Systems/Circuits

Auditory and Visual Modulation of Temporal Lobe Neurons in Voice-Sensitive and Association Cortices

Catherine Perrodin,[1] Christoph Kayser,[1,2] Nikos K. Logothetis,[1,3] and Christopher I. Petkov[1,4]

[1]Department of Physiology of Cognitive Processes, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany, [2]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, G12 8QB, United Kingdom, [3]Division of Imaging Science and Biomedical Engineering, University of Manchester, Manchester, M13 9PT, United Kingdom, and [4]Institute of Neuroscience, Newcastle University Medical School, Henry Wellcome Building, Newcastle upon Tyne, NE2 4HH, United Kingdom

Effective interactions between conspecific individuals can depend upon the receiver forming a coherent multisensory representation of communication signals, such as merging voice and face content. Neuroimaging studies have identified face- or voice-sensitive areas (Belin et al., 2000; Petkov et al., 2008; Tsao et al., 2008), some of which have been proposed as candidate regions for face and voice integration (von Kriegstein et al., 2005). However, it was unclear how multisensory influences occur at the neuronal level within voice- or face-sensitive regions, especially compared with classically defined multisensory regions in temporal association cortex (Stein and Stanford, 2008). Here, we characterize auditory (voice) and visual (face) influences on neuronal responses in a right-hemisphere voice-sensitive region in the anterior supratemporal plane (STP) of Rhesus macaques. These results were compared with those in the neighboring superior temporal sulcus (STS). Within the STP, our results show auditory sensitivity to several vocal features, which was not evident in STS units. We also newly identify a functionally distinct neuronal subpopulation in the STP that appears to carry the area's sensitivity to voice identity related features. Audiovisual interactions were prominent in both the STP and STS. However, visual influences modulated the responses of STS neurons with greater specificity and were more often associated with congruent voice-face stimulus pairings than STP neurons. Together, the results reveal the neuronal processes subserving voice-sensitive fMRI activity patterns in primates, generate hypotheses for testing in the visual modality, and clarify the position of voice-sensitive areas within the unisensory and multisensory processing hierarchies.

Key words: **audiovisual; congruency; face; multisensory; primate; voice**

Introduction

Social interactions often depend upon the receiver forming a coherent multisensory representation of voice and face content in communication signals. In primates, the temporal lobe contains voice-sensitive (Belin et al., 2000; von Kriegstein and Giraud, 2004; Petkov et al., 2008) and face-sensitive areas (Sergent et al., 1992; Tsao et al., 2008). Although functional magnetic resonance imaging (fMRI) studies in humans have shown voice/face multisensory interactions (von Kriegstein et al., 2005; Blank et al., 2011), primate face- and voice-sensitive neurons have been studied in their respective dominant sensory modalities, leaving unclear how fMRI multisensory influences relate to neuronal responses.

Audiovisual input is thought to be processed along sensory pathways that become progressively more feature-specific along the multisensory hierarchy (Schroeder et al., 2003; Ghazanfar and Schroeder, 2006; Kayser and Logothetis, 2007; Werner and Noppeney, 2010a). For instance, cross-modal interactions near primary auditory areas strongly depend on spatiotemporal stimulus alignment (Ghazanfar et al., 2005; Bizley et al., 2007; Lakatos et al., 2007) and can be influenced by attention (Schroeder and Foxe, 2005; Ghazanfar and Schroeder, 2006; Kayser et al., 2008; Lakatos et al., 2009). Cross-modal influences appear to become more feature-specific (Stein and Stanford, 2008; Werner and Noppeney, 2010a) in classically defined multisensory regions within the superior temporal sulcus (STS) (Barraclough et al., 2005; Werner and Noppeney, 2010b), intraparietal cortex (Linden et al., 1999; Avillac et al., 2007; Chen et al., 2013), and prefrontal cortex (Fuster et al., 2000; Sugihara et al., 2006; Romanski, 2007; Müller et al., 2011). For example, audiovisual interactions in the anterior STS are sensitive to cross-modal "object" features (Bruce et al., 1981; Calvert et al., 2000; Schroeder and Foxe, 2002; Beauchamp et al., 2004b) and reflect cross-modal stimulus feature congruency or informativeness (Barraclough et al., 2005; Dahl et al., 2010; Werner and Noppeney, 2010a).

Received July 1, 2013; revised Oct. 22, 2013; accepted Nov. 22, 2013.

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

This work was supported by the Max-Planck Society (C.P., C.K., C.I.P., N.K.L.), the Swiss National Science Foundation (C.P.), and the Wellcome Trust (C.I.P.). This work is dedicated to C. Stamm, who provided expert veterinary care. We thank M. Munk for encouragement and support.

The authors declare no competing financial interests.

This article is freely available online through the *J Neurosci* Author Open Choice option.

Correspondence should be addressed to Dr. Christopher I. Petkov, Institute of Neuroscience, Newcastle University, Framlington Place, Newcastle upon Tyne, NE2 4HH, United Kingdom. E-mail: chris.petkov@ncl.ac.uk.

DOI:10.1523/JNEUROSCI.2805-13.2014

Voice- and face-sensitive areas are positioned several anatomical stages beyond primary auditory/visual cortex but are lower in the multisensory processing hierarchy compared with regions, such as the STS (Maunsell and Newsome, 1987; Sergent et al., 1992; Petkov et al., 2008; Kikuchi et al., 2010). Such areas are not prominent in current models of multisensory convergence, unlike early sensory and association cortices (Stein and Stanford, 2008).

To clarify the position of voice/face areas in the multisensory processing hierarchy, we studied the specificity of neuronal auditory and audiovisual processing in an fMRI-identified voice-sensitive cluster, in the right anterior supratemporal plane (STP) of two Rhesus macaques. Results were compared with those from neurons in an adjoining multisensory region in the upper bank of the anterior STS. We first characterized the auditory sensitivity of neuronal responses to features such as vocalization "call type" and "voice identity" in communication signals. Using dynamic face and voice stimuli, we also quantified cross-modal interactions. The results supported the hypothesis that neurons in voice-sensitive STP, a high-level auditory area, are strongly involved in auditory analysis of vocal features. Visual influences in this area, although prominent, were less specific than in the STS, which, on the other hand, was less sensitive than the STP to auditory features.

Materials and Methods

Subjects

Two adult male Rhesus macaques (*Macaca mulatta*) participated in these experiments (S1: 10 years old, 17 kg; S2: 11 years old, 9 kg). The animals were in two separate group-housed colonies and worked with two different human scientists. Thus, some of the macaques' conspecifics and one of the human scientists would be familiar to one of the subjects but not the other and vice versa. This aspect of the subjects' environment was used in our experimental design to evaluate familiarity effects (see below). 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 carefully selected set of "coo" and "grunt" vocalizations by rhesus monkeys, and recordings of humans imitating monkey "coo" vocalizations (for details on the different caller-related factors included in the stimulus set, see Experimental design). All videos were recorded in the same sound-attenuated booth with the same lighting configuration, ensuring that each video had similar auditory and visual background. The stimuli were filmed while monkeys spontaneously vocalized, seated in a primate chair. The videos were acquired at 25 frames per second (640 × 480 pixels), 24 bits resolution, and compressed using Indeo video5. The audio tracks were acquired at 48 kHz and 16 bits resolution in stereo (PCM format). We selected the stimuli to ensure that the callers' head position and eye gaze direction were similar across all videos played within one experimental run. 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. 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. Image contrast and luminance for each channel (RGB) were normalized in all videos using Adobe Photoshop CS2. The video clips were 960 and 760 ms in duration, respectively, for the two main experiments (see below).

Auditory stimuli consisted of vocalizations that 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 2 free-field speakers (JBL Professional), which were positioned at ear level 70 cm from the head and 50 degrees to the left and right. Sound presentation was calibrated using a condenser microphone (Brüel and Kjær, 4188) and sound level meter (Brüel and Kjær, 2238 Mediator) to ensure a linear (±4 dB) transfer function of sound delivery (between 88 Hz 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 SPL within a sound-attenuating chamber (Illtec). The duration of the vocalizations was, on average, 402 ± 111 ms (mean ± SD; range: 271–590 ms).

Experimental design

For the main experiment, we recorded neural responses to 10 vocalizations each presented in auditory (A), visual (V), and audiovisual (AV) modalities (see Fig. 1*B*, examples) plus three incongruent audiovisual combinations (AVi) of pairs of some of these stimuli and two acoustically manipulated vocalizations. For data acquisition and analysis purposes, we broke up the large number of stimuli (35) into two subsets, each presented separately during the recordings, with at least 12 repeats of each stimulus condition. The first type of analysis focused on 3 subsets of 4 stimuli organized in the following 2 × 2 balanced factorial designs: AVCallType/Identity, AVSpecies/Familiarity, and AVCallerSize. The second type of analysis pooled the stimuli to assess visual influences on auditory responses across the datasets. The congruency conditions were analyzed separately (AVCongruency). Finally, a subset of the recordings also included acoustically manipulated vocalizations (Phase-scrambled vocalizations). It was not possible to balance all of the factors in a combined factorial analysis.

AVCallType/Identity. This subset of stimuli varied the "voice identity" (monkey 1, M1; and monkey 2, M2) and "call type" (coo/grunt) factors in a 2 × 2 factorial design by combining a "coo" and a "grunt" vocalization from two different callers. The familiarity of the callers was balanced across the subjects: one caller was familiar to the first subject but unfamiliar to the second, and vice versa.

AVSpecies/Familiarity. This 2 × 2 factorial design evaluates differences in the factors: "caller species" (monkey/human) and "caller familiarity to the listener" (familiar/unfamiliar). The species specificity was established by contrasting 2 monkey "coo" vocalizations and 2 humans imitating a monkey "coo" call. The caller familiarity factor was established in the following way: we considered a stimulus of a (monkey) caller familiar to the listener (subject) if they were living in the same colony with considerable audiovisual contact (shared or neighboring cage). The unfamiliar caller came from a different colony. Thus, we were able to "cross" the familiarity factor between the two monkeys by carefully selecting the stimuli used in the stimulus set (i.e., one individual was familiar to one subject and unfamiliar to the other, and vice-versa). We applied the same strategy when selecting the human stimuli: both humans used in the stimulus set are researchers, each of whom was working closely with one subject, and having limited contact with the other subject. Moreover, by asking humans to imitate monkey "coo" calls, we sought to obtain audiovisual human vocalizations that were similar to the conspecific calls in their low-level dynamic visual and acoustic characteristics but were unmistakably human voices. After pooling the neural responses from both subjects in the analysis, this meant that the "familiar" and "unfamiliar" categories actually contain the same acoustic stimuli, but familiarity is crossed with respect to the subject.

