Behavioral/Systems/Cognitive

Dual Neural Routing of Visual Facilitation in Speech Processing

Luc H. Arnal,[1] Benjamin Morillon,[1] Christian A. Kell,[1,2] and Anne-Lise Giraud[1,3]

[1]Inserm U960, Laboratoire de Neurosciences Cognitives, Département d'Etudes Cognitives, Ecole Normale Supérieure, F-75005 Paris, France, [2]Brain Imaging Center, Department of Neurology, Goethe-University, D-60590 Frankfurt am Main, Germany, and [3]Centre de Neuroimagerie de Recherche, Hôpital de la Pitié-Salpêtrière, F-75013 Paris, France

Viewing our interlocutor facilitates speech perception, unlike for instance when we telephone. Several neural routes and mechanisms could account for this phenomenon. Using magnetoencephalography, we show that when seeing the interlocutor, latencies of auditory responses (M100) are the shorter the more predictable speech is from visual input, whether the auditory signal was congruent or not. Incongruence of auditory and visual input affected auditory responses ~20 ms after latency shortening was detected, indicating that initial content-dependent auditory facilitation by vision is followed by a feedback signal that reflects the error between expected and received auditory input (prediction error). We then used functional magnetic resonance imaging and confirmed that distinct routes of visual information to auditory processing underlie these two functional mechanisms. Functional connectivity between visual motion and auditory areas depended on the degree of visual predictability, whereas connectivity between the superior temporal sulcus and both auditory and visual motion areas was driven by audiovisual (AV) incongruence. These results establish two distinct mechanisms by which the brain uses potentially predictive visual information to improve auditory perception. A fast direct corticocortical pathway conveys visual motion parameters to auditory cortex, and a slower and indirect feedback pathway signals the error between visual prediction and auditory input.

Introduction

Psychological and neurophysiological data show that visual speech improves auditory speech recognition and processing (Sumby and Polack, 1954; von Kriegstein et al., 2008). In most ecological settings, auditory input lags visual input, i.e., mouth movements and speech-associated gestures, by ~150 ms (Chandrasekaran et al., 2009). This lag allows the brain to anticipate auditory signals, resulting in speeding up early cortical auditory responses (Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007). Physiological and anatomical studies in humans and monkeys indicate several routes by which visual input might influence auditory information processing (Kayser et al., 2007; Driver and Noesselt, 2008; Schroeder et al., 2008). The most common view is that the visual system indirectly reaches the auditory system via a feedback from "supramodal" areas (Ghazanfar et al., 2005) in which auditory and visual inputs converge. As it responds to both auditory and visual inputs, and more specifically to audiovisual speech combinations (Calvert and Campbell, 2003; Miller and D'Esposito, 2005), the middle part of the superior temporal sulcus (STS) is the most likely feedback provider to auditory cortex in a speech context. (Calvert and Campbell, 2003; Beauchamp et al., 2004a,b; Hertrich et al., 2007; Ghazanfar et al., 2008; Kayser and Logothetis, 2009). Current views on multisensory integration (Driver and Noesselt, 2008; Schroeder et al., 2008) however suggest that there might be at least one additional cortical pathway by which visual input could influence auditory processing (Fig. 1*A*), a direct corticocortical input to auditory cortex from visual cortex (Falchier et al., 2002; Rockland and Ojima, 2003; Cappe and Barone, 2005). Here, we assume that we can distinguish the contribution of direct corticocortical versus feedback pathways (Fig. 1, pathways 1 and 2) by exploring the degree of specificity of auditory processing facilitation by visual input.

We investigated the two possible routing of a visual signal on auditory speech processing by recording early visual (M170V) and auditory (M100A) evoked responses to natural (congruent) and to nonmatching (incongruent) audiovisual syllables using magnetoencephalography (MEG). First, we measured viseme specificity (dependence on lip movements associated with a syllable) of M100A facilitation; second, we examined the dependence of this effect upon audiovisual congruence (neural mismatch). We hypothesized that pathways 1 and 2 should both induce a facilitation that depends on those specific mouth movements that are being used for pronunciation (i.e., viseme-dependent facilitation), but should yield distinct incongruence effects. Direct corticocortical projections arising from visual cortex convey visual information, e.g., visual motion, but no detailed phonological information. pathway 1 is hence expected to induce either no neural mismatch or a mismatch effect that does not exhibit viseme dependency. Conversely, as pathway 2 originates in a

Received July 3, 2009; revised Sept. 8, 2009; accepted Sept. 21, 2009.

A.-L.G. is funded by Centre Nationale de la Recherche Scientifique. We thank the staff of the Centre de Neuroimagerie de Recherche and the Magnetoencephalography Center (Hôpital de la Pitié-Salpêtrière, Paris), in particular Antoine Ducorps and Denis Schwarz. We are also grateful to Virginie van Wassenhove, Catherine Tallon-Baudry, Goulven Josse, Daniel Abrams, Andreas Kleinschmidt, Pascal Barone, and Andrej Kral for their valuable scientific input.

Correspondence should be addressed to Luc H. Arnal, Laboratoire de Neurosciences Cognitives, Département d'Etudes Cognitives, Ecole Normale Supérieure, 29 rue d'Ulm, F-75005 Paris, France. E-mail: luc.arnal@ens.fr.

DOI:10.1523/JNEUROSCI.3194-09.2009

Copyright © 2009 Society for Neuroscience 0270-6474/09/2913445-09$15.00/0

region which, among other functions, underpins audiovisual integration of phonological inputs (Beauchamp et al., 2004a; Barraclough et al., 2005; Hickok and Poeppel, 2007), it is expected to induce a viseme-dependent neural mismatch.

We further distinguished between pathways 1 and 2 using a complementary approach. We collected functional MRI data using similar stimuli as in the MEG experiment, with a slightly different paradigm (1) to assess viseme dependency of hemodynamic responses in motion visual cortex and middle STS and (2) to examine the neuroanatomical plausibility of each route using functional connectivity.

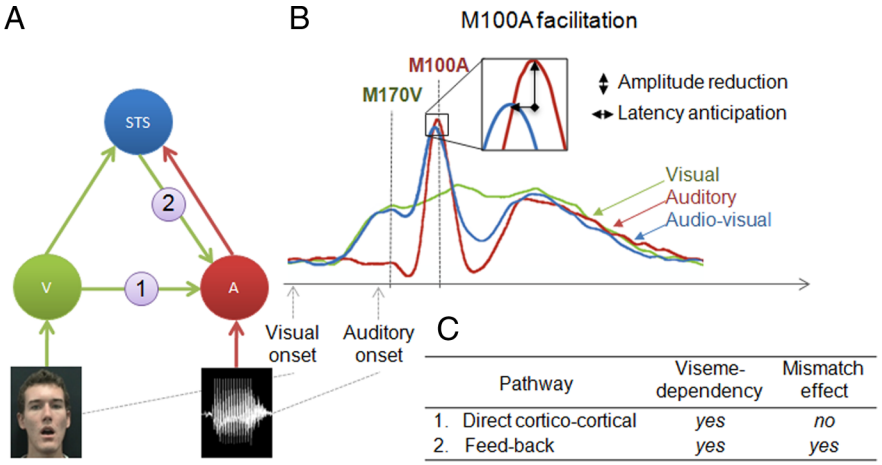

Figure 1. Neuroanatomical model of auditory facilitation by concurrent visual input and related predictions. ***A***, Two anatomical pathways are proposed for the routing of visual information (green arrows) to auditory areas (A) (red arrows representing routing of auditory information): pathway (1) is a direct corticocortical pathway from visual cortices (V) and pathway (2) is a feedback pathway from multisensory STS. ***B***, Time course of evoked components for auditory (red), visual (green), and AV (blue) stimuli. Neuronal facilitation is assessed by measuring amplitude reduction and latency anticipation of the M100A peak in the AV–V versus A conditions. ***C***, Predictions on the origin of M100A facilitation as a function of (1) viseme dependency and (2) mismatch when audio and visual syllables are not congruent.

