Audio-visual combination of syllables involves time-sensitive dynamics following from fusion failure

Sophie Bouton[1,2,3], Jaime Delgado-Saa[1,4], Itsaso Olasagasti[1], and Anne-Lise Giraud[1]*

[1]*Department of Basic Neuroscience, University of Geneva, Biotech Campus, 9, Chemin des Mines, Geneva 1211, Switzerland*

[2]*Centre de Recherche de l'Institut du Cerveau et de la Moelle Epinière & Centre de Neuro-imagerie de Recherche, Paris, 75013, France*

[3]*Laboratoire Dynamique du Langage, CNRS & Université de Lyon UMR 5596, 69007 Lyon, France*

[4]*Biomedical Signal Processing and Artificial Inteligence Laboratory, Universidad del Norte, Barranquilla, Colombia*

Sophie Bouton and Jaime Delgado-Saa contributed equally to this work.

*Correspondence and requests for materials should be addressed to ALG. (anne-lise.giraud@unige.ch)

Abstract (137 words)

In face-to-face communication, audio-visual (AV) stimuli can be fused, combined or perceived as mismatching. While the left superior temporal sulcus (LSTS) is admittedly the locus of AV integration, the process leading to *combination* is unknown. Analysing behaviour and time-/source-resolved human MEG data, we show that fusion and combination both involve early detection of AV physical features discrepancy in the LSTS, but that this initial registration is followed, in combination only, by the activation of AV asynchrony-sensitive regions (auditory and inferior frontal cortices). Based on dynamic causal modelling and neural signal decoding, we further show that AV speech integration outcome primarily depends on whether the LSTS quickly converges or not onto an existing multimodal syllable representation, and that combination results from subsequent temporal re-ordering of the discrepant AV stimuli in time-sensitive regions of the prefrontal and temporal cortices.

Keywords

Audio-visual integration, Combination, McGurk effect, Neural dynamics, Audio-visual asynchrony.

Introduction

Screen-based communication poses specific challenges to our brain for integrating audiovisual (AV) disparities due to either asynchronies between audio and visual signals (e.g. facetime, skype) or to mismatching physical features (dubbed movies). To make sense of discrepant audio-visual speech stimuli, our brain mostly focuses on the auditory input, which is taken as ground truth, and tries to discard the disturbing visual one. In some specific cases, however, AV discrepancy goes unnoticed and the auditory and visual inputs are implicitly *fused* into a percept that corresponds to none of them. More interestingly perhaps, discrepant AV stimuli can also be *combined* into a composite percept where simultaneous sensory inputs are perceived sequentially. These two distinct outcomes can experimentally be obtained using "McGurk effect"[1], where an auditory /aba/ dubbed onto a facial display articulating /aga/ elicits the perception of a fused syllable /ada/, while an auditory /aga/ dubbed onto a visual /aba/ typically leads to a mix of the combined syllables /abga/ or /agba/. What determines whether AV stimuli are going to be fused[2–4] or combined[5], and the underlying neural dynamics of such a perceptual divergence is not known yet.

Audio-visual speech integration draws on a number of processing steps distributed over several cortical regions, including auditory and visual cortices, the left posterior temporal cortex, and higher-level language regions of the left prefrontal[6,7] and anterior temporal cortex[8,9]. In this cortical hierarchy, the left superior temporal sulcus (LSTS) plays a central role in integrating visual and auditory inputs from the visual motion area (mediotemporal cortex, MT) and the auditory cortex (AC)[10–15]. The LSTS is characterized by relatively smooth temporal integration properties that enables it to cope with the natural asynchrony between auditory and visual speech inputs, i.e. the fact that orofacial speech movements often start before the sounds they produce[4,16,17]. Although the LSTS responds better when auditory and visual speech are perfectly synchronous[18,19], its activity can cope with large temporal discrepancies, reflecting a broad temporal window of integration in the order of the syllable length (up to ~260 ms)[20]. This large window of integration can even be pathologically stretched to about 1s in subjects suffering from autism spectrum disorder[21]. Yet, the detection of shorter temporal AV asynchronies is possible and takes place in other brain regions, in particular in the dorsal premotor area and the inferior frontal gyrus[22–25].