AVCallerSize. This subset of stimuli tested the influence of the "caller body size" (large/small) and "caller familiarity" (familiar/unfamiliar) factors, and contained 4 coo calls by four different callers. Callers defined as large were 10 and 12 years old, and weighed 10 and 16 kg, respectively. "Small" callers were 6 and 7 years old, and weighed 8 and 9 kg. The familiarity factor was crossed between participants as described above.

AVCongruency. To test the specificity of visual influences to behaviorally relevant (matching) voice-face pairs and whether they would be disrupted in response to mismatching voice-face pairs, we included incongruent audiovisual combinations of stimuli. We designed 3 mismatched audiovisual "control" pairs, created by combining auditory and visual versions of the previously detailed vocalizations. In particular, two mismatched pairs violated the "caller species" congruency of the stimu-

lus, by combining (1) a human voice with a monkey face and (2) a monkey voice with a human face. The third audiovisual mismatch violated the temporal synchrony of a monkey voice-face pair, by delaying the onset of the voice with a 340 ms temporal delay (see Fig. 7*A* for a schematic of the design). The global sensitivity to voice-face congruency was assessed by pooling together the 3 incongruent controls and comparing with the congruent versions. We also analyzed each of the 3 controls separately to assess the impact of individual congruency violations on neuronal responses.

Phase-scrambled vocalizations. To test whether units would be sensitive to disruption of the temporal pattern of vocalizations, some of our experiments also included phase-scrambled versions of a coo and grunt call from the AVCallType/Identity stimulus set. This acoustic manipulation was implemented by scrambling the phase of the vocalization stimuli in the Fourier domain, which removes the temporal envelope structure of the sounds while preserving the overall frequency spectrum. Thus, this last stimulus subset varied the "call type" (coo/grunt) and the "acoustic manipulation" factors (original vs phase-scrambled vocalizations).

Behavioral paradigm

Recordings were performed in a darkened and sound-insulated booth (Illtec, Illbruck Acoustic) while the animals sat in a primate restraint chair in front of a 21-inch color monitor. 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. Successful completion of a trial resulted in a juice reward. A trial began with the appearance of a central fixation spot. The eye position was measured using an infrared eye-tracking system (iView X RED P/T, SensoMotoric Instruments). 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 1200 ms stimulation period, and a 300 ms poststimulus recording time. Intertrial intervals were at least 1800 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. During the stimulation period, a visual stimulus (video sequence only), an auditory stimulus (audio track only, black screen) or an audiovisual stimulus was presented. The visual stimuli (dynamic, vocalizing primate faces) covered a visual field with a 15° diameter. The stimuli and stimulus conditions (such as modality) were randomly selected for presentation. Each stimulus presentation was repeated 12 times. Subject 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 (Perrodin et al., 2011). Subject 2 was accustomed from participating in the prior fMRI study 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 (4–6 consecutive trials, 8–12 s, 8–20° diameter fixation window), aborting the trial if eye movements breached this area. Only data from successfully completed trials in both animals were analyzed further.

Electrophysiological recording procedures

The two macaques had previously participated in fMRI experiments to localize their voice-preferring regions, including the anterior voice identity sensitive clusters (Petkov et al., 2008; Perrodin et al., 2011). A combination of neurological targeting software, fMRI, and stereotactic coordinates of the voice cluster centers, including postmortem histology at the end of the experiments, were used to guide or ascertain the electrophysiological recording electrodes to the voice-sensitive clusters in each animal (for details on the targeting procedures, see Perrodin et al. (2011)).

A custom-made multielectrode system was used to independently advance up to 5 epoxy-coated tungsten microelectrodes (FHC; 0.8–2 MOhm impedance). Electrophysiological signals were amplified using an Alpha Omega amplifier system (Alpha Omega), filtered between 4 Hz and 10 kHz (4-point Butterworth filter) and digitized at a 20.83 kHz sampling rate. Further details on the recording procedures have been reported previously (Perrodin et al., 2011).

The electrodes were advanced to the MRI-calculated depth of the anterior auditory cortex on the STP through an angled grid placed on the recording chamber. The coordinates of each electrode along the AP and mediolateral axes were noted, as were the angle of the grid and the depth of the recording sites. During a recording session, each electrode was advanced toward the STP. Auditory LFP and/or spiking activity for recording was identified as follows. Experimental recordings were initiated if at least one electrode had LFP 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 appeared 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. Unit responses were often obtained at two depths along one electrode penetration track, but most of our electrodes were inserted with anterior angles (5–15°). Thus, successive recording sites would not necessarily sample from the same neuronal microcolumn. Indeed, no obvious patterns of sound selectivity were seen when comparing neuronal responses at the two depths.

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 (i.e., the occurrence of $>$2 mm of white matter between auditory cortex and STS) and the emergence of visual evoked potentials at deeper recording sites.

Electrophysiological data preprocessing

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). For many sites, spike-sorting could extract well-isolated single-unit activity. We characterized clusters as single units if the waveform signal-to-noise ratio was larger than 4 (signal-to-noise ratio = average waveform peak amplitude/average waveform SD), combined with a clear refractory period ($<$1.5% of the total number of spikes occurring in the first 1.5 ms after a spike). For other sites where the spike-sorting did not yield well-separated clusters, the activity was combined into multiunit activity. To increase statistical power, for the spiking activity results we combined single and multiunit clusters for analysis, except when otherwise noted. Spike times were saved at a resolution of 1 ms. Peristimulus time histograms were obtained using 5 ms bins and 10 ms Gaussian smoothing (full-width at half-maximum).

Neuronal populations and subpopulations

We studied neuronal responses to our main audiovisual experiment within two brain regions: STP units are defined as the set of auditory responsive units sampled across the voice-sensitive fMRI cluster in the anterior STP. STS units are defined as the set of auditory responsive units sampled from recording sites in the upper bank of the STS (below the STP).

In addition to the main audiovisual experiment, the same units were also probed with a previously described, auditory "voice localizer" containing three categories of complex natural sounds: (1) conspecific monkey vocalizations (MVocs), (2) heterospecific animal vocalizations (AVocs), and (3) natural and environmental sounds (NSnds) (Petkov et al., 2008; Perrodin et al., 2011). Within the STP, we defined two subpopulations, based on the response of the units to the different sound categories: Voice/vocalization-sensitive units (VS) are defined as units that responded maximally to the MVocs category. Non–voice-sensitive units (non-VS) are defined as units that responded maximally to either the AVocs or the NSnds categories.

Data analysis

A significant response to sensory stimulation was determined by comparing the response amplitude of the average response to the response variability during the baseline period. The average response was normalized to standard deviation (SD) units with respect to baseline (i.e., *z*-scores), and a response was considered significant if the *z*-score exceeded 2 SDs during a continuous period of at least 50 ms during stimulus presentation. A unit was considered auditory responsive if its

Table 1. Summary of the sample size (number of responsive units) for each analysis and the different neuronal (sub)populations considered[a]

				Neuronal population			
				STP			
Type of analysis	Figure	No. of stimuli considered (modality)	Stimulus subset	All units	VS	Non-VS	STS
1. Stimulus factors	2, 5	4 (A, AV)	AVCallType/Identity	95	24	21	24
2. Stimulus factors	2	4 (A, AV)	AVSpecies/Familiarity	84	26	25	31
3. Stimulus factors	2, 5	4 (A, AV)	AVCallerSize	76	21	21	22
4. Visual influences	3, 4	10 (A, V, AV)	All vocalizations	159			67
5. AVCongruency	6, 7	9 (A, AV, AVi)	AVCongruency	123			57
		3 (A, AV, AVi)	Control 1	38			18
		3 (A, AV, AVi)	Control 2	41			20
		3 (A, AV, AVi)	Control 3	44			19
6. Phase-scrambling		4 (A)	Phase-scrambled vocalizations	26	7	8	0

[a]Rows identify the type of analysis in which each neuronal population was used and are referred to in Results. Also identified are the figures within which the particular analyses were reported.

activity breached this threshold for any of the experimental sounds in the considered set of auditory or audiovisual stimuli (Table 1). When characterizing sensory responses and visual influences in the main audiovisual experiment, we included all units that responded to sensory stimulation in any modality (A, auditory only; V, visual only; or AV, audiovisual) to account for the likely bimodal nature of some units, across the STS in particular. For subsequent analyses (factorial analyses: AVCallType/Identity, AVSpecies/Familiarity, AVCallerSize, Phase-scrambled vocalizations, and audiovisual controls: AVCongruency), we only included units responsive to an auditory stimulus presentation (A, AV or AVi), to have a fair comparison between auditory responsive neuronal populations in the STP and STS.

Units in the anterior STP have been previously shown (Kikuchi et al., 2010; Perrodin et al., 2011) to be highly stimulus-selective, and to respond to a minority of the presented vocalizations. Thus, we chose for each analysis to only include the units that significantly responded to at least one stimulus in the relevant subset. This approach accounts for the high stimulus selectivity of units in the studied area and prevents the inclusion of unresponsive units, yet yields different sample sizes for each of the stimulus subset considered, which are summarized in Table 1. It is important to note that analyses restricted to subcategories of units (e.g., non-VS units with a significant sensitivity to voice identity) can result in small sample sizes that might not be representative in the results from both of the animals. To confirm that the results were supported by data from each animal, all analyses were also performed on each monkey's dataset separately. It was observed that all of the main conclusions regarding STP/STS sensitivity to the auditory factors, multisensory influences, and the STP's voice identity sensitivity in the subpopulations of VS and non-VS units are supported by the results from each animal. This also justified pooling the data from the two animals for analysis and reporting here.

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 by first computing the mean response for each stimulus across trials. The peak of the stimulus response was calculated, and the response amplitude was defined as the average response in a 400 ms window centered on the peak of the stimulus response.

The auditory response onset latency was computed for each unit by taking the average auditory response to all sounds the unit was responsive to, calculating a *z*-score relative to the baseline firing rate, and identifying the first time point after sound onset where the response exceeded our response criterion (2 SD for at least 50 consecutive milliseconds).

Multisensory interactions were assessed individually for each unit with a significant response to sensory stimulation (A, V, or AV). We considered a sensory responsive unit "visually influenced" if it was classified as either "bimodal" or "nonlinear multisensory." Bimodal units were defined as exhibiting a significant response to both A and V presentations of a stimulus. A unit was termed "nonlinear multisensory" if its response to the audiovisual stimulus was significantly different from 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 (Stanford et al., 2005; Kayser et al., 2008; Dahl et al., 2009): 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 step for $N = 1000$ 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 (FDR) correction for multiple comparisons was applied to all p values (Benjamini, 1995). All statistical tests used were two-tailed.