Materials and Methods

Ethics statement

All subjects gave written informed consent to take part in these studies that were approved of by the local ethics committee (Comité Consultatifs de Protection des Personnes se prêtant à des Recherches Biomédicales Paris-Cochin, # RBM 01-04).

Participants

Thirty-four French native subjects without known neurological or sensory disorder participated in two behavioral pilot experiments and two neuroimaging experiments. Fifteen participants (eight females; age range: 20–53 years) took part in the behavioral experiments. Fifteen other subjects (right-handed, 10 females; age range: 20–28 years) participated in the MEG experiment, and 16 in the fMRI experiment (right-handed, 7 females; age range: 21–26 years). Twelve of them participated in both neuroimaging studies.

Stimuli and behavioral studies

Audiovisual, audio-only, and visual-only stimuli were extracted from digital videos of a male speaker pronouncing consonant/vowel (CV) syllables (C /a/ syllables) (supplemental Figs. 1 and 2, available at www.jneurosci.org as supplemental material). Videos were edited in Adobe Premier Pro into a 720/576 pixel movie with a digitization rate of 60 images/s (1 frame = 16.7 ms). Stereo soundtracks were digitized in Adobe Audition at 44.1 kHz with 16-bit resolution. Stimulus presentation was coordinated with Presentation software (Neurobehavioral Systems).

The two behavioral experiments, referred to as "predictability" and "incongruence" experiments, served to establish a gradient of visual predictability and perceived incongruence, respectively, which we subsequently used in the MEG study. In the predictability pilot experiment participants performed a five-alternative forced-choice experiment in which they were asked to repeat syllables randomly presented in the visual modality only. Mean recognition rates were used as an index of visual predictability (van Wassenhove et al., 2005). The pilot incongruence study was conducted in the same subjects, who were then asked to quantify perceived incongruence of AV congruent (AVc) or incongruent (AVi) pairs on a four-step (0–3) subjective scale (Bernstein et al., 2008a). At the end of each trial, subjects were asked to verbally report which syllable had been perceived. Incongruent combinations yielding McGurk fused or combined illusory percepts (McGurk and MacDonald, 1976; van Wassenhove et al., 2005) were excluded. We used identical selection criteria for the stimuli used in the fMRI study, and individual recognition scores were assessed online during fMRI recordings.

We selected CV syllables with different places of articulation to enhance a gradient in their visual predictive power. The visual and auditory tracks of each syllable were combined to yield the 4 following conditions: auditory (A), visual (V), AVc, and AVi. In the A condition the sound track was presented with a video of a still face, and in the V condition the speaking face was presented in silence. Incongruent pairs of syllables were created by dubbing the visual track randomly with a nonmatching auditory track. Stimulus mean duration was 5.3 s for the MEG experiment, and 11 s for the fMRI study including varying interstimulus intervals. In both MEG and fMRI designs trials began with a fixation cross on a black screen located at the center of mouth to prevent gaze shift when the face appeared. Videos lasted 2 s, with auditory onset (AO) 1 s after the video began. Details on stimuli and experimental design are provided additionally in supplemental Figures 1 and 2, available at www.jneurosci.org as supplemental material.

MEG study

Experimental procedures. Participants sat at a distance of 1 m from the monitor, the movie subtending 10.5° (horizontal) and 8.5° (vertical) visual angles. Videos were projected on a white screen with a Mitsubishi X120 videoprojector, in a dimly lit room. Sounds were presented at a comfortable hearing level individually set before the experiment (mean = 30 dB sensation level) via Promold earphones (International Aquatic Trade).

During MEG recordings, subjects performed an unrelated target detection task. They were presented with six possible syllables (/ga/, /ta/, /la/, /pa/, /ja/, /fa/) and were asked to report by keypress whether the presented syllable was a /fa/ (not included in the visual prediction gradient) regardless of the input condition (A, V, or AV) or perceived audiovisual congruence (AVc, AVi). To prevent eye movements, subjects were asked to fixate the cross and blink only after giving their motor response (after the video). Thus, only fewer than 5% of trials were contaminated by eye movement artifacts and were excluded. Stimuli were presented in a pseudorandomized order, with 54 repetitions of each.

Recordings. Continuous cerebral activity was recorded with a whole-head MEG system (Omega 151, CTF Systems), with 151 axial gradiometers over the scalp, at a sampling rate of 1250 Hz and low-pass filtered online at 300 Hz. Three small coils were attached to reference landmarks on the participant's face: at the left and right preauricular points and at the nasion. At the beginning of each block, the head position relative to the coordinate system of the MEG helmet was calculated from the position of those coils to register possible head movements during the experimental session. Eye movements and blinks were monitored with four ocular electrodes (Viasys Healthcare). They were automatically marked when they exceeded the mean by 2 SDs. This technique however does not detect microsaccades. One supplementary electrode was used to monitor cardiac activity.

Data processing. Data preprocessing, analysis, and visualization were performed using in-house software (http://cogimage.dsi.cnrs.fr/logiciels/).

We rejected off-line trials that were contaminated by eye or head movement, muscle contractions, or electromagnetic artifacts. Artifacts related to cardiac activity were eliminated by using a heartbeat trace matched filter. Two subjects were excluded from the data analysis due to poor recording quality. High-pass (0.15 Hz) and low-pass (30 Hz) filters were applied to the continuous recorded signal. Event-related fields (ERFs) were obtained by averaging epochs on a 3 s interval surrounding AO (2 s before and 1 s after) and baseline corrected (still face) on the interval (−900; −300 ms) relative to AO, to ensure that the correction occurred before lip movement onset.

Behavioral and MEG data analysis. Visual predictability was assessed by measuring recognition rates of each syllable when presented in the visual modality only and tested by repeated-measures ANOVA (factor: viseme; five levels: /ga/, /ta/, /la/, /pa/, /ja/). For each audiovisual combination we also tested for potential interactions between predictability and perceived congruence (factor: viseme; five levels: /ga/, /ta/, /la/, /pa/, /ja/, and congruence; two levels: AVi vs AVc pairs).