The LSTS and the IFG regions hence exhibit different functions in AV speech integration, depending on their temporal integration properties[26]. Interestingly, a relative resilience to asynchrony could confer the LSTS a specific sensitivity to the incongruence of *physical* features across A and V modalities. A key function of the LSTS could hence be to resolve AV speech feature discrepancies[27] via a process requiring a double sensitivity to canonical visual motion (lip movements) and auditory spectrotemporal (formant transitions) cues.

To characterize the mechanism(s) underlying integration of A and V physical speech features (in the LSTS), we previously developed a generative predictive coding model[28] that explored whether cross-modal predictions and prediction errors could be utilized to combine speech stimuli into different perceptual solutions, corresponding to *fused*, i.e., /ada/, or *combined*, i.e., /abga/, percepts (supp. Note 1). The model showed that considering the temporal patterns in a 2nd acoustic formant/lip aperture two-dimensional (2D) feature space is sufficient to qualitatively reproduce participants' behaviour for *fused*[13,29] but also *combined* responses[28]. Simulations indicated that fusion is possible, and even expected, when the physical features of the A and V stimulus, represented by the 2nd formant and lip in the model, are located in the neighbourhood of an existing 2D syllable representation. This is the case for the canonical McGurk stimulus, which falls in the neighbourhood of /ada/, when the input corresponds to the visual features of /aga/ and auditory features of /aba/. Conversely, audio-visual stimuli having no valid syllable representation in their 2nd formant/lip neighbourhood (Figure 1A) lead to the sensation that the two (quasi) simultaneous consonants /b/ and /g/ are being pronounced sequentially, e.g. the combination percepts /abga/ or /agba/[30–32].

These theoretical data lend support to the recent proposal that the recognition process is very similar for congruent AV stimuli and for AV *fusion*[33], but different for AV combination[5]. However, the mechanisms leading to such a diverging dynamics have never been characterized. Here we test two alternative hypotheses: 1) *combination* readily involves fine detection of AV asynchrony (outside the LSTS, most likely in the left IFG) and the percept arises on-line from the real order with which each phoneme is detected (Figure 1B, right panel), or 2) the *combination* process is triggered by the failure of AV *fusion* in the LSTS, and combination results from a post-hoc reconstruction of the most plausible AV sequence. The latter hypothesis has two experimental implications. The first one is that the

delay to converge on a plausible solution should be longer in *combination* than *fusion*, because the latter requires solving AV discrepancy by explicitly serializing auditory and visual inputs into a complex consonant transition (Figure 1B). Extra processing time for *combination* relative to *fusion* should manifest both in reaction times and in the timing of neural events. The second implication is that combination should emerge from early AV comparison in the LSTS and only subsequently involve auditory and (articulatory) prefrontal cortices, to produce an ordered and articulable new composite syllable. In this process, the LSTS is expected to have a pivotal role and we should therefore observe enhanced functional interactions between the LSTS and these additional brain regions during *combination* relative to *fusion*.