The direction and amplitude of the deviation from additivity were quantified using the following index: $\text{Additivity} = 100 \times (AV - (A + V))/(A + V)$, where A, V, and AV reflect the baseline-corrected response amplitude, averaged in a 400 ms window. Positive (negative) values of the additivity index indicate superadditive (subadditive) visual modulations of the auditory response.

The time course of superadditive and subadditive visual modulation was estimated as follows. For each nonlinear multisensory unit, we computed the $AV - (A + V)$ difference between the audiovisual and summed unimodal responses. This difference time course was averaged separately for units that showed enhanced or suppressed responses. We then converted it into SD from baseline. The onset (respectively offset) of the multisensory effect was defined for each unit, as the first (last) time point during which the time course breached 2 SD of its baseline level for at least 10 ms.

Friedman's two-way ANOVA was performed on the responses to different subsets of 4 stimuli, each organized in a 2×2 factorial design, to explore the effects of several stimulus factors on neuronal responses (see Experimental design). The nonparametric tests were chosen to account for our non-normally distributed data. More specifically, repeated-measures (RM-) ANOVAs were used to test for differences in two levels of one factor of interest (column factor) while accounting for potential effects of the second factor (row factor/nuisance effect). For each of the studied subsets, all units that significantly responded to the auditory presentation of at least one of the four stimuli in the set were included in the analysis. For the population analyses, the nonparametric RM-ANOVA tests used trial-averaged response amplitudes as observations, with the different responsive units as replicates. When looking at individual units, the analyses of variance used trial-based response amplitudes as observations, with the different trials as replicates (n = 8–12 repetitions). Neuronal responses tested were the auditory response amplitudes (A) and the magnitude of the nonlinear audiovisual influences (measured as the absolute value of the additivity index; see previous paragraph and Fig. 4). The population response was considered to be sensitive to a given factor if the analysis was significant at $p < 0.05$. Individual units were considered to be sensitive to a given factor if the analysis was significant at $p < 0.05$, with an FDR correction for multiple comparisons. We only report results for which a given factor significantly modulated the responses of at least 5% of the tested units. Wilcoxon rank sum tests were used for *post hoc* comparisons of the Friedman ANOVA.

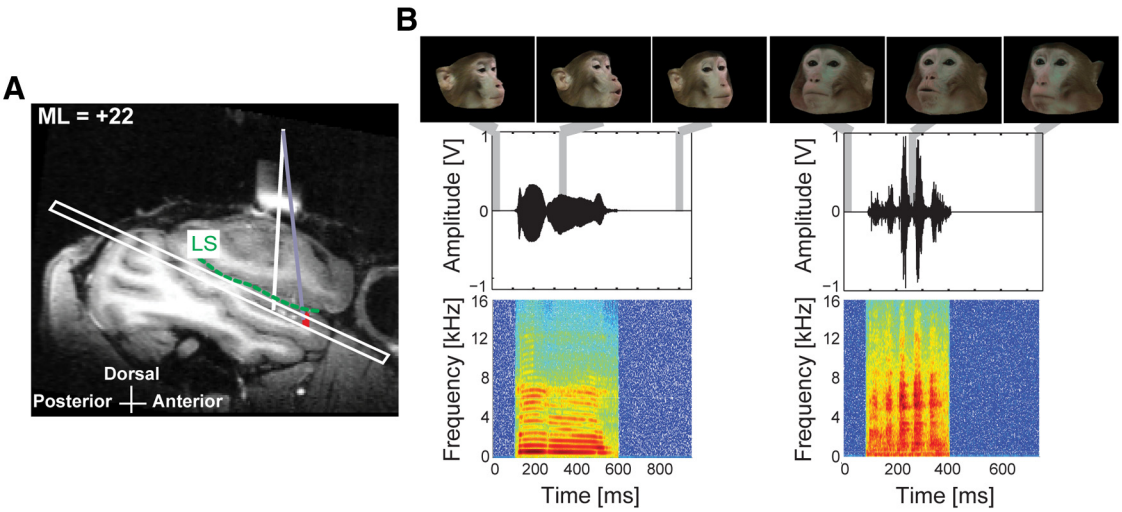

Figure 1. Localization of recording sites and audiovisual voice-face stimuli. ***A***, Sagittal structural magnetic resonance image (MRI) of the liquid-filled recording chamber (white bar above brain), with vertical white line projecting to the STP below the lateral sulcus (LS). The image is located at +22 mm mediolateral (ML) using the Frankfurt-zero standard. Stereotaxic coordinates and a neurosurgical targeting system guided electrode placement to the anterior fMRI voxels (red) with a strong preference to conspecific voices over other complex natural sounds. ***B***, Two examples of audiovisual rhesus macaque vocalizations used for stimulation: a coo (left) and a grunt call (right). The video starts at the onset of mouth movement. 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.

Scheirer-Ray-Hare tests (Sokal and Rohlf, 1995) were performed to assess interaction effects between identified stimulus features and the two brain areas (STP and STS) on the neuronal population responses. They are nonparametric multifactorial analyses of variance (a generalization of the Kruskal–Wallis test) for assessing the impact of the 2 factors: identified stimulus feature and brain area (STP vs STS), including their interactions.

Characterization of auditory temporal response profiles was done by taking, for each unit, the average auditory response to the two grunt calls from the AVCallType/Identity stimulus subset. After calculating a *z*-score relative to the baseline firing rate, we divided the poststimulus time into three intervals of equal duration (25–175, 175–325, 325–475 ms after sound onset) and identified in which interval(s) the response exceeded 2 SD for at least 10 consecutive milliseconds. Responses were characterized as "phasic onset responses" if they breached the criterion during the first one or two time intervals. Responses were characterized as "offset responses" if they breached the criterion during the third interval only, or during the third interval and any one of the other two. Reponses were characterized as "sustained" if they breached the criterion during the middle interval only, or during all three.

Spontaneous firing rates were computed for well-isolated single units (SU) by computing the average firing rate in a 400 ms window preceding stimulus onset (before baseline subtraction). The waveform duration of units' action potentials was computed as the time between the trough and the peak of each action potential waveform (Mitchell et al., 2007).

Results

We targeted neurons for extracellular electrophysiological recordings in a previously identified voice-preferring fMRI cluster (Petkov et al., 2008; Perrodin et al., 2011) in the right hemisphere of two rhesus macaques (Fig. 1*A*). This area resides in anatomical areas Ts1/Ts2 on the supratemporal plane, which are anatomical areas anterior to the tonotopically organized auditory core and belt fields (Kaas and Hackett, 2000; Petkov et al., 2008; Perrodin et al., 2011). The anterior STP receives particularly dense afferent inputs from the adjacent anterior belt and parabelt areas (Galaburda and Pandya, 1983; Romanski et al., 1997; Hackett et al., 1998) as well as afferent input from area TPO in the STS (Cipolloni and Pandya, 1989). We also recorded from a population of auditory responsive units in the upper bank of the anterior STS, ventral to the STP recording sites.

Sensitivity of auditory responses to stimulus features

Unlike the well-described auditory fields along the STP, the role of the anterior STP in auditory processing has been less investigated (Kikuchi et al., 2010; Perrodin et al., 2011). Until recently, whether the anterior STP belonged to auditory cortex in rhesus macaques was uncertain based solely on anatomical studies (Galaburda and Pandya, 1983; Hackett et al., 1998). In contrast, the STS is a classically multisensory association cortex, thought to be involved in multisensory representations. One hypothesis is that neurons in the STP, more so than those in the STS, would be primarily involved in the analysis of auditory features, including distinguishing between voice identity or call type aspects. Alternatively, the STS is known to have auditory responsive clusters of neurons and thus might be engaged in both auditory and visual feature analysis (Beauchamp et al., 2004a; Dahl et al., 2009). Our results supported the former hypothesis and identified a neuronal subpopulation particularly sensitive to voice identity in the population of STP units that were recorded from.

Our experimental design factorially varied vocal features, such as call type, caller identity, species, size, and familiarity to the listener (AVCallType/Identity, AVSpecies/Familiarity, and AVCallerSize; see Materials and Methods). Using 2-way nonparametric Friedman's ANOVA tests on the population of STP units, we initially tested whether a global preference to certain auditory factors was observable across the STP (Table 1, rows 1–3, column 5 for sample sizes related to these analyses). Within the STP, the call type factor significantly modulated the population of auditory responsive units ($n = 95$ auditory responsive STP units; Friedman RM-ANOVA; main effect of call type: $\chi^2(1) = 14.3$, $p = 2.50 \times 10^{-4}$). Here, grunts elicited larger responses than coo calls (Fig. 2*A*). Representation of the several other vocal factors in our experimental design was not evident in the population response ($p > 0.05$), presumably because of heterogeneity in the responses of individual units, which we next studied using Friedman's ANOVA (FDR-corrected) on individual unit responses.

The unit analyses showed that across the STP considerable proportions of units were significantly sensitive to the following