ERF analysis focused on visual facilitation effects on M100A. Thus, for each subject, we pooled the three left temporal channels in which auditory M100 amplitude was maximal. This provided an objective criterion to focus the analysis on M100 source, as we probed visual facilitation on this component. Peak latency and amplitude were extracted for M100A (from 50 to 130 ms relative to AO) in A, AVc, and AVi conditions. To compare these three conditions, time series corresponding to the V condition were subtracted from AV time series. For each syllable, we calculated latency and amplitude differences between A and AV-V M100A peaks. Viseme specificity of latency anticipation and amplitude change of M100A was assessed by computing two-way ANOVAs with repeated-measurement factors as follows: viseme (five levels: /ga/, /ta/, /la/, /pa/, /ja/) and congruence (two levels: AVc, AVi). If there was a significant ($p < 0.05$) viseme dependency of latency and/or amplitude on M100A facilitation, we further tested whether this effect was due to visual predictability by correlating viseme-dependent facilitation with the related behavioral recognition rate (Pearson's regression analysis). Significance was assumed at $p < 0.05$.

To identify the level at which visual and auditory inputs are compared, we searched for a main effect of incongruence at the topographical level by comparing grand-averaged AVc (average of all AVc syllables) with AVi (average of all AVi combinations) conditions. Paired *t* tests (two tailed, AVi vs AVc) were performed on each MEG sensor to identify the spatiotemporal windows at which incongruence yielded significant effects. Sensors showing the highest *T* values (minimum $T > 3$) during a period exceeding 20 ms were computed together. The mean magnitude of the neuromagnetic activity induced by each condition for each window was then extracted and repeated-measures ANOVAs were performed with the following factors: congruence (two levels: AVc, AVi) and viseme (five levels: /ga/, /ta/, /la/, /pa/, /ja/). We checked that these incongruence effects were related to prediction error (discrepancy between expected and incoming auditory input, i.e., perceive incongruence) using a regression analysis in which amplitude differences between AVi and AVc signals were compared with the perceived differences between AVi and AVc conditions. For each of the time windows we obtained, interaction effects between predictability and incongruence were tested by entering amplitude of the evoked response for each syllable as dependent variable into a univariate general linear model with visual predictability as covariate and incongruence as random variable.

MEG source estimations. Cortical current density time series were estimated individually at each of the 10,000 sources normally distributed over the cortical surface, by using the linear minimum-norm estimator available in the BrainVisa software (http://brainvisa.info/). For each condition, source estimations were individually projected on the MNI template, by matching spatial positions of three coils with corresponding landmarks on the template. Current source time series were then normalized individually by computing *Z* scores with respect to their baseline (−600 ms; −300 ms) for each condition, before subsequently averaging each condition across subjects. *Z* score maps were then thresholded above $Z = 10$ for subsequent analysis and interpretation. We considered cortical activations significant when they exceeded this threshold (deviating from baseline with $p < 0.01$, uncorrected). Time course of mean ($n = 15$) *Z* scores maps at the group level are shown in the supplemental videos, available at www.jneurosci.org as supplemental material (decelerated over the −200 to 500 ms time period). Visualization of spatiotemporal cortical activity was optimized by setting the maximum threshold at 90% of the maximum amplitude of each of the three main components [M170V (−200; 50 ms), M100A (50; 180 ms), and M400 (180; 500 ms), delays are given relative to AO].

fMRI study

Stimuli. Methods of stimulus acquisition, edition, and presentation were the same as those used in the MEG experiment. Specific stimulus combinations are provided in supplemental Figure 2*B*, available at www.jneurosci.org as supplemental material. During MRI acquisition, subjects lay comfortably supine and wore headphones for noise protection and delivery of acoustic stimuli. Visual stimuli were presented on a screen and viewed through a mirror.

Experimental procedures. During fMRI recordings, subjects were presented videos of a speaker's face pronouncing six syllables (supplemental Fig. 2, available at www.jneurosci.org as supplemental material). The soundtrack was either turned off (V) or on, in which case it could either be congruent to the video (AVc) or incongruent (AVi). A video of the speaker's still face with soundtrack (A), or without (still-face video; null), served as control conditions. The task was to determine whether the seen and/or heard stimulus corresponded to a written syllable presented subsequently. Individual syllable predictability scores (recognition rate of each syllable when presented in the visual modality only) were assessed online during fMRI acquisition. These scores were computed for each syllable and normalized within subjects to be related to the fMRI time series.

fMRI measurements and data processing. Functional images were collected with a Siemens Allegra 3.0 T scanner by acquisition of 980 volumes (four sessions of 245 volumes) of a gradient echo planar imaging (EPI) sequence. Images were parameterized as follows: matrix size, 64×64; voxel size, $3 \times 3 \times 3$ mm; echo time, 30 ms; repetition time, 2400 ms. A functional image volume comprised 39 contiguous slices which ensured that the whole brain was within the field of view. Additionally, we acquired a T1 sequence to exclude subjects with possible structural brain anomalies.

Imaging data were processed and analyzed with SPM5 (Wellcome Department of Imaging Neuroscience, University College of London, UK, http://www.fil.ion.ucl.ac.uk/spm/software/spm5/). EPI images were spatially preprocessed (realignment, normalization; smoothed with an 8 mm full width at half-maximum isotropic Gaussian kernel) using standard parameters of SPM5. The data were analyzed in the framework of the general linear model. Auditory, visual, audiovisual (AVc and AVi), and null (still-face video) conditions were modeled independently for each syllable as boxcar functions of 2 s and convolved with a classical hemodynamic response function. We analyzed the contrast: speaking faces (V) for all syllables > still faces (null) to determine which regions responded specifically to orofacial movements (speech motion localizer). To probe brain regions that responded to visual predictability, V, AVc, and AVi events conditions were weighted with normalized individual predictability scores for each viseme. Contrasts were calculated at the first level and entered in a second level analysis, with subjects treated as random variable (one-sample *t* test, $p < 0.001$ uncorrected).

Psychophysiological interaction analyses. Functional connectivity was assessed using a procedure implemented in SPM5 for the study of psychophysiological interactions (PPIs) (Friston et al., 1997). This technique detects changes in the coupling between two brain regions, depending on a factor, here visual syllable predictability. The PPI were computed across the three regions of interest (ROIs): left motion-sensitive cortex, auditory cortex including Heschl's gyrus, and middle STS. The individual time series were extracted from the peak voxel that responded in the appropriate functional contrast in motion-sensitive cortex and in auditory cortex within a radius of 10 mm from the group maxima in the anatomical borders of the probe region. Due to too-widespread individual responses, the ROI in the STS was used only as a target. The regressor for the PPI was computed individually as the product of the extracted time series and a vector coding for the parametric

effect of visual predictability (individual recognition score for each syllable). In addition to the PPI regressor and the region-of-interest time series, our model included the main effects of parametric predictability and effects of no interest (movement regressors). We probed an effect of visual predictability on functional connectivity by testing (t test) for a correlation between activity in each region and PPI regressors. We assessed functional connectivity pairwise across left motion-sensitive cortex, auditory cortex, and STS and report maxima appearing within 2-cm-radius spheres surrounding these regions of interest with a threshold of $p \leq 0.01$, uncorrected.

Results

Gradient of visual syllable predictability

In the MEG experiment we used five syllables (/ga/, /ta/, /la/, /pa/, /ja/) that were randomly presented to the subjects in auditory (sound plus still face), visual (silent video), and audiovisual conditions (supplemental Fig. 1, available at www.jneurosci.org as supplemental material). When presented in the visual modality only, in a pilot behavioral experiment, these five syllables yielded recognition rates ranging from 38 to 99.3% (repeated-measures ANOVA: $F_{(4,56)} = 83.604$, $p = 0.000$) (Fig. 2). Such a broad gradient of visual predictability was a requirement to use these stimuli to assess viseme dependency in neural effects.