Figure 1 *(legend on next page)*

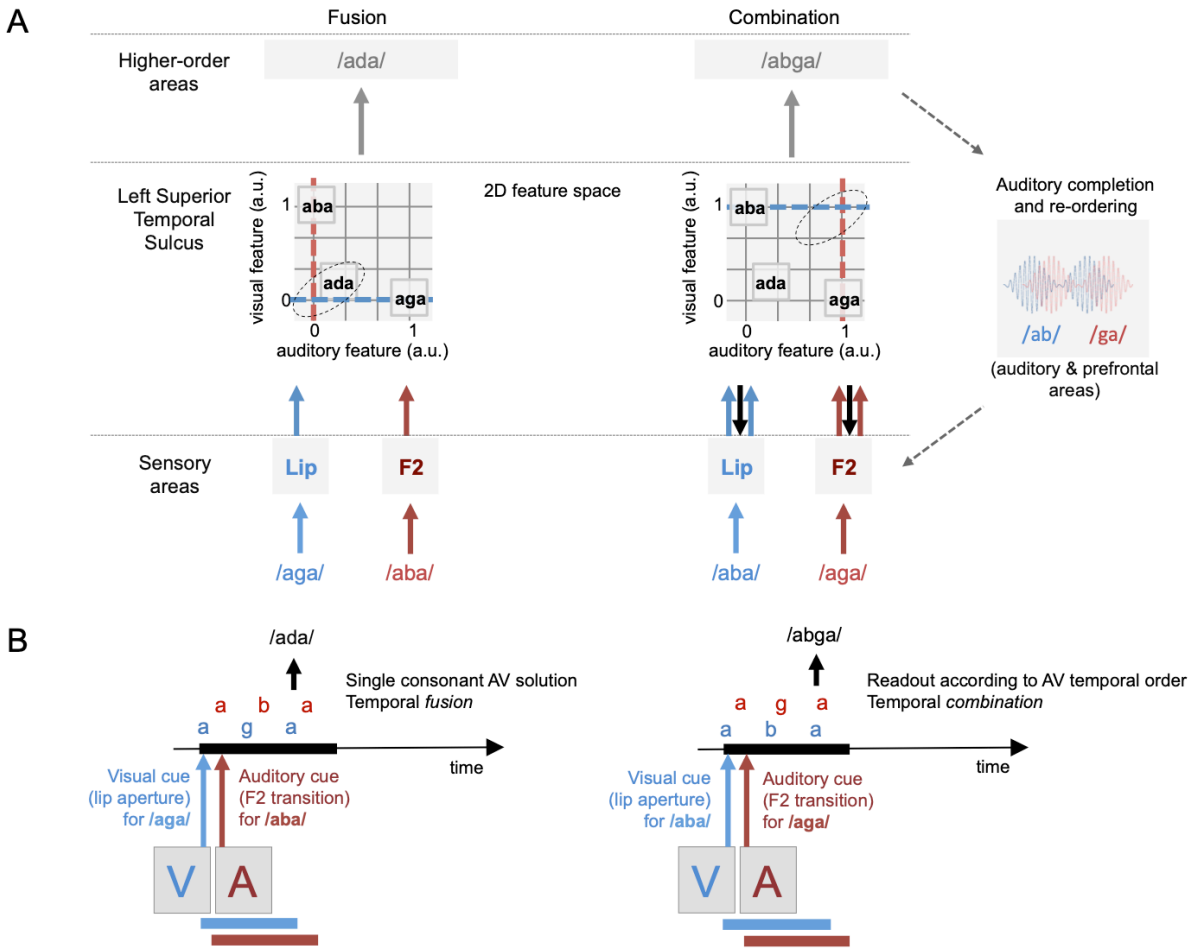

Figure 1. (A) Proposed neurophysiological mechanisms for *fusion* versus *combination*. We posit that after being processed by primary auditory and motion sensitive areas (bottom row), AV inputs converge in the left Superior Temporal Sulcus (LSTS, middle row) that works as a multidimensional feature space, here reduced to a simple 2D-space in which lip-motion and 2nd speech formant are the main dimensions. The LSTS is relatively insensitive to AV asynchrony (as depicted in B), but encodes both physical inputs in the 2D-space, converging on the most likely cause of a common speech source given these inputs. In the visual /aga/ - auditory /aba/ condition, coordinates in the 2D space fall close to those of the existing syllable /ada/, which is picked as solution such that the subject senses no conflict. In the visual /aba/ - auditory /aga/ condition, the absence of existing /aXa/ solution at coordinates crossing triggers post-hoc reconstruction of the most likely cause of the inputs via a complex consonant transition /abga/ (true temporal sequence), with occasional temporal inversions of the sound sequence /agba/[31]. Both combination outputs require additional interaction with time sensitive (prefrontal and auditory) brain regions. Grey arrows represent the LSTS output as readout by higher order areas. Blue and red arrows represent visual and auditory inputs, respectively. (B). Discrepant audio (A) and visual (V) syllabic speech units /aXa/ are represented within a critical time-window for integrating them as a single item coming from the same source. The auditory percept is either a McGurk fusion /ada/ (left) or a combination percept /abga/ (right).