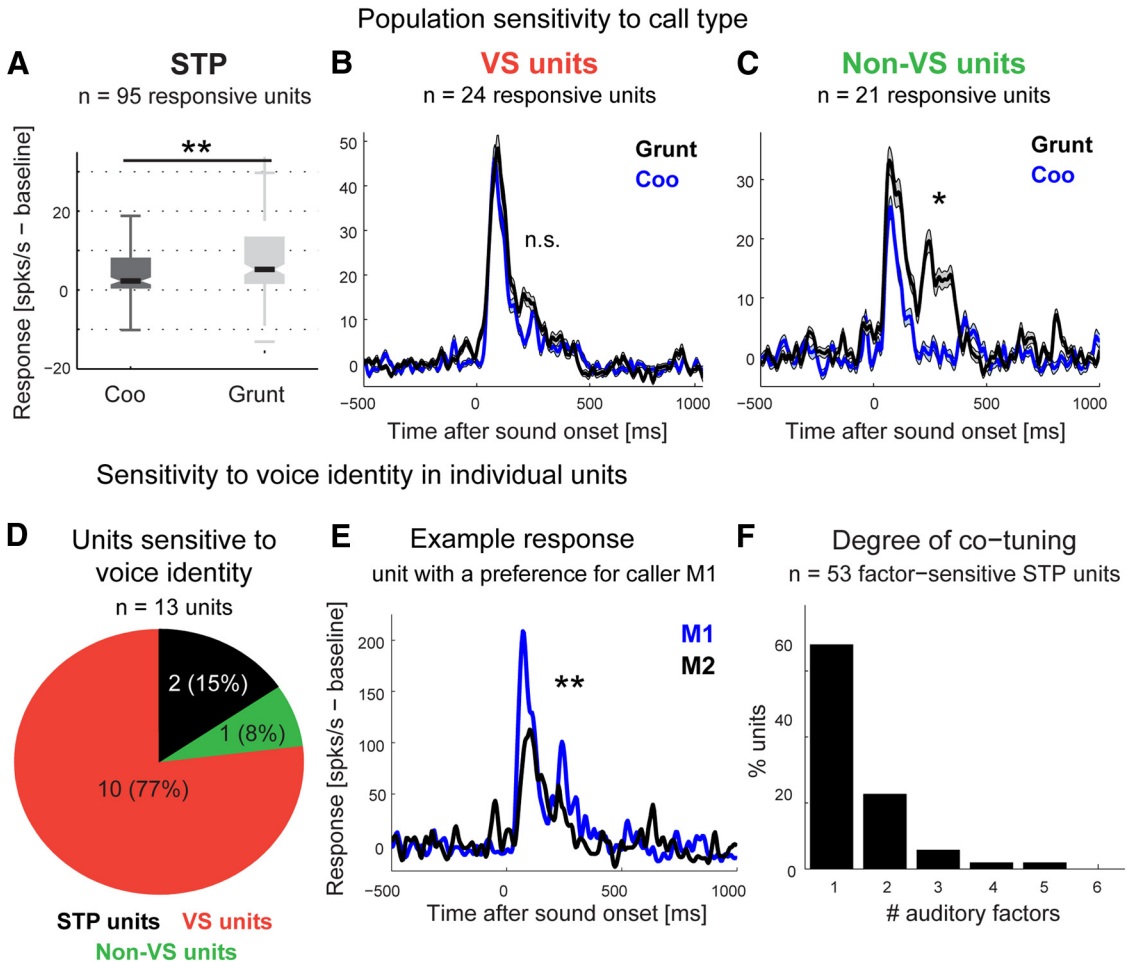

Figure 2. Auditory processing of call type and voice identity in STP neuronal subpopulations. ***A***, *Post hoc* comparisons (Wilcoxon rank sum test) of the effect of the call type factor on the median auditory responses across the population of STP units ($n = 95$ auditory responsive units). Boxplots represent the median, upper, and lower quartiles of the population auditory responses. ***B***, Grand average responses to the two coo (blue trace) and the two grunt (black trace) call type exemplars by callers M1 (monkey 1) and M2 (monkey 2), averaged across the population of VS units ($n = 24$ auditory responsive units). Traces represent mean ± SEM. ***C***, Grand average responses to coo and grunt calls, averaged across the population of non-VS units ($n = 21$ auditory responsive units). ***D***, Functional characterization of units showing a significant effect for voice identity in the STP ($n = 13$ auditory responsive units). ***E***, Example response of a voice identity sensitive unit from the VS subpopulation, with stronger responses to vocalizations by caller M1. Each trace is the average response to a coo and a grunt call type exemplar from one caller (blue represents M1; black represents M2). ***F***, Degree of co-tuning to different auditory factors in STP units. Asterisks indicate significant effects of the specified factor in a balanced 2-way nonparametric Friedman's test, with different auditory responsive units (***A–C***) or trials (***E***) as repetitions. $^{**}p < 0.01$ (Wilcoxon rank sum test). $^{*}p < 0.05$ (Wilcoxon rank sum test). n.s., Not significant.

factors, ordered by the proportions of units seen, as follows: (1) call type, (2) caller species (human or monkey), (3) voice identity, (4) familiarity, and (5) caller size. First, sensitivity to call type was evident in the auditory responses of 23% of auditory responsive units across the STP (call type factor significant in 22 of 95 of auditory responsive STP units; $p < 0.05$, FDR-corrected). Second, 20% of neurons showed sensitivity to the species of the caller (monkey or human) that produced "coo" calls (17 of 84 auditory responsive STP units; $p < 0.05$, FDR correction). Third, a considerable proportion of STP units were also modulated by the voice identity factor (13 of 95 = 14% of auditory responsive units; $p < 0.05$, FDR-corrected; Fig. 2*D*). Here, for example, neuronal responses were more similar across the two acoustically distinct "coo" and "grunt" call types produced by the same individual than they were to the same call type produced by different individuals. A significant sensitivity to the familiarity factor was seen in 11% of neurons, for all stimuli in which this factor was available (18 of 160 auditory responsive STP units; pooled results from the AVSpecies/Familiarity and AVCallerSize subsets; Table 1, rows 2 and 3). Finally, the body size factor modulated 9% of STP units (7 of 76 auditory responsive STP units; $p < 0.05$, FDR correction). We also noted that a minority of the STP units sensitive to a given auditory factor (12 of 53 = 23% of factor-sensitive STP units) displayed co-tuning to combinations of two vocal features. However, the majority of STP units (36 of 53 = 68%) were sensitive to one feature, showing little co-tuning to the auditory factors (Fig. 2*F*).

In contrast to the STP units, none of the stimulus factors significantly modulated the population of STS units (Friedman's ANOVA, all $p > 0.05$; Table 1, rows 1–3, column 8 for sample sizes related to these analyses). In addition, auditory responses of STS units were sensitive to much fewer auditory factors: caller size (4 of 22 = 18% of auditory responsive STS units) and call type (4 of 24 = 17%; all other factors $p > 0.05$; FDR correction). None of the STS units showed any co-tuning and were sensitive to one acoustic feature at a time. The spontaneous firing rates of well-isolated single units did not differ between both recording locations (STP: $n = 60$ SU, 4.6 ± 0.9 spks/s (mean ± SEM), STS: $n = 23$ SU, 3.5 ± 0.5 spks/s; paired-sample t test, $p > 0.4$).

Testing the acoustical factor sensitivity of the STP and STS directly, we observed that auditory features modulated a significantly larger number of STP units than STS units (χ^2 test =

372.25, $p < 0.001$). Together, our results indicate auditory sensitivity to a number of features in communication signals in STP neurons, which was not evident in the STS.

Functionally distinct auditory neuronal subpopulations in the STP: voice identity versus call type

To further probe the auditory processes of neuronal subpopulations in the STP, we evaluated neuronal responses to the different auditory factors by subdividing the STP neuronal population based on response properties assessed in a "voice localizer" experiment (Petkov et al., 2008; Perrodin et al., 2011), which was conducted separately. Using the localizer experiment, we identified VS versus non-VS units (Perrodin et al., 2011), as follows: VS units were characterized by a categorical response preference for conspecific vocalizations from many different callers (MVocs) over other types of vocalizations or complex natural sounds (Table 1, rows 1–3, column 6 for sample sizes related to the factorial analyses on VS units). Non-VS units were defined as units preferentially responding to heterospecific vocalizations or natural sounds (Table 1, rows 1–3, column 7).

The VS units alone seemed to account for much of the observed STP sensitivity to voice identity. The auditory response of 42% (10 of 24) of VS and 5% (1 of 21) of non-VS units was significantly modulated by the voice identity factor ($p < 0.05$, FDR correction; Fig. 2*E*, example response). Notably, from the 13 voice identity-sensitive units identified in the STP, most belonged to the subset of VS units (10 of 13 = 77%; Fig. 2*D*, red section).

Sensitivity to the call type factor was prominent in both the VS and non-VS units. The call type factor significantly modulated 24% (5 of 21) of non-VS neurons, and 38% (9 of 24) of VS neurons. Interestingly, the subset of non-VS units, but not that of VS units (Fig. 2*B*), also displayed a population preference for grunts over coos (main effect of call type on the population of non-VS units: $\chi^2_{(1)} = 4.94$, $p = 0.026$; Fig. 2*C*), which is comparable with the overall STP population call type effect. Both VS and non-VS units reflected the remaining vocal factors in fairly comparable proportions (caller species: 23%, 6 of 26 of VS units and 32%, 8 of 25 of non-VS units, respectively; caller size: 19%, 4 of 21 and 14%, 3 of 21 of units, respectively; caller familiarity: 19%, 9 of 47 and 11%, 5 of 46 of units, respectively). Figure 2 illustrates the differential neuronal representations of call type and voice identity features in these subpopulations of STP neurons: A prominent sensitivity to call type is observed in both subpopulations, and grunts elicited larger responses than coos in non-VS units (Fig. 2*C*). However, the units sensitive to voice identity were more likely to belong to units classified as VS (Fig. 2*D*).

Further characterization of VS and non-VS units' acoustic response properties revealed that both neuronal populations differed in their sensitivity to temporal dynamics of vocalizations. We first investigated the units' sensitivity to acoustic control stimuli that randomized the phases of the vocalization stimuli, which was designed to disrupt the temporal envelope of the vocalization stimuli but preserve the overall frequency spectrum (for similar manipulations, see Petkov et al., 2006, 2008). We compared the auditory responses of the units to intact versus phase-scrambled versions of two vocalizations (Table 1 row 6 for sample sizes related to this analysis). Units with a significant sensitivity to the acoustic manipulation were only found in the non-VS subset (3 of 8 = 38% units), whereas none of the VS units responded differentially to the original and the phase-scrambled vocalizations. This suggests that VS units are less likely to be affected by disruptions in the temporal dynamics of the vocalizations, consistent with the notion that a prominent acoustical cue of voice identification is present in the spectral filtering of the vocal tract (Fitch and Fritz, 2006, Ghazanfar et al., 2007), which our acoustical control preserves in the vocalization stimuli.

Next, we studied the temporal response profiles of individual unit responses to grunts, as vocalizations with strong sound envelope modulation (see an example vocalization in Fig. 1*B*, average responses (black traces) in Fig. 2*B*,*C*, and Table 1 row 1 for sample sizes related to this analysis). The majority of VS units (13 of 24 = 54% of auditory responsive VS units), but few non-VS units (7 of 21 = 33% of auditory responsive non-VS units) displayed phasic-onset responses. In contrast, the majority of non-VS units (12 of 21 = 57%) showed sustained responses, whereas such responses were less prominent in the VS units (9 of 24 = 38%). Phasic-offset responses proportions were comparable between the VS and non-VS units (respectively, 2 of 24 = 8%; 2 of 21 = 10%). Thus, the typical response profiles differed between VS and non-VS subsets, with VS units favoring phasic-onset responses, and sustained/envelope-following responses in non-VS units (χ^2 test = 5.9, $p = 0.015$). Finally, complementing the analysis on temporal response profiles, we computed the timing of the peak spiking response. We found that responses of non-VS units peaked later after sound onset than those of VS units (VS: mean peak latency = 134 ± 10.8 ms, non-VS: mean = 189 ± 16.0 ms, paired-sample *t* test: $p = 0.0031$). This confirms the prominence of early, onset-type responses by VS units. In contrast, non-VS units preferentially responded with sustained temporal profiles.