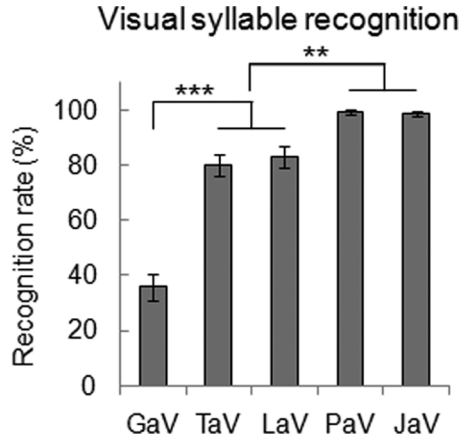

Figure 2. Predictability of syllables /ga/, /ta/, /la/, /pa/, and /ja/, presented visually (V) and ordered by increasing predictability. The predictive power of five visual syllables was assessed by measuring recognition rates in 15 subjects. Error bars indicate SEM. $^{**}p < 0.01$, $^{***}p < 0.001$.

Viseme dependency of M100 facilitation

Early cortical evoked responses (EEG P2) typically have shorter latencies in audiovisual relative to auditory condition (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007). We confirmed decreased latency of event-related MEG responses, detectable as early as 100 ms poststimulus (M100A) when facial movements accompany corresponding speech sounds ($F_{(1,14)} = 22.7$, $p = 0.000$) (Fig. 3*A*). This anticipation was viseme-dependent ($F_{(4,56)} = 4.35$, $p = 0.003$) (Fig. 3*B*), and strictly followed the visual predictability gradient established behaviorally. Latency reduction was stronger for syllables that are associated with large and unambiguous mouth movements (high visual predictability: /ja/) than for syllables associated with ambiguous facial movement (low visual predictability: /ga/) (Pearson's $r = 0.30$, $p = 0.009$). Anticipation of M100A by visual input was also accompanied by a change in M100 amplitude ($F_{(4,56)} = 4.06$, $p = 0.006$) (Fig. 3*C*) that affected all syllables ($F_{(1,14)} = 8.84$, $p = 0.010$), without precisely following the behavioral visual predictability gradient ($r = 0.11$, $p = 0.332$) (Fig. 3*C*).

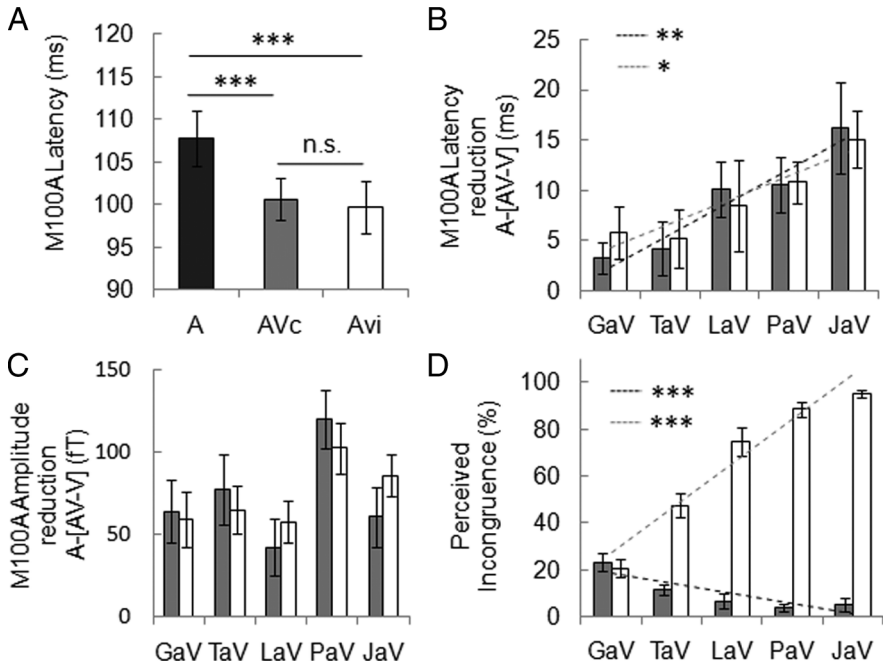

Figure 3. Facilitation of early auditory response by visual input. ***A***, Auditory evoked response (M100A) latency (A, dark bar) was globally reduced by visual syllables whether they matched the sound (AVc, gray bar) or not (AVi, white bar). ***B***, M100A latency reduction [A-(AV–V)], represented as a function of visual predictability (Fig. 2), shows a significant viseme dependency but no effect of incongruence. M100 latency reduction is proportional to visual predictability in both AVc (black dashed line) and AVi (gray dashed line) combinations. No significant difference between AVc and AVi regression slopes was found. ***C***, M100A amplitude change (positive values correspond to a reduction of M100A in AV–V vs A condition) indicates a significant effect of syllables but no viseme dependency or incongruence effect. ***D***, Perceived incongruence for AVc and AVi combinations. Note that comparisons focus on the visual syllable (for example PaAVc is compared with PaAVi, e.g., PaV/GaA) (supplemental Fig. 1*B*, available at www.jneurosci.org as supplemental material). Perceived incongruence for AVi pairs correlates positively with visual predictability (gray dashed line), whereas perceived incongruence for AVc pairs correlates negatively with visual predictability (black dashed line, interaction significant). Error bars indicate SEM. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

M100 insensitivity to mismatch

We examined whether direct corticocortical and feedback schemes can be distinguished on the basis of incongruence effects. As auditory and visual speech information converge on the STS (Calvert et al., 1997), feedback signal from the STS to auditory cortex should convey a phonologically more detailed prediction than one from visual areas, which should merely convey the amount and timing of facial movements. Because the more detailed the predictive signal, the stronger the mismatch when expectation is not met, we assume audiovisual mismatch to interact with facilitation proportionally to audiovisual incongruence. Thus, we expected audiovisual incongruence to interact with the early facilitation effect only if the latter was mediated by a feedback route (pathway 2), but not if was driven by a direct input from visual regions (pathway 1).