Results

To address whether *combining* AV stimuli results from direct sensitivity to AV asynchrony or from *fusion* failure in the LSTS, we used the two canonical McGurk conditions, which give rise to either the *fused* percept 'ada' or 'ata', or a *combined* solution 'abga', 'agba', or 'apka', 'akpa'. Although these stimuli are artificial (see [34–36] for ecologically valid audiovisual stimulation), they allow for a rigorous parameterization of the various AV speech integration outcomes.

We first run two behavioural experiments carried out in distinct groups of participants. Both experiments involved vowel-consonant-vowel syllables of the type aXa denoted /aXa/ for audio and [aXa] for visual. These AV stimuli were used across three different conditions: (i) a *congruent* condition in which auditory and visual inputs corresponded to the same syllable (stimuli /ada/ + [ada] and stimuli /ata/ + [ata]), and two *incongruent* conditions in which auditory and visual inputs could give rise to either (ii) a *fusion* percept (stimuli /aba/ + [aga] and stimuli /apa/ + [aka]) or (iii) a *combination* percept (stimuli /aga/ + [aba] and stimuli /aka/ + [apa]) (Figure 2C). All stimuli were video clips showing either a female or a male articulating aXa stimuli belonging to the phoneme family 'bdg' or the phoneme family 'ptk'. In a first experiment, 20 participants performed a repetition task. Instructions were the same as those given in the McGurk & MacDonald article (1976): participants watched the videos and were asked to repeat what they "heard" as fast as possible (Figure 2A), with no restrictions on the pronounced syllable. Detailed behavioural analyses are presented in the Methods section.

AV *combination* takes more time than *fusion*

We compared the task's dependent variables (response of interest rate and response time) between the three conditions (*congruent*, *fusion* and *combination*) using two repeated-measures ANOVAs (see Methods, Figure 3, S1 Table, S2 Table). In the repetition task (behavioural experiment 1), the report rate for each response of interest (i.e. the percentage of 'ada' and 'ata' responses in the *fusion and congruent* conditions, and the percentage of 'abga', 'apka', 'agba' and 'akpa' responses in the *combination* condition) differed across conditions ($F(2,38) = 275.51$, $P < .001$), irrespective of the consonant family ($F < 1$) or speaker gender ($F < 1$) (S2 Table). Subjects reported more 'ada' and 'ata'

responses in the *congruent* than *fusion* condition ($t(19) = 8.45$, $P < .001$) (Figure 3A, left panel), but the mean rates for each response of interest were not different across *fusion* and *combination* conditions ($t(19) = 0.69$, $P > .20$). As expected, in the *fusion* condition participants mostly reported fused and auditory-driven responses. In the *combination* condition they reported mostly combined responses, but also auditory and visually-driven responses (S2 Table). In the three conditions, only response times (RTs) associated with the responses of interest were analysed, i.e., 'ada' and 'ata' responses in the *congruent* and *fusion* conditions, and 'abga'-'apka'-'agba'-'akpa' responses in the *combination* condition. RTs differed between conditions ($F(2,38) = 6.92$, $P = .001$): the syllable heard in the *fusion* condition was repeated as fast as in the *congruent* condition ($t < 1$, Cohen $d = 0.002$), whereas the delay to repeat the syllables heard was longer in the *combination* condition than in the *congruent* and *fusion* conditions ($t(19) = 3{,}98$, $P < .001$, Cohen $d = 0.61$, difference *combination – congruent* = 198 *ms*; $t(19) = 3.78$, $P < .001$, Cohen $d = 0.59$ difference *combination – fusion* = 191 *ms*, respectively) (Figure 3A, right panel), irrespective of the consonant family ($F < 1$) or speaker gender ($F < 1$) (S2 Table). Thus, RTs indicate that subjects were slower to integrate mismatching audio-visual inputs when they elicited a combined rather than a fused, percept. Importantly, repetition time for 'ada' or 'ata' was equal whether it arose from congruent or incongruent syllables, showing that AV incongruence was not at the origin of the slower RT for combination.