Beyond these differences, similarities were observed between the VS and non-VS subsets of units. First, VS and non-VS units displayed comparable stimulus selectivity to monkey vocalizations, animal vocalizations, and natural sounds. Overall, units from both subsets responded strongly to a select 17% of the presented complex sounds (6 of 36, median number of sounds eliciting response amplitudes larger than the half-maximum response), with no differences in selectivity between the VS and non-VS units for any of the sound categories. Computing the average auditory response onset latency separately for VS and non-VS units in the STP showed no differences (paired-sample *t* test, $p > 0.5$): the VS units had auditory response latencies of 100 ± 29.6 ms after sound onset (mean ± SEM), and non-VS units responses started at 92 ± 19.7 ms. Auditory response latencies of STS units were 109 ms ± 22.4 ms and did not differ from those in the STP (paired-sample *t* test, $p > 0.5$). Finally, we compared the action potential waveform durations of the single units. Action potential waveform durations from both neuronal subsets followed a bimodal distribution, but the proportions of narrow versus broad duration spikes did not differ between VS and non-VS subpopulations (χ^2 test, $p > 0.05$).

In summary, we identify two functionally distinct neuronal subpopulations in the STP using a previously described "voice localizer". VS units seem to support the area's sensitivity to voice identity. In relation to non-VS units, the VS units were less sensitive to disruptions in the temporal structure of the vocalizations and showed a preponderance of phasic responses.

Visual influences on auditory activity

We assessed to what extent auditory responses are modulated by visual information from face content. Neurons in the anterior STP have not, to our knowledge, been previously probed with audiovisual stimuli (Fig. 1*B*, example stimuli). The STS has been better studied in this regard.

In the STP, we recorded spiking activity in response to either auditory or visual input ($n = 159$ single units and multiunits respon-

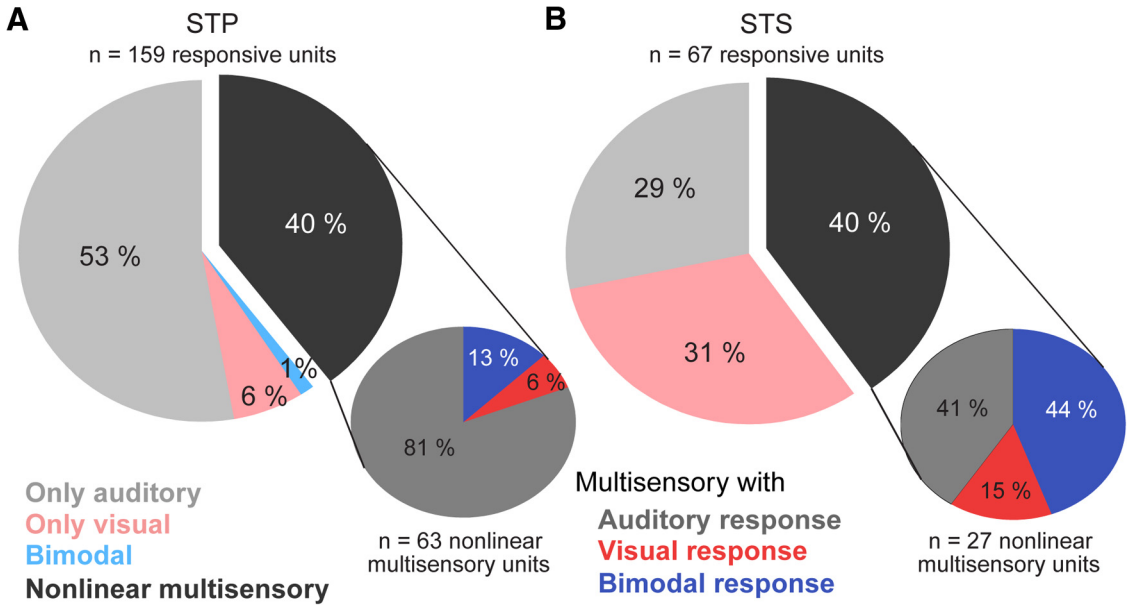

Figure 3. Visual influences on auditory responses in STP and STS units. ***A***, Summary of the type of sensory responses in STP units ($n = 159$ sensory responsive single units and multiunits). Inset, Distribution of sensory responsiveness of the subset of visually modulated units (main pie chart, dark gray area). Colored sections indicate units with cross-modal responses. ***B***, Summary of visual influences in STS units ($n = 67$ sensory responsive units).

sive to auditory, visual, or audiovisual stimuli; Table 1 row 4, column 5). The majority of sensory responsive STP units were auditory (84 of 159 = 53%; Fig. 3*A*, light gray section), whereas very few responded to purely visual stimulation (10 of 159 = 6%; Fig. 3*A*, pink section). Unlike the primarily auditory responses in the STP, STS units were as likely to be purely auditory (19 of 67 = 29% of sensory responsive STS units; Table 1 row 4, column 8; Fig. 3*B*, light gray section) as purely visual (21 of 67 = 31%; Fig. 3*B*, pink section).

In the STP, 41% of responsive units demonstrated different types of visual influences, defined by either nonlinear multisensory responses or bimodal responses (Fig. 3*A*). Bimodal units are defined as showing significant responses to both the auditory and visual stimuli. The majority of the visual influences was characterized by audiovisual responses that significantly deviated from the sum of the responses to both unimodal stimuli (63 of 159 = 40% of sensory responsive STP units; *z* test between trial-based AV responses and a bootstrapped sample of possible A + V summations, $p < 0.01$, FDR correction; Fig. 3*A*, dark gray section). Nonlinear audiovisual interactions consisted of both superadditive (AV > A + V, 43% of nonlinear multisensory units) and subadditive (AV < A + V, 57% of nonlinear multisensory units) visual influences. Figure 4*A*, *B* shows some exemplary unit responses that were either superadditive (Fig. 4*A*) or subadditive (Fig. 4*B*). Other types of visual influences were less common in the STP, such as a few bimodal units with responses to both the auditory and the visual stimuli (10 of 67 = 15% of visually modulated STP units; Fig. 3*A*, blue sections).

In the STS, nonlinear multisensory interactions were apparent in 40% of the sensory responsive units (27 of 67; Fig. 3*B*, dark gray section). However, in contrast to the STP, a large portion of the nonlinear multisensory units were also bimodal (12 of 27 = 44% of nonlinear multisensory units; Fig. 3*B*, blue section in the small pie), and a larger proportion responded to the visual stimuli (4 of 27 = 15%; Fig. 3*B*, red section in the small pie; χ^2 test on the numbers of auditory, visual and bimodal STP and STS units showing nonlinear visual interactions: $\chi^2 = 146.79$, $p < 10^{-6}$).

A time-resolved analysis of the timing of the visual effect revealed that, in the STP, the onset latency of the nonlinear visual modulation occurred at 108 ± 13.3 ms (mean ± SEM) after sound onset and was similar for superadditive and subadditive influences. On average, the visual effect lasted for 400 ms (offset at 508 ± 21.9 ms after sound onset; Fig. 4*C*). In the STS, the nonlinear visual modulation started at 120 ± 20.4 ms and lasted for 368 ms (average offset at 488 ± 28.6 ms after sound onset). Despite a trend for later onset of visual modulation in the STS compared with the STP, the differences in visual effect latency and duration were not significant.

In summary, whereas the proportion of nonlinear visual modulation was similar between the STP and the STS, visually influenced units reflecting direct cross-modal convergence through bimodal responses were more prominent in the STS.

Sensitivity of visual influences to features in communication signals

Having observed a general prominence of audiovisual influences on STP units, we asked whether audiovisual interactions would be modulated by certain communication signal-related stimulus features. In the population of auditory responsive STP units, we found no significant impact of any of the stimulus factors on the amplitude of visual modulation (Fig. 5*A*,*B*; Table 1 rows 1–3, column 5 for sample sizes related to these analyses). However in the auditory responsive population of STS units, visual influences were modulated by the stimulus factors, despite the above observation that the auditory responses of these neurons did not reflect any auditory stimulus features (Table 1, rows 1–3, column 8 for sample sizes). The magnitude of the nonlinear audiovisual response in STS units was significantly modulated by voice identity and caller size (Friedman's RM-ANOVA on auditory responsive STS units; main effect "voice identity": $\chi^2_{(1)} = 9.93$, $p = 0.0016$; Fig. 5*C*; main effect "caller size": $\chi^2_{(1)} = 3.97$, $p = 0.046$; Figure 5*D*).

To assess whether visual influences on neuronal populations from both brain areas were differentially sensitive to stimulus features, we extended our nonparametric ANOVA to include a brain area factor and its interaction with the stimulus factor. This revealed a significant interaction between the voice identity and area factors (Scheirer-Ray-Hare test, $p = 0.0072$), confirming

