20 Fries, P. *et al.* (2001) Rapid feature selective neuronal synchronization through correlated latency shifting. *Nat. Neurosci.* 4, 194–200
21 Canolty, R.T. *et al.* (2006) High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313, 1626–1629
22 Freeman, W.J. and Rogers, L.J. (2002) Fine temporal resolution of analytic phase reveals episodic synchronization by state transitions in gamma EEGs. *J. Neurophysiol.* 87, 937–945
23 Lakatos, P. *et al.* (2004) Attention and arousal related modulation of spontaneous gamma-activity in the auditory cortex of the cat.. *Brain Res. Cogn. Brain Res.* 19, 1–9
24 Steriade, M. *et al.* (1996) Synchronization of fast (30–40 Hz) spontaneous oscillations in intrathalamic and thalamocortical networks. *J. Neurosci.* 16, 2788–2808
25 Bragin, A. *et al.* (1995) Gamma (40–100 Hz) oscillation in the hippocampus of the behaving rat. *J. Neurosci.* 15, 47–60
26 Drullman, R. (1995) Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.* 97, 585–592
27 Luo, H. and Poeppel, D. (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010
28 Ghazanfar, A.A. and Hauser, M. (2001) The auditory behavior of primates: a neuroethological perspective. *Curr. Opin. Neurobiol.* 11, 712–720
29 Singh, N.C. and Theunissen, F. (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* 114, 3394–3411
30 Munhall, K.G. *et al.* (2004) Visual prosody and speech intelligibility: head movement improves auditory perception. *Psychol. Sci.* 15, 133–137
31 Ahissar, E. *et al.* (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13367–13372
32 Luo, H. and Poeppel, D. (2007) Phase patterns of neuronal responses reliably discriminate speech in auditory cortex. *Neuron* 54, 1001–1010
33 Munhall, K.G. *et al.* (1996) Temporal constraints on the McGurk effect. *Percept. Psychophys.* 58, 351–362
34 Poeppel, D. *et al.* Speech perception at the interface of neurobiology and linguistics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* (in press)
35 van Wassenhove, V. *et al.* (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607

36 Ghazanfar, A.A. *et al.* (2005) Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* 25, 5004–5012
37 Belin, P. *et al.* (2004) Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* 8, 129–135
38 Partan, S.R. and Marler, P. (2005) Issues in the classification of multimodal communication signals. *Am. Nat.* 166, 231–245
39 Andrew, R.J. and Huber, E. (1972) *Evolution of facial expression.* Arno Press
40 Rosenfeld, S.A. *et al.* (1979) Face recognition in the rhesus monkey. *Neuropsychologia* 17, 503–509
41 Marler, P. (1998) Animal communication and human language. In *Origin and Diversification of Language* (Jablonski, N.G. and Aiello, L.C., eds), pp. 1–19, California Academy of Sciences
42 Ghazanfar, A.A. and Logothetis, N. (2003) Facial expressions linked to monkey calls. *Nature* 423, 937–938
43 Kamachi, M. *et al.* (2003) Putting the face to the voice: matching identity across modality. *Curr. Biol.* 13, 1709–1714
44 Biben, M. *et al.* (1989) Contour variables in vocal communication between squirrel monkey mothers and infants. *Dev. Psychobiol.* 22, 617–631
45 Schroeder, C.E. *et al.* (1998) A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. *Cereb. Cortex* 8, 575–592
46 Logothetis, N.K. *et al.* (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157
47 Driver, J. and Spence, C. (1998) Crossmodal attention. *Curr. Opin. Neurobiol.* 8, 245–253
48 Hackett, T.A. *et al.* (2007) Multisensory convergence in auditory cortex II. Thalamocortical connections of the caudal superior temporal plane. *J. Comp. Neurol.* 502, 924–952
49 Romanski, L.M. *et al.* (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136
50 Smiley, J.F. *et al.* (2007) Multisensory convergence in auditory cortex I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *J. Comp. Neurol.* 502, 894–923
51 Galaburda, A.M. and Pandya, D.N. (1983) The intrinsic architectonic and connectional organization of the superior temporal region of the rhesus monkey. *J. Comp. Neurol.* 221, 169–184
52 Jones, E.G. (1998) Viewpoint: The core and matrix of thalamic organization. *Neuroscience* 85, 331–345
53 Schroeder, C.E. *et al.* (2001) Somatosensory input to auditory association cortex in the macaque monkey. *J. Neurophysiol.* 85, 1322–1327
54 Fu, K.M. *et al.* (2004) Timing and laminar profile of eye position effects on auditory responses in primate auditory cortex. *J. Neurophysiol.* 92, 3522–3531
55 Schroeder, C.E. and Foxe, J.J. (2002) The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Res. Cogn. Brain Res.* 14, 187–198
56 Jones, E.G. (2001) The thalamic matrix and thalamocortical synchrony. *Trends Neurosci.* 24, 595–601
57 Schroeder, C.E. *et al.* (2003) Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *Int. J. Psychophysiol.* 50, 5–17
58 Schroeder, C.E. and Foxe, J.J. (2004) Multisensory convergence in early cortical processing. In *The Handbook of Multisensory Processes* (Calvert, G.Aa *et al.*, eds), pp. 295–309, MIT Press
59 Hauser, M.D. *et al.* (1993) The role of articulation in the production of rhesus monkey, *Macacca mulatta*, vocalizations. *Anim. Behav.* 45, 423–433
60 Lakatos, P. *et al.* (2005) Timing of pure tone and noise-evoked responses in macaque auditory cortex. *Neuroreport* 16, 933–937
61 Chen, C.M. *et al.* (2007) Functional anatomy and interactions of fast and slow visual pathways in maccaque monkeys. *Cereb. Cortex* 17, 1561–1569
62 Schroeder, C.E. *et al.* (2004) Human-simian correspondence in the early cortical processing of multisensory cues. *Cogn. Process* 5, 140–151