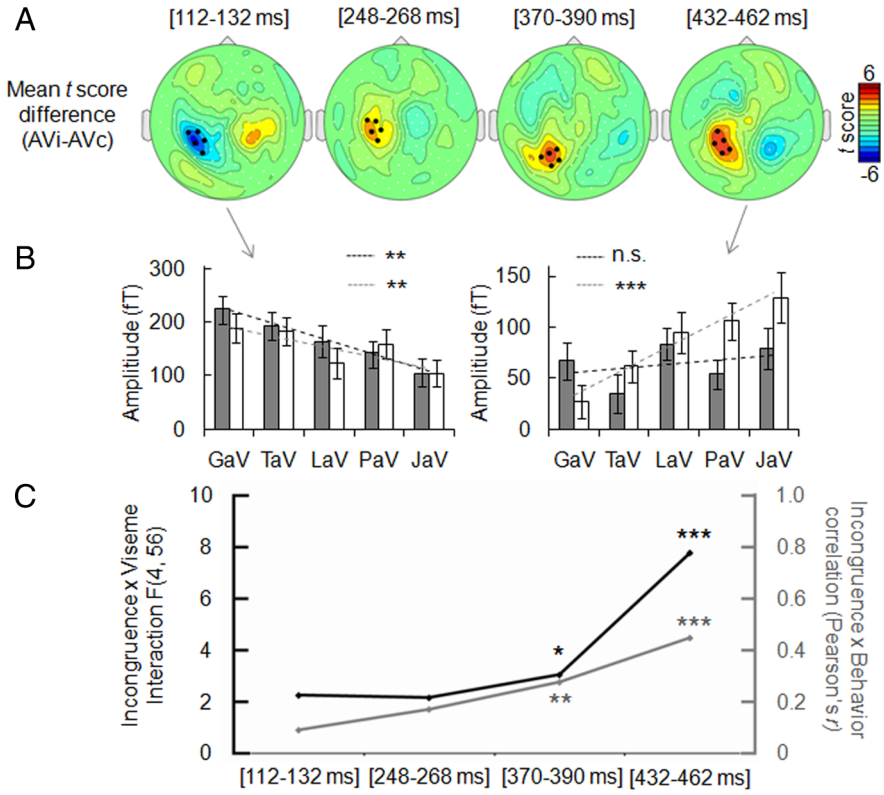

Figure 4. Effect of incongruence on ERFs across time. ***A***, Scalp topographies within the four time windows in which neural incongruence effect was detected (paired *t* test; grand average of AVi vs AVc conditions, with the overall sum of stimuli in AVi and AVc conditions physically the same) (supplemental Fig. 1*A*, available at www.jneurosci.org as supplemental material). ***B***, Effect of incongruence on viseme dependency of neural response amplitude, tested across those five selected sensors (black dots on topographies) showing a maximal effect, within the two extreme time windows. Dark and light gray dashed lines represent the correlations between amplitude and predictability in AVc and AVi conditions respectively. ***C***, Parallel between neural responses and behavioral reports related to incongruence. Left axis (dark line) indicates incongruence-by-viseme interaction *F* values (significant for the last 2 time windows) at the ERF level. The gray line shows that correlation values (Pearson's *r*, right axis) between ERF amplitude differences and perceived incongruence difference for each AVi versus AVc pair also increases over time. Error bars indicate SEM. n.s., Nonsignificant effect. $*p < 0.05$, $***p < 0.001$.

We also measured changes on auditory M100 amplitude and latency induced by the presence of visual syllables, when visual and auditory inputs were incongruent. Incongruence was generated by randomly combining the five visual and sound tracks of the videos, while excluding McGurk combinations, which resulted in five distinct levels of perceived incongruence (Fig. 3*D*). In the second pilot behavioral experiment, we established that visual predictability determined the perceived level of incongruence in physically incongruent syllables. In other words, the most predictive viseme yielded the strongest incongruence sensation (Pearson's $r = 0.78$, $p = 0.000$) (Fig. 3*D*). We further observed that physically congruent syllables could also evoke an incongruent percept when visual information was ambiguous, i.e., in the least predictable visual syllables ($r = -0.53$, $p = 0.000$) (Fig. 3*D*). When comparing responses evoked by incongruent and congruent stimuli, we did not detect neural mismatch effects on M100 peak neither in amplitude [$F_{(1,14)} = 0.03$, $p = 0.864$) (Fig. 3*C*), nor in latency ($F_{(1,14)} = 0.01$, $p = 0.906$) (Fig. 3*A*,*B*)], which confirms previous observations (Stekelenburg and Vroomen, 2007). Audiovisual mismatch did not alter viseme dependency of M100 facilitation, neither in amplitude ($F_{(4,56)} = 3.57$, $p = 0.011$) (Fig. 3*C*) nor in latency ($F_{(4,56)} = 3.80$, $p = 0.008$) (Fig. 3*B*). This result implies that speech was facilitated only as a function of the physical characteristics of the visual input and regardless of its discordance with the auditory input. Our results hence suggest that facilitation corresponds to a viseme-dependent signal that reflects the relative predictability of facial motion, but conveys only imprecise phonological information. Alternatively facilitation could occur at a too early stage of auditory processing to take full benefit of a phonological prediction.

Time course of neural mismatch

As we found no effect of audiovisual incongruence on M100 and no interaction between incongruence and M100 latency shortening, we examined the time course of neural mismatch in MEG responses to determine when physical mismatch between auditory and visual stimuli was reflected in cortical responses. When we compared the time course of responses evoked by incongruent versus congruent stimuli, we found that significant amplitude changes were evoked by the mismatch between auditory and visual inputs (Fig. 4). The earliest mismatch effect ($F_{(1,14)} = 25.4$, $p = 0.000$) was detected ~120 ms after voice onset (Fig. 4*A*), i.e., 20 ms after the facilitation effect. As we also detected a 20 ms difference between the M170 peak to visual syllables in motion-sensitive cortex and the M170 peak in the STS (supplemental Fig. 3, available at www.jneurosci.org as supplemental material), these results are compatible with a phonological signal arising from the STS with a delay relative to the motion signal. Of note, auditory cortex and STS peaked simultaneously in response to visual syllables. This suggests that a motion signal splits between two targets, the auditory cortex and the STS, and is compatible with a secondary feedback subsequently reaching the auditory cortex from the STS. Neural mismatch increased for another 300 ms showing three more maxima at ~250 ms ($F_{(1,14)} = 8.84$, $p = 0.010$), 370 ms ($F_{(1,14)} = 13.39$, $p = 0.003$) and 460 ms ($F_{(1,14)} = 18.94$, $p = 0.001$) (Fig. 4*A*), indicating at least three more steps in which visual prediction and bottom-up auditory signal are being compared. A repeated 100 ms delay between successive mismatch maxima is suggestive of iterative loops of interactions between motion-sensitive cortex, auditory cortex and the STS. At each iteration, the strength of the interaction between viseme predictability and neural mismatch increased. This interaction effect became statistically significant and additionally correlated with perceived incongruence at ~350 ms (Fig. 4*B*,*C*). Dynamic source reconstructions show that auditory cortex/STS/motion cortex loops occur when audiovisual convergence is not reached, i.e., when syllables are presented visually only (supplemental Video 1, available at www.jneurosci.org as supplemental material), and when audio and visual tracks do not match (supplemental Video 2, available at www.jneurosci.org as supplemental material). When ambiguity is low, i.e., in response to auditory only or to congruent stimuli, neural activity flows toward anterior regions in the ventral temporal cortex (supplemental Video 3, available at www.jneurosci.org as supplemental material). Although source recon-

structions are only qualitative evidence, the timing of neural mismatch supports a secondary feedback signal from the STS to the auditory cortex.