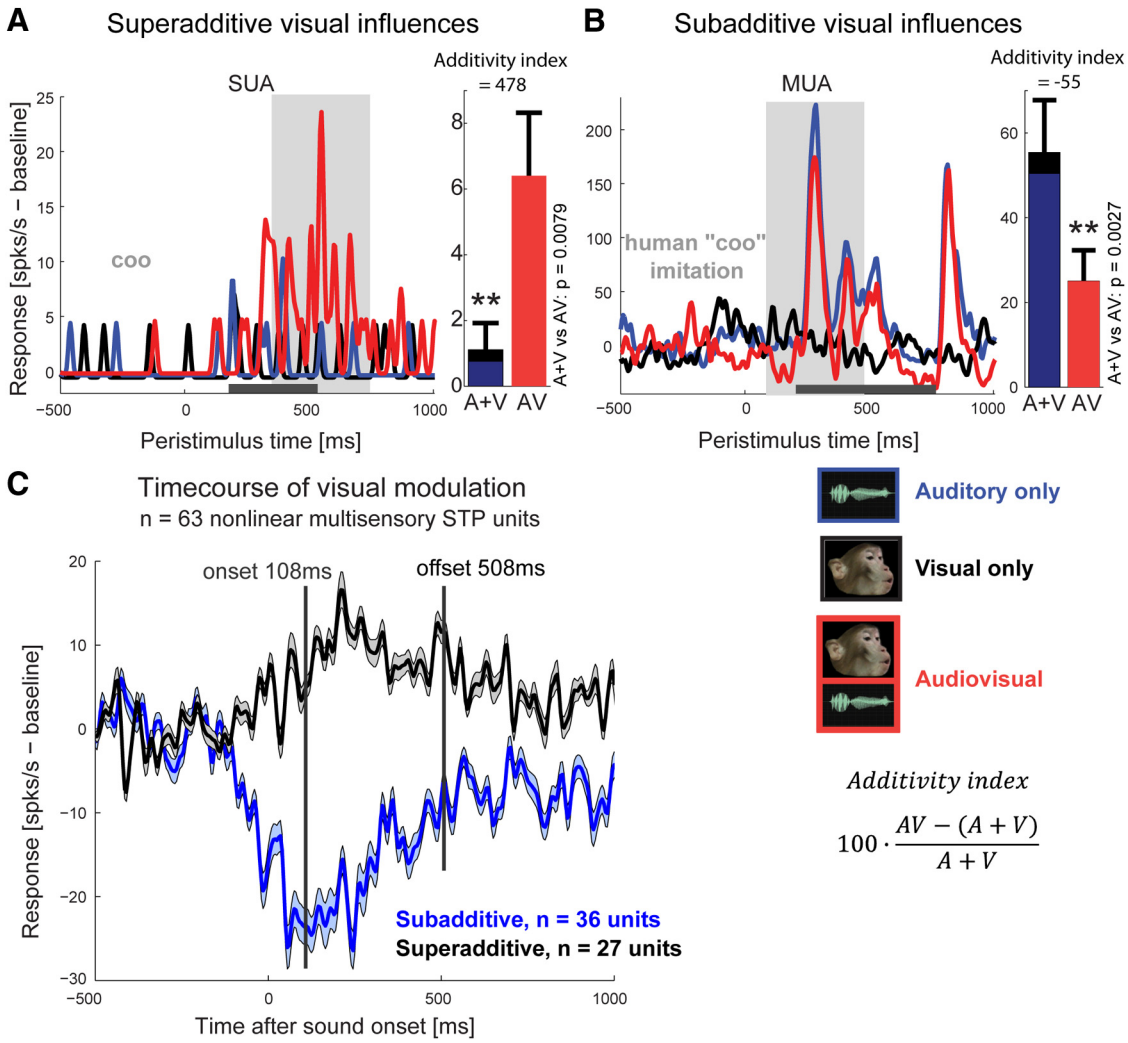

Figure 4. Visual influences on auditory responses in STP units. ***A***, Example response: single-unit activity (SUA) displaying superadditive visual modulation of the auditory response. The horizontal gray line indicates the duration of the auditory stimulus, and the light gray box represents the 400 ms response window in which the response amplitude was computed. Bar plots indicate the response amplitudes in the 400 ms response window. Data are mean ± SEM. *p* values refer to significantly nonlinear audiovisual interactions, defined by comparing the audiovisual response with all possible summations of auditory and visual responses: AV vs (A + V). $^{**}p < 0.01$ (*z* test). The additivity index values displayed quantify the audiovisual deviation from linear summation, in percentage of the sum of unimodal inputs. ***B***, Example response: multiunit activity (MUA) displaying subadditive visual modulation of the auditory response. ***C***, Time course of visual modulation (AV − (A + V)), separately for superadditive and subadditive units. Data are mean ± SEM.

that the effect of voice identity on the visual modulation depends on the brain area studied. A trend for an interaction between caller size and brain area failed to reach significance ($p = 0.081$).

These results suggest a double dissociation between STP and STS units on the cross-modal sensitivity to specific stimulus features, at least for the voice identity factor: STP units' cross-modal visual influences did not seem sensitive to different stimulus characteristics. In contrast, the cross-modal effects in the STS were modulated by stimulus-related features.

Cross-sensory sensitivity to audiovisual congruency

We tested the specificity of visual interactions using a set of incongruent audiovisual stimuli that paired the original auditory stimulus (voice) with a mismatched visual (face) context (AV-Congruency, see Materials and Methods for details; for sample sizes related to this analysis, see Table 1 row 5). We hypothesized that the visual influences in the STP and STS would differently depend on the congruency of voice-face pairs. One prediction is that the visual influences in the STS would, more so than the STP, depend upon the congruency of face-voice pairings, which was supported by the results.

On the whole, visual influences on auditory responsive STP units were as likely to occur in response to incongruent versions of the voice-face pairs as to congruent stimuli (distribution of modulated units not differing from uniformity, χ^2 test: $\chi^2 = 0.74$, $p = 0.69$; Fig. 6*A*). In contrast, audiovisual interactions in the STS showed a strong sensitivity to stimulus congruency, with a significant majority of units modulated by a congruent version of the audiovisual pairs, and reduced visual influences in response to incongruent pairs (χ^2 test against uniformity: $\chi^2 = 12.29$, $p = 0.0021$; Fig. 6*B*). The sensitivity to voice-face congruency significantly differed between both brain areas (χ^2 test comparing the distribution of STP and STS modulated units: $\chi^2 = 18.67$, $p = 8.83 \times 10^{-5}$). These differences in sensitivity to congruent/incongruent stimulus relationships did not affect the types of multisensory influences, which remained constant and

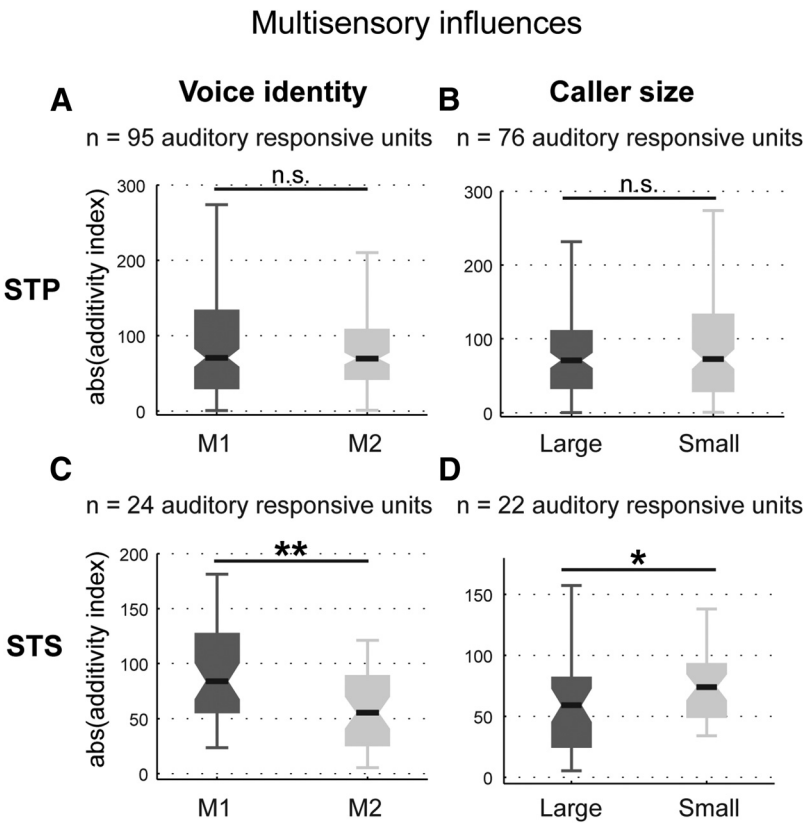

Figure 5. Impact of communication signal features on the magnitude of visual modulation. ***A***, ***C***, *Post hoc* comparisons (Wilcoxon rank sum test) of the effect of the voice identity factor (caller 1, M1; vs caller 2, M2) on the median amplitude of the audiovisual nonlinearity (absolute value of the additivity index; Fig. 4) across the population of STP units (***A***, $n = 95$ auditory responsive units) and the STS units (***C***, $n = 24$ auditory responsive units). ***B***, ***D***, Effect of the caller size (large vs small) on the amplitude of the visual modulation across the population of STP units (***B***, $n = 76$ auditory responsive units) and STS units (***D***, $n = 22$ auditory responsive units). The boxplots represent the median, upper, and lower quartiles of the population rectified additivity index values. Asterisks indicate significant effects of the investigated factor in a balanced 2-way nonparametric Friedman's test, with different auditory responsive units as repetitions. $^{**}p < 0.01$ (Wilcoxon rank sum test). $^{*}p < 0.05$ (Wilcoxon rank sum test). n.s., Not significant.

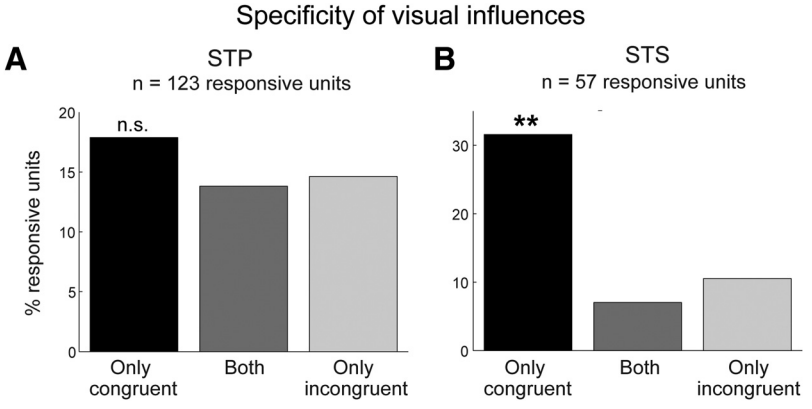

Figure 6. Effects of voice-face congruency on STP and STS unit responses. ***A***, Distribution of visually modulated STP units ($n = 123$ units responding to a least one of the three auditory vocalizations used in the "AVCongruency" subset of stimuli; Fig. 7A). ***B***, Visual influences in STS units ($n = 57$ auditory responsive units). $^{**}p < 0.01$ (resulting from a χ^2 test comparing the numbers of visually modulated units for each of the three categories to a uniform distribution). n.s., Not significant.

primarily subadditive in both STP and STS units (proportions of subadditive influences during congruency/incongruency: respectively, 62% and 66% in the STP; 80% and 77% in the STS).

To better pinpoint whether some of these effects were sensitive to different types of audiovisual congruency violations, we next analyzed responses to each of the mismatched controls separately. The audiovisual controls violated either the caller species congruency (controls 1 and 2), or temporal congruency (control 3) (Fig. 7*A*). Control stimulus pairing 1 combined a human mimicking a monkey "coo" vocalization with the video of a conspecific monkey mouthing a "coo" vocalization. Control 2 paired a monkey "coo" with the video of the human mimicking a monkey "coo" vocalization. Control 3 introduced a temporal asynchrony between two original congruent stimuli: a monkey "coo" vocalization was paired with the corresponding monkey face mouthing the coo call, but with a 340 ms auditory lag.

Noticeably, violating the caller species congruency (controls 1 and 2) was most disruptive on visual influences, in both the STP and the STS. Despite the STP units not being strongly influenced by the congruency of all of the stimuli (Fig. 6*A*), units were significantly sensitive to the congruency violation of a human face replacing the monkey face in this pairing (Control 2; χ^2 test on distribution of modulated units for control 1 compared with uniformity: $\chi^2 = 7.0$, $p = 0.030$; Fig. 7*C*; example response, Fig. 7*B*). STS units were also sensitive to the cross-species violation, but instead for Control 1 where a human face was replaced with a monkey face (χ^2 test: $\chi^2 = 9.50$, $p = 0.0087$; Fig. 7*D*). The temporal asynchrony, on the other hand, whereby a monkey voice onset was delayed by 300 ms relative to the monkey dynamic face onset, did not have an obvious impact (Fig. 7, Control 3).