Although, this first experiment confirmed our proposal that participants should be faster to fuse than to combine AV stimuli, the effect could be biased by the difficulty to plan and articulate a more complex (double consonant) combination than a (single consonant) fusion syllable. To address this potential issue, we ran a second experiment in which 16 new participants performed a pairing task (behavioural experiment 2), where each trial included a written syllable followed by a video clip. Participants had to identify whether a syllable displayed on the screen before the video matched the syllable that was subsequently heard (Figure 2B). Interestingly, the score for each responses of interest in each condition was similar to that of the repetition task (Figure 3B, left panel). Subjects were better at identifying a *congruent* than an *incongruent* stimulus ($F(2,30) = 30.26$, $P < .001$). In addition, participants matched the written syllable with the video faster in the *congruent* and *fusion* conditions than in the *combination* condition ($F(2,30) = 6{,}84$, $P < .001$, partial η^2

= 0.92; difference *combination – congruent* = 206 ms, $t(15) = 3.06$, $P = 0.08$, Cohen = 0.58; difference *combination – fusion* = 186 *ms,* $t(15) = 2.82$, $P = 0.018$, Cohen = 0.54; difference *fusion – congruent* = 20 ms, $t(15) = 0.39$, $P = 0.694$, Cohen = 0.08), confirming our previous findings (Figure 3B, right panel), and showing that the extra delay for combination does not lie in added articulatory complexity. These data overall suggest that AV discrepancy was more easily solved in the *fusion* than in the *combination* condition, and that the integration of incongruent AV stimuli presumably relies on different neuronal processes depending on whether individuals end-up fusing or combining conflicting AV inputs.