Visual motion sources

The temporal dynamics of evoked visual responses provides additional arguments to distinguish between direct corticocortical/pathway 1 and feedback/pathway 2 (supplemental Fig. 3, available at www.jneurosci.org as supplemental material). The earliest evoked response to visually presented syllables (M170) (Fig. 1*B*) peaked in motion-sensitive visual cortex (in accord with Besle et al., 2008) earlier than in the STS (paired *t* test, $p = 0.025$). Our results thus tend to point to motion-sensitive cortex as the most plausible origin of an early visual prediction signal, further arguing for direct corticocortical pathway 1. Although we emphasized the detection of motion response by using a still-face video as a baseline, we also observed a weak source in V1/V2, which could reflect a complementary influence of earlier visual areas that have been shown to project on auditory cortex in animals (Rockland and Ojima, 2003; Cappe and Barone, 2005). Despite their high anatomical plausibility these projections unlikely contribute to M100A facilitation in a viseme-dependent manner.

The amplitude of M170 over the occipital sensors in response to silent videos increased as a function of syllable visual predictability (Pearson's $r = 0.46$, $p = 0.000$). The amount of information reaching the auditory cortex from motion-sensitive cortex could thus determine the strength of the facilitation induced by visual signal in this natural audiovisual context. Yet, we could not observe significant statistical relationship between M170V amplitude and the strength of the facilitation, neither on amplitude nor on latency of auditory M100. This may stem from the fact that visually driven correlations between M170 and M100 responses were masked by inter-individual variability even after normalization. MEG source reconstruction on the other hand, is moderately reliable to localize the origin of a visual predictive signal (Fig. 5*A*). Whether response magnitude in motion-sensitive cortex determines the amount of visual facilitation therefore remains unclear.

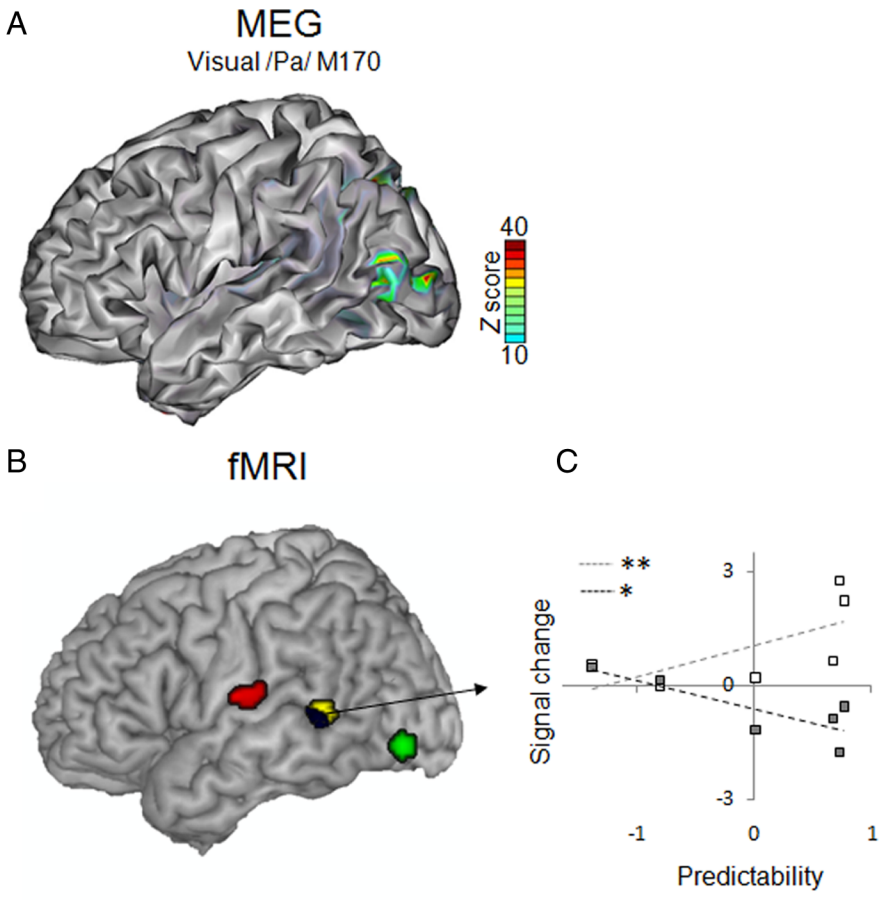

Figure 5. Surface renderings of MEG sources and fMRI activations. ***A***, Source reconstruction of M170 peak measured in response to the viseme /pa/ shows that early activity related to lips movements emerges in the temporo-occipital cortex (visual motion cortex, as separately assessed by functional localizer). ***B***, Summary of fMRI findings: parametric increase with syllable visual predictability (green blob) overlaps with the sources of M170 shown in ***A***. Functional connectivity was assessed using PPI with visual syllable recognition rates as the psychological variable. There was a parametric increase of functional connectivity between visual motion cortex and auditory regions surrounding Heschl's gyrus (red blobs). The middle STS showed the opposite effect, i.e., a decrease of functional connectivity as a function of visual predictability (yellow blob and blue blob) when using both visual motion and auditory cortices as seed regions. ***C***, Activity in STS also reflects the amount of prediction error, showing a signal increase for incongruent stimuli (white squares) and a signal decrease for congruent stimuli (gray squares), proportionally to visual predictability. $^*p < 0.05$, $^{**}p < 0.01$.

Effect of visual predictability on hemodynamic responses

We used fMRI with similar stimuli as in the MEG experiment (supplemental Fig. 2, available at www.jneurosci.org as supplemental material), but instead of detecting a target syllable subjects had to answer whether the presented stimulus corresponded or not to a syllable proposed in written, e.g., "Pa?". In the fMRI time series, we probed brain regions in which hemodynamic activity increased parametrically with visual predictability as assessed during MRI acquisition using individual recognition rates for each syllable (visual only). In a whole brain analysis, increasing visual recognition rates of syllables correlated with bilateral activity of lateral extrastriate occipital cortex (Fig. 5*B*, green blob, Table 1*a*), which according to its coordinates could correspond to V5/hMT+ (Malikovic et al., 2007). We confirmed the sensitivity to motion of this region by a speech motion functional localizer (see Materials and Methods). This effect is consistent with the above-described correlation between visual predictability and M170 amplitude. Visual predictability enhanced functional connectivity between left motion-sensitive cortex and the left perisylvian region, i.e., rolandic operculum (Fig. 5*B*, red blob, Table 1*d*), supramarginal gyrus ($p = 0.022$), and medial auditory posterior insula ($p = 0.029$). That the target of connectivity from motion-sensitive cortex was not located within the primary auditory cortex fits with intracortical recordings in humans (Besle et al., 2008), and with anatomical studies in animals showing that fibers arising from motion-sensitive cortex target the belt but not the primary auditory region (Cappe and Barone, 2005). These results are compatible with our MEG data, and further show that the region that is contacted by visual inputs is normally involved in both expressive and receptive phonology (Hickok and Poeppel,