These results show that visual influences in the STS were more specific to congruent voice-face stimuli than those in the STP, although the STP showed some congruency sensitivity to one of the control stimuli used.

Discussion

Our results revealed a number of double dissociations between the auditory and multisensory processing in STP and STS neurons: we observed considerable sensitivity in a "voice" region in the right STP to several auditory features in communication signals (such as call type, caller species, voice identity, and caller familiarity). This was not the case for auditory responses in the STS, an association area. Moreover, an unexpected finding was that the sensitivity to voice identity was primarily supported by a subpopulation of auditory STP units identified using a "voice localizer," whereas the sensitivity to call type was prominent throughout the STP. The results also reveal, to our

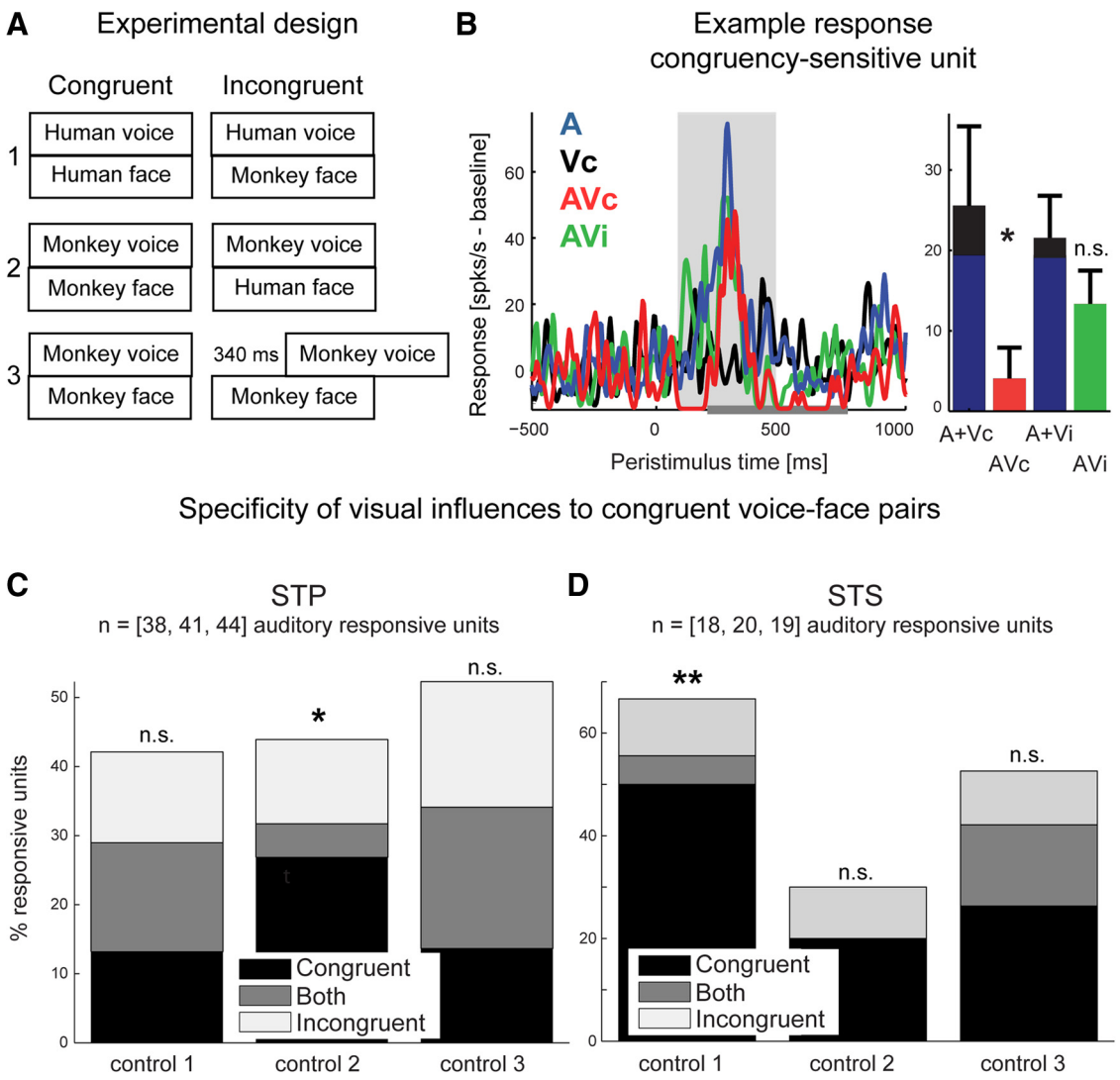

Figure 7. Specific violations of voice-face congruency and their impact on cross-sensory modulation. ***A***, Design of the AVCongruency stimulus subset, containing three congruency violations within primate voice/face pairs. ***B***, Example response of a unit in which visual influences were sensitive to audiovisual stimulus congruency: a congruent, but not an incongruent, visual stimulus significantly modulated the auditory response. The plot shows spiking activity in response to the auditory stimulus alone (A), the congruent visual stimulus alone (Vc), the congruent audiovisual (AVc), and the incongruent audiovisual (AVi) pairs. The horizontal gray line indicates the duration of the auditory stimulus, and the light gray box represents the 400 ms response window in which the response amplitude was computed. Bar plots indicate the response amplitudes in the 400 ms response window (mean ± SEM). The symbols refer to significantly nonlinear audiovisual interactions, defined by comparing the audiovisual response with all possible summations of auditory and visual responses: AVc vs (A + Vc) and AVi vs (A + Vi). *$p < 0.05$ (z test). n.s., Not significant. ***C***, Summary of the specificity of the visually modulated STP units for each of the 3 stimulus pairs tested. Bar plots indicate the percentage of auditory responsive units ($n = 38$ units responding to control 1, $n = 41$ for control 2, and $n = 44$ for control 3) that showed significant nonadditive audiovisual interactions in response to the congruent pair only (black bars), the incongruent pair only (light gray bar), or that integrated both the congruent and the incongruent stimuli (dark gray bar). ***D***, Summary of the specificity of the visually modulated STS units for each of the 3 stimulus pairs tested. Bar plots indicate the percentage of auditory responsive units ($n = 18$ for control 1, $n = 20$ for control 2, and $n = 19$ for control 3). **$p < 0.01$, *$p < 0.05$ (resulting from a χ^2 test comparing the numbers of visually modulated units for each category to a uniform distribution). n.s., Not significant.

knowledge, first evidence of cross-modal influences on spiking activity in a voice/face cluster. However, STS neurons had stronger responses to visual stimulation, a greater variety of cross-sensory response types, and stronger sensitivity to stimulus congruency than STP neurons. This study clarifies the position of STP and STS regions in unisensory and multisensory processing hierarchies and generates hypotheses for testing of face-sensitive clusters.

Auditory sensitivity of the anterior STP and voice-sensitive neurons

Temporal lobe voice-sensitive areas (also known as temporal voice areas [TVA]) are identified by their preference for voice versus non-voice stimulation. This contrast results in a number of fMRI-identified TVA clusters in humans or monkeys (Belin et al., 2002; Poremba et al., 2004; von Kriegstein and Giraud, 2004; Petkov et al., 2008). Among these TVA clusters, the one in the right anterior temporal lobe (superior-temporal gyrus/STS in humans: von Kriegstein et al., 2003; STP in monkeys: Petkov et al., 2008) in particular seems to be sensitive to voice identity. This was shown using either voice identity fMRI adaptation experiments in humans and monkeys (i.e., holding the call type constant and varying the callers: Belin and Zatorre, 2003; Petkov et al., 2008) or by selective attention to voice versus speech content in humans (von Kriegstein et al., 2003). Initial neuronal recordings from the anterior voice-sensitive cluster in monkeys have shown that "voice cells", classified analogously to "face cells" in the visual domain (Perrodin et al., 2011), can be sensitive to both voice identity and call type. A comparable dual sensitivity has also

been found in some of the auditory responsive neurons of the ventrolateral prefrontal cortex (Romanski et al., 2005). However, using a multifactorial design in this study and subdividing the STP neuronal population into conspecific VS and non-VS units, we observed that the strong call type sensitivity is prominent throughout the STP, including these two subpopulations. Yet the VS neurons, unlike the non-VS neurons, seemed to carry the voice identity factor sensitivity observed in the STP.

This observation is interesting, in light of recent human neuroimaging work (Belin and Zatorre, 2003; Formisano et al., 2008) and recordings from neuronal populations in patients (Mesgarani and Chang, 2012; Zion Golumbic et al., 2013), suggesting that speech and voice content are processed in largely overlapping temporal lobe regions. Such an overlap brings up the question of how neuronal representations to these different features in communication signals are segregated. Attentional selection has been highlighted as a key mechanism for this process (von Kriegstein et al., 2003; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013). However, our results in macaques using passive auditory stimulation (with the animals performing a simple visual task) show that there is already some evidence for functional segregation of at least voice identity sensitivity at the neuronal level in the anterior TVA cluster.

Visual influences along the multisensory processing hierarchy: STP versus STS

Our results also show evidence for robust visual modulation of auditory neuronal responses at a voice-sensitive area in the anterior STP. Approximately 40% of units in the STP demonstrated visual influences, seen mostly as nonlinear visual modulation of auditory responses. Other audiovisual studies have reported visual influences in ~12% of auditory responsive units either in monkey posterior core/belt auditory areas (Kayser et al., 2008) or several ferret auditory cortical fields (Bizley et al., 2007). The visual influences that we observed in the STP are similar to the numbers reported by Ghazanfar et al. (2005), who used dynamic voice and face stimuli to identify 40% and 35% of visually modulated units in the belt and core fields, respectively. Comparably, in visual area TE of the inferotemporal cortex, ~24% of visually responsive units are modulated by cross-sensory input (Kaposvari et al., 2011). Potential sources of modulatory visual "face" input into the auditory STP include corticocortical projections from visual areas (Bizley et al., 2007; Blank et al., 2011) and feedback projections from higher association areas, such as the frontal lobe, including the voice/face sensitive ventrolateral PFC (Romanski et al., 1999a, b), and the STS (Pandya et al., 1969; Kaas and Hackett, 1998).

In the STS, the proportion of nonlinear multisensory influences was comparable to that in the STP, with cross-modal effects modulating ~40% of sensory responsive units. However, we found that audiovisual interactions in the STS were more likely to be mediated by direct cross-modal convergence than in the STP. This observation is in line with previous electrophysiological studies that have reported multimodal neuronal proportions of 36–38% (Benevento et al., 1977; Bruce et al., 1981) and 53% (Dahl et al., 2009) in the anterior STS, and is consistent with studies highlighting the STS as an association cortical region that is a prominent target for both auditory and visual afferents (Seltzer and Pandya, 1994; Beauchamp et al., 2004a).