Figure 2 *(legend on next page)*

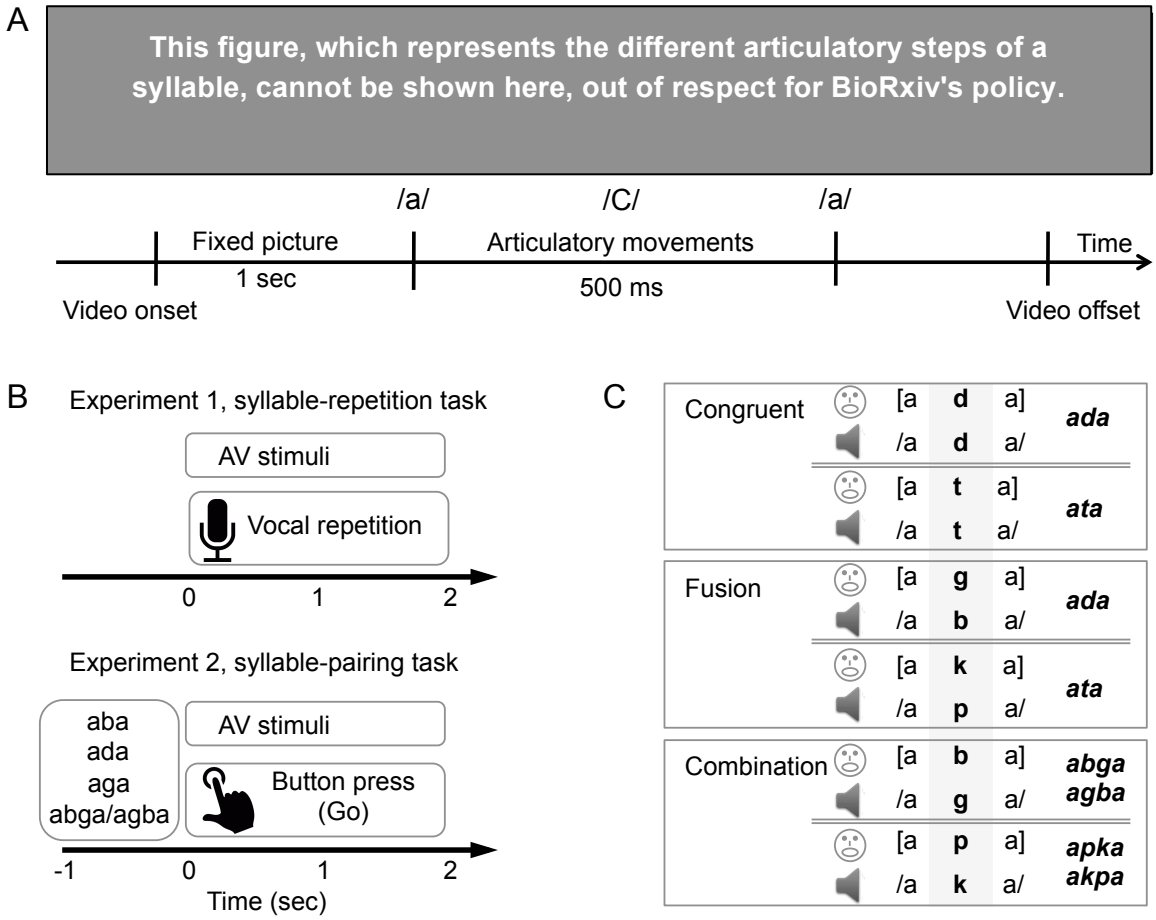

Figure 2. (A) Typical time course of audio-visual stimuli. (B) Example trials from experiments 1 and 2. Experiment 1: trials started with a 1s fixation period, followed by a videoclip showing a speaker pronouncing a syllable. Participants had to repeat the syllable they "heard" as fast as possible. Experiment 2: trials started with a 1s written syllable, followed by a short videoclip showing a speaker pronouncing a syllable. Participants were instructed to press a button as fast as possible if the written syllable matched the syllable they perceived from the AV videoclip. (C) Experimental conditions used in the behavioural and MEG experiments. The same three conditions, labelled '*congruent*', '*fusion*', and '*combination*' were used in the behavioural and neuroimaging experiments. In each condition, stimuli combined a video track and an audio track, and used two consonant families: either 'bdg' or 'ptk'. Right panel: the answer of interest (expected response depending on the exact AV stimulus combination) is shown in bold-italic.