Table 1. Peak coordinates of activated clusters in the fMRI experiment

Parameter/anatomical description	BA	MNI coordinates			*Z* score	*p* value
		x	*y*	z		
Main parametric effects of predictability						
a. Positive covariation between individual predictability scores and responses to V stimuli						
Visual motion cortex	19	−38	−80	−6	4.18	0.000
		38	−68	6	4.07	0.000
R. superior occipital gyrus	17	16	−96	0	4.03	0.000
L. inferior frontal gyrus (p.Orbit.)	47	−28	30	−16	4.26	0.000
b. Negative covariation between individual predictability scores and responses to AVc stimuli						
Superior temporal sulcus	22	−58	−52	12	3.46	0.000
		−52	−46	6	3.25	0.001
	21	56	−40	0	3.85	0.000
L. inferior frontal gyrus (p.Operc.)	44	−58	10	26	3.83	0.000
L. insula		−46	4	0	4.15	0.000
Putamen		−24	8	8	3.93	0.000
		36	6	−4	3.55	0.000
c. Positive covariation between individual predictability scores and responses to AVi stimuli						
L. superior temporal sulcus	22	−56	−50	6	4.00	0.000
L. inferior frontal gyrus (p.Operc.)	44	−50	14	14	3.87	0.000
Precentral gyrus	6	−38	−2	44	3.77	0.000
		48	4	48	3.82	0.000
R. inferior parietal lobule	40	42	−42	36	3.48	0.000
R. cerebellum (VI)		36	−50	−28	3.84	0.000
PPI with predictability from visual motion cortex						
d. Individual increasing predictability scores						
L. auditory/rolandic operculum	43	−48	−12	16	2.60	0.01
e. Individual decreasing predictability scores						
L. superior temporal sulcus	21	−66	−48	8	2.34	0.01
PPI with predictability from Heschl's gyrus						
f. Individual increasing predictability scores						
L. visual motion cortex	19	−36	−68	−2	2.31	0.01
g. Individual decreasing predictability scores						
L. superior temporal sulcus	21	−48	−38	6	2.45	0.01

BA, Broadmann area; R, right hemisphere; L, left hemisphere; p.Operc., pars opercularis; p.Orbit, pars orbitalis. p thresholds (uncorrected): for a., b., and c., $p < 0.001$; for d., e., f., and g., $p \leq 0.01$. For PPI analysis, reported are maxima appearing within a 2-cm-radius sphere surrounding the following regions of interest visual motion cortex [−38, −80, −6], Heschl's gyrus [−50, −20, 2], and the STS [−54, −50, 6].

2007). Note that PPIs cannot permit to infer directionality. However, as connectivity varied as a parametric function of visual input neural information likely flows from motion-sensitive cortex to auditory cortex.

In contrast, an increase in functional connectivity between left motion-sensitive cortex and the middle STS was found when visual predictability decreased, i.e., when visual syllable ambiguity increased (Fig. 5*B*, yellow blob, Table 1*e*). Functional connectivity also increased between the left STS and the auditory cortex [near Heschl's gyrus according to cytoarchitectonic templates (Morosan et al., 2001)] when viseme ambiguity increased (Fig. 5*B*, blue blob, Table 1*g*). As the STS could not be taken as a source for the functional connectivity analysis, we could not determine whether a feedback from the STS targets the same or a different region than the corticocortical projection from motion-sensitive cortex. However, enhanced neural activity for visual ambiguity was confirmed in this region when directly tracking regions in which activity decreased with visual predictability in natural stimuli (AV congruent stimuli) (Fig. 5*C*, gray squares) ($r = -0.21$, $p = 0.038$) (Table 1*b*), and also when probing effects that covaried positively with predictability when auditory and visual inputs did not match (AV incongruent stimuli) (Fig. 5*C*) (white squares, $r = 0.29$, $p = 0.004$) (Table 1*c*).

Discussion

We used MEG and fMRI to distinguish the contribution of two possible neural connectivity patterns by which visual speech may facilitate auditory responses. Our MEG design takes advantage of the 150 ms natural delay between visual and auditory onset in natural speech to track independently predictive visual signals and their cross-modal facilitation effect. By selecting visual syllables associated with increasing recognition rates (van Wassenhove et al., 2005) we implemented an incremental visual prediction and confirmed that the early auditory cortical MEG component (M100) is sped up by the presence of congruent facial movements (van Wassenhove et al., 2005) proportionally to their visual predictive power. While MEG provided a good temporal resolution to detect an early visual effect in auditory cortex, we used fMRI data to more precisely localize the auditory target of this effect and to explore the role of a potential feedback from the STS to auditory cortex in audiovisual integration.

Viseme dependency of M100 facilitation denotes projections that selectively convey visual motion cues, which could be the case for both a direct corticocortical path from visual to auditory cortex (pathway 1) (Falchier et al., 2002; Rockland and Ojima, 2003; Cappe and Barone, 2005) and a feedback path from the STS (pathway 2). Yet, as viseme-specific facilitation of early auditory response M100 was not influenced by audiovisual incongruence, we assume that visual input drives a fast prediction that does not depend on any audiovisual comparison, hence should not arise from a multimodal region in which auditory and visual speech input are being integrated, e.g., the STS.

FMRI data confirmed that the source of this effect is the motion-sensitive cortex, as syllable visual predictability enhanced connectivity between motion-sensitive cortex and auditory regions. Several authors propose that the direct corticocortical pathway corresponds to a modulatory input to auditory cortex

from visual cortex targeting agranular layers (Kayser et al., 2007, 2008; Lakatos et al., 2008; Schroeder et al., 2008). This input could reset the phase of ongoing oscillatory activity, thus bringing the auditory system into a periodically receptive, excitatory state during which auditory neural response is enhanced (Lakatos et al., 2007). In this framework motion-sensitive cortex could tune neuronal activity in auditory cortex to the forthcoming sound track, with little phonological accuracy, but precise timing. Phase resetting of auditory cortex by visual syllables occurring periodically at the frequency of jaw movements, i.e., the syllabic (theta) rhythm, could account for the behavioral benefit of viewing speaker's lip movements (Sumby and Polack, 1954; Chandrasekaran et al., 2009). Given the specificity of the observed effect we believe that a syllable-dependent phase reset from motion-sensitive cortex is a plausible interpretation of our data. Phase resetting, is only one of several possible mechanisms by which visual input could facilitate auditory processing. First, it is not excluded that eye movements (that were only partly monitored in this experiment) could phase-reset auditory activity thereby structuring and facilitating the auditory response (Chandrasekaran et al., 2009; Melloni et al., 2009). Second, our results do not prove that input from motion-sensitive cortex is modulatory rather than of feedforward nature. While enhanced firing synchronization can be directly detected in single and multiunit recordings [electrocorticography, local field potential (LFP) recordings], visual "facilitation" of auditory evoked fields recorded from the scalp could reflect other physiological mechanisms as, for instance, neural priming. A sharper response accompanied or not with a reduced latency could reflect that less neurons are responding to the repeated input (Grill-Spector et al., 2006).

Here, we observed facilitation jointly on M100 amplitude and latency. While latency reduction depended on visual predictability, amplitude reduction did not. It is possible that amplitude reduction is underpinned by a different neural mechanism than latency reduction. M100 amplitude reduction could be determined by the number of phonological solutions (ambiguity) corresponding to a given viseme, e.g., two solutions for a visual /ja/ versus up to six alternatives for a visual /ga/, while latency reduction could be driven by the timing of the mouth movement onset. This remains unclear, as we did not find a correlation between M100 latency reduction and visual onset ($r = -0{,}2$; $p = 0{,}88$).