Behaviorally, intermediate noise levels yield the strongest audiovisual benefits (Ross et al., 2007). Although our study did not manipulate sensory noise, it would be interesting for future studies to evaluate the impact of noise on multisensory processes (Kayser et al., 2007). In this regard, we hypothesize that cross-modal influences in the STS and STP would be similarly affected.

Beyond the proportions and types of audiovisual interactions, we also tested the specificity of visual influences to stimulus congruency, using a set of incongruent voice-face pairings. Our data show that visual influences on STP units were relatively generic to different pairings of dynamic primate faces and voices and were not strongly disrupted by mismatched audiovisual stimulus pairs. The exception was some sensitivity to a species incongruency affecting a conspecific caller, suggesting that STP units can in some cases tune out mismatching visual information during the processing of conspecific sounds. Ghazanfar et al. (2005) noted sensitivity to a congruency violation pairing a voice with an artificial visual mouth movement in caudal auditory cortex. It remains possible that, with such or other stimulus conditions, the STP may have been more strongly sensitive to violations of audiovisual congruency. However, even in this case, our STP and STS results would predict a relative difference between the form and/or preponderance of audiovisual congruency sensitivity between auditory cortex and the STS. Our STS results are consistent with Dahl et al. (2010), who also reported congruency-sensitive auditory influences on visual responses in the monkey lower-bank STS.

The role of the STP in the unisensory and multisensory hierarchies

Our results comparing auditory and audiovisual analysis of communication signals between the neurons in the STP and STS are relevant for models of multisensory processing and would suggest some revision or updating of current notions. The voice-sensitive STP characterized here represents a hierarchically higher-level auditory association cortex, positioned at the later stage of a ventral auditory cortical processing stream (Rauschecker et al., 1997; Romanski et al., 1999b; Petkov et al., 2008; Kikuchi et al., 2010). However, such regions do not yet feature in models of multisensory processing. For instance, direct interactions between voice and face recognition units are proposed in models of person perception (Ellis et al., 1997; Campanella and Belin, 2007) and are supported by recent tractography data in humans (Blank et al., 2011). However, the specificity of cross-sensory influences is not clear, leaving uncertain whether (1) association areas are the primary sites for multisensory integration, (2) most cortical and many subcortical regions are multisensory in presumably comparable ways, and/or (3) there is clearly a multisensory processing hierarchy that relates in certain ways to anatomical processing hierarchies. Also, these possibilities are not mutually exclusive (for review, see, e.g., Ghazanfar and Schroeder, 2006; Campanella and Belin, 2007; Stein and Stanford, 2008).

Our results certainly contribute to the set of regions in the auditory cortical processing hierarchy that are influenced by visual input and motivate the hypothesis for testing in the visual modality that at least the anterior face-sensitive IT cluster has prominent cross-sensory modulation. However, the results also suggest that such auditory/visual regions, perhaps because they are primarily engaged in sensory analysis in the dominant modality, have less specificity regarding cross-sensory influences. This might prevent disruption of vocal or facial analysis during incongruent cross-sensory situations (e.g., looking at an individual that is not the one vocalizing). On the other hand, association cortical areas, such as the STS, appear to lose the fidelity of unisensory processes but seem to be well involved in resolving cross-sensory conflict. Overall, our results are consistent with reversed gra-

dients of functional specificity in unisensory processing and multisensory influences, along their respective hierarchical processing pathways.

References

Avillac M, Ben Hamed S, Duhamel JR (2007) Multisensory integration in the ventral intraparietal area of the macaque monkey. J Neurosci 27:1922–1932. CrossRef Medline

Barraclough NE, Xiao D, Baker CI, Oram MW, Perrett DI (2005) Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J Cogn Neurosci 17:377–391. CrossRef Medline

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004a) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190–1192. CrossRef Medline

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

Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14:2105–2109. CrossRef Medline

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309–312. CrossRef Medline

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13:17–26. CrossRef Medline

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

Benjamini YHY (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.

Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17:2172–2189. CrossRef Medline

Blank H, Anwander A, von Kriegstein K (2011) Direct structural connections between voice- and face-recognition areas. J Neurosci 31:12906–12915. CrossRef Medline

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

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of cross-modal binding in the human heteromodal cortex. Curr Biol 10:649–657. CrossRef Medline

Campanella S, Belin P (2007) Integrating face and voice in person perception. Trends Cogn Sci 11:535–543. CrossRef Medline

Chen A, Deangelis GC, Angelaki DE (2013) Functional specializations of the ventral intraparietal area for multisensory heading discrimination. J Neurosci 33:3567–3581. CrossRef Medline

Cipolloni PB, Pandya DN (1989) Connectional analysis of the ipsilateral and contralateral afferent neurons of the superior temporal region in the rhesus monkey. J Comp Neurol 281:567–585. CrossRef Medline

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924–11932. CrossRef Medline

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. CrossRef Medline

Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. Br J Psychol 88:143–156. CrossRef Medline

Fitch WT, Fritz JB (2006) Rhesus macaques spontaneously perceive formants in conspecific vocalizations. J Acoust Soc Am 120:2132–2141. CrossRef Medline

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what?" Brain-based decoding of human voice and speech. Science 322:970–973. CrossRef Medline

Fuster JM, Bodner M, Kroger JK (2000) Cross-modal and cross-temporal association in neurons of frontal cortex. Nature 405:347–351. CrossRef Medline

Galaburda AM, Pandya DN (1983) The intrinsic architectonic and connectional organization of the superior temporal region of the rhesus monkey. J Comp Neurol 221:169–184. CrossRef Medline

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278–285. CrossRef Medline

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

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

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

Kaas JH, Hackett TA (1998) Subdivisions of auditory cortex and levels of processing in primates. Audiol Neurootol 3:73–85. CrossRef Medline

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci U S A 97:11793–11799. CrossRef Medline

Kaposvari P, Csibri P, Csete G, Tompa T, Sary G (2011) Auditory modulation of the inferior temporal cortex neurons in rhesus monkey. Physiol Res 60 [Suppl 1]:S93–S99.

Kayser C, Logothetis NK (2007) Do early sensory cortices integrate cross-modal information? Brain Struct Funct 212:121–132. CrossRef Medline

Kayser C, Petkov CI, Augath M, Logothetis NK (2007) Functional imaging reveals visual modulation of specific fields in auditory cortex. J Neurosci 27:1824–1835. CrossRef Medline

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

Kikuchi Y, Horwitz B, Mishkin M (2010) Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. J Neurosci 30:13021–13030. CrossRef Medline

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279–292. CrossRef Medline

Lakatos P, O'Connell MN, Barczak A, Mills A, Javitt DC, Schroeder CE (2009) The leading sense: supramodal control of neurophysiological context by attention. Neuron 64:419–430. CrossRef Medline

Linden JF, Grunewald A, Andersen RA (1999) Responses to auditory stimuli in macaque lateral intraparietal area: II. Behavioral modulation. J Neurophysiol 82:343–358. Medline

Maunsell JH, Newsome WT (1987) Visual processing in monkey extrastriate cortex. Annu Rev Neurosci 10:363–401. CrossRef Medline

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236. CrossRef Medline

Mitchell JF, Sundberg KA, Reynolds JH (2007) Differential attention-dependent response modulation across cell classes in macaque visual area V4. Neuron 55:131–141. CrossRef Medline

Müller VI, Habel U, Derntl B, Schneider F, Zilles K, Turetsky BI, Eickhoff SB (2011) Incongruence effects in cross-modal emotional integration. Neuroimage 54:2257–2266. CrossRef Medline

Pandya DN, Hallett M, Kmukherjee SK (1969) Intra- and interhemispheric connections of the neocortical auditory system in the rhesus monkey. Brain Res 14:49–65. CrossRef Medline

Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. Curr Biol 21:1408–1415. CrossRef Medline

Petkov CI, Kayser C, Augath M, Logothetis NK (2006) Functional imaging reveals numerous fields in the monkey auditory cortex. PLoS Biol 4:e215. CrossRef Medline

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK (2008) A voice region in the monkey brain. Nat Neurosci 11:367–374. CrossRef Medline

Poremba A, Malloy M, Saunders RC, Carson RE, Herscovitch P, Mishkin M (2004) Species-specific calls evoke asymmetric activity in the monkey's temporal poles. Nature 427:448–451. CrossRef Medline

Rauschecker JP, Tian B, Pons T, Mishkin M (1997) Serial and parallel processing in rhesus monkey auditory cortex. J Comp Neurol 382:89–103. CrossRef Medline

Romanski LM (2007) Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. Cereb Cortex 17 [Suppl 1]:i61–i69.

Romanski LM, Giguere M, Bates JF, Goldman-Rakic PS (1997) Topographic organization of medial pulvinar connections with the prefrontal

cortex in the rhesus monkey. J Comp Neurol 379:313–332. CrossRef Medline

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

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999b) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2:1131–1136. CrossRef Medline

Romanski LM, Averbeck BB, Diltz M (2005) Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J Neurophysiol 93:734–747. CrossRef Medline

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147–1153. CrossRef Medline

Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. Curr Opin Neurobiol 15:454–458. CrossRef Medline

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

Schroeder CE, Smiley J, Fu KG, McGinnis T, O'Connell MN, Hackett TA (2003) Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. Int J Psychophysiol 50: 5–17. CrossRef Medline

Seltzer B, Pandya DN (1994) Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol 343:445–463. CrossRef Medline

Sergent J, Ohta S, MacDonald B (1992) Functional neuroanatomy of face and object processing: a positron emission tomography study. Brain 115: 15–36. CrossRef Medline

Sokal RR, Rohlf FJ (1995) Biometry. New York: W.H. Freeman.

Stanford TR, Quessy S, Stein BE (2005) Evaluating the operations underlying multisensory integration in the cat superior colliculus. J Neurosci 25:6499–6508. CrossRef Medline

Stein BE, Stanford TR (2008) Multisensory integration: current issues from the perspective of the single neuron. Nat Rev Neurosci 9:255–266. CrossRef Medline

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

Tsao DY, Moeller S, Freiwald WA (2008) Comparing face patch systems in macaques and humans. Proc Natl Acad Sci U S A 105:19514–19519. CrossRef Medline

von Kriegstein K, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22:948–955. CrossRef Medline

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17:48–55. CrossRef Medline

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17: 367–376. CrossRef Medline

Werner S, Noppeney U (2010a) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci 30:2662–2675. CrossRef Medline

Werner S, Noppeney U (2010b) Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. Cereb Cortex 20:1829–1842. CrossRef Medline

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." Neuron 77:980–991. CrossRef