Figure 3

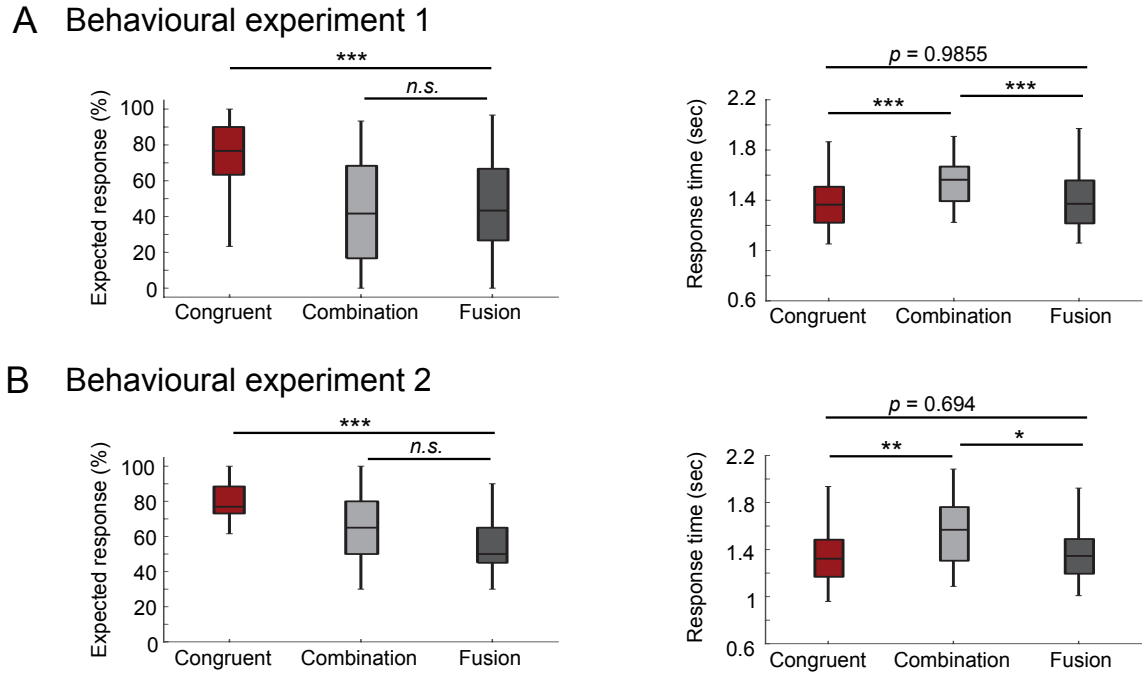

Figure 3. (A) Dependent variables from behavioural experiment 1, syllable-repetition task. (B). Dependent variables from behavioural experiment 2, syllable-pairing task. (A & B, left panels) Rate (%) of responses of interest in each condition, i.e., 'ada' or 'ata' responses in the *congruent* and *fusion* conditions, 'abga', 'agba', 'apka' or 'akpa' responses in the *combination* condition. (A & B, right panels) Response time for responses of interest in each AV condition. In (A & B) error bars correspond to s.t.d. Three stars indicate a significant difference at $P < .001$, two stars indicate a significant difference at $P < .01$, and n.s. indicate a non-significant difference ($P > .05$).

Global brain dynamics of AV integration

To investigate the neural underpinnings of AV *fusion* and *combination*, we recorded brain activity during perception of congruent and incongruent AV stimuli using magnetoencephalography (MEG). Participants watched videos showing a speaker pronouncing a syllable and reported which syllable they heard among 5 alternatives (i.e., 'aba', 'ada', 'aga', 'abga', 'agba' in the 'bdg' family, and 'apa', 'ata', 'aka', 'apka', 'akpa' in the 'ptk' family). Subject's responses were purposely delayed to avoid temporal overlap between perceptual/decisional processes and motor effects due to button press. Response times hence do not constitute relevant data here, and we only consider the report rates (Figure 4A), which were calculated for each condition including congruent responses (i.e., /ada/ or /ata/) in the *congruent* condition, fused responses (i.e., 'ada' or 'ata') in the *fusion* condition, and combined responses (either VA, i.e., 'abga' or 'apka', or AV, i.e., 'agba' or 'akpa') in the *combination* condition. To address whether the components of the AV integration brain network were primarily sensitive to asynchrony or to AV physical features (formants and lip motion), and how these two variables contribute to fusion versus combination, we also varied the delay between audio and visual syllables. We used 12 different stimulus onset asynchronies over a temporal window ranging from -120 ms audio lead to 320 ms audio lag (40 ms step), a window corresponding to a range in which fusion responses are expected to dominate over the auditory driven responses[4]; hence maximizing *fusion* reports (Figure 4A). In this task, the response of interest rate differed between conditions ($