When visual and auditory signals were not congruent, the evoked response was enhanced and delayed (mismatch effect). This effect was observed repeatedly, four times, every 100 ms from 120 ms on. That mismatch was observed 20 ms after M100, and does not interfere with the viseme dependency of M100 facilitation, is compatible with a possible additional, feedback mechanism following the direct visual-to-auditory motion signal. The anatomical location of mismatch effects in the MEG dataset varied in time but is generally compatible with the STS being the central point of a periodic audiovisual comparison and a periodic feedback to auditory cortex. Mismatch, hence audiovisual comparison, occurred repeatedly every 100 ms, i.e., around 10 Hz, a frequency that is close to the beta band over which STS and auditory cortex are found to interact using intracortical recordings in animals (Kayser and Logothetis, 2009). The long 100 ms delay between each audiovisual comparison, relative to a three relay synaptic delay, could reflect cumulated integration times in auditory cortex, STS, and motion-sensitive cortex. FMRI results indicate with higher topographical precision than MEG that the STS serves as a comparator as its activity strictly followed the degree of perceived audiovisual incongruence, and negatively correlated with visual predictability. This double observation speaks to the theory of predictive coding that stipulates that neural activity (as assessed with LFP recordings and fMRI) reflects the difference between predicted and incoming signals, resulting in lower activity when a signal is correctly anticipated (Friston, 2005). This principle is compatible with a more synchronized neuronal spiking that makes global neural activity appear overall sharper when a stimulus is primed (Grill-Spector et al., 2006). If we assume that signals from motion-sensitive cortex synchronize auditory cortex neural activity each time a syllable is being pronounced, feedback signals from the STS on the other hand might essentially reflect the attempt to synchronize audio and visual inputs when they do not match. The current experiments only allow us to speculate about the physiological mechanism by which visual prediction and prediction error modulate auditory cortex activity. The most parsimonious account of our observations would be that direct (pathway 1) and indirect (pathway 2) routes yield distinct effects using a similar mechanism.

(1) The direct route might convey direct phase resetting by motion-sensitive cortex resulting in tuning auditory cortex to the upcoming sound regardless of whether auditory and visual inputs match (unsupervised mechanism).

(2) The indirect route might drive phase resetting via the STS in a distinct manner, depending on whether auditory and visual inputs do or do not match. (a) If auditory and visual stimuli do match, the STS receives convergent input from both modalities, and hence sends back a focal feedback to auditory and motion-sensitive cortices, resulting in audiovisual fine-tuning, and possibly further speed-up of auditory processing. We presumably cannot detect such a focal effect in evoked responses that reflect neural synchronization of large neuronal populations. (b) If auditory and visual inputs do not match, the total number of activated STS neurons should be larger than in case (a), i.e., approximately the sum of those recruited by visual and by auditory input. Therefore, the feedback signal should target a larger portion of auditory cortex, hence also improving audiovisual tuning. This tuning process could be successful, e.g., in McGurk, or fail resulting in mismatch perception. Our MEG results, showing four maxima in the neural mismatch, with only the last two significantly correlated to perceived incongruence, likely indicate that it takes several motion cortex/STS/auditory cortex loops (over ~500 ms) for incongruence to be stably inferred.

This proposal needs to be tested using intracortical recordings in humans and animal models, but altogether both MEG and fMRI results converge to show a fast direct visual-to-auditory predictive mechanism, immediately followed by a secondary feedback involving the STS as a central audiovisual comparator, and working in dual loops with the auditory cortex on one side and the motion-sensitive cortex on the other side (supplemental Video 4, available at www.jneurosci.org as supplemental material).

Although this study focused on cortical circuits within the temporal lobe, i.e., on the STS, Schroeder et al. (2008), proposed three additional cerebral sources that could phase reset auditory cortical neurons: the nonspecific thalamus (and possibly also the visual thalamus), the parietal and the prefrontal cortices. As our model only tests for the direct corticocortical and STS feedback pathways we do not rule out a contribution of these other regions. Negative results were obtained when we tested for a possible viseme-dependent response in the thalamus in the fMRI dataset, but given the limitations of the approach, a thalamic contribution to the facilitation effect is not excluded. Regarding parietal and prefrontal input to auditory cortex, dynamic source reconstruction of responses to visual syllables showed that each of these

regions responded shortly after stimulus presentation (yet later than motion-sensitive cortex) (supplemental Fig. 3, available at www.jneurosci.org as supplemental material) (Bernstein et al., 2008b). The STS was the latest of these three regions to respond, and it peaked simultaneously with the auditory cortex ~20 ms after the motion-sensitive occipital cortex. The timing of activation in the prefrontal cortex is not compatible with a feedback to auditory cortex, as it peaks too late, but a feedback from the parietal cortex cannot be ruled out. MEG data alone thus suggest that the fastest effect is a direct visual to auditory one (pathway 1), but do not clearly distinguish between several additional feedback sources (pathway 2). Analyses of the fMRI data, on the other hand, showing (1) enhanced connectivity from the STS with both motion-sensitive cortex and the auditory cortex when predictability decreases, and (2) an increase in activity with visual "ambiguity," confirm the involvement of the STS as a major functional interface between visual and auditory cortices, and a more specific role in audiovisual speech perception than the parietal cortex.

References

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.

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

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

Bernstein LE, Lu ZL, Jiang J (2008a) Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. Brain Res 1242:172–184.

Bernstein LE, Auer ET Jr, Wagner M, Ponton CW (2008b) Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39:423–435.

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234.

Besle J, Fischer C, Bidet-Caulet A, Lecaignard F, Bertrand O, Giard MH (2008) Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. J Neurosci 28:14301–14310.

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

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596.

Cappe C, Barone P (2005) Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. Eur J Neurosci 22:2886–2902.

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

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11–23.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749–5759.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

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.

Ghazanfar AA, Chandrasekaran C, Logothetis NK (2008) Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J Neurosci 28:4457–4469.

Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 10:14–23.

Hertrich I, Mathiak K, Lutzenberger W, Menning H, Ackermann H (2007) Sequential audiovisual interactions during speech perception: a whole-head MEG study. Neuropsychologia 45:1342–1354.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

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

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.

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

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.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320:110–113.

Malikovic A, Amunts K, Schleicher A, Mohlberg H, Eickhoff SB, Wilms M, Palomero-Gallagher N, Armstrong E, Zilles K (2007) Cytoarchitectonic analysis of the human extrastriate cortex in the region of V5/MT+: a probabilistic, stereotaxic map of area hOc5. Cereb Cortex 17:562–574.

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748.

Melloni L, Schwiedrzik CM, Rodriguez E, Singer W (2009) (Micro)Saccades, corollary activity and cortical oscillations. Trends Cogn Sci 13:239–245.

Miller LM, D'Esposito M (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci 25:5884–5893.

Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K (2001) Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13:684–701.

Rockland KS, Ojima H (2003) Multisensory convergence in calcarine visual areas in macaque monkey. Int J Psychophysiol 50:19–26.

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A (2008) Neuronal oscillations and visual amplification of speech. Trends Cogn Sci 12:106–113.

Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19:1964–1973.

Sumby WH, Polack I (1954) Perceptual amplification of speech sounds by visual cues. J Acoust Soc Am 26:212–215.

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102:1181–1186.

von Kriegstein K, Dogan O, Grüter M, Giraud AL, Kell CA, Grüter T, Kleinschmidt A, Kiebel SJ (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105:6747–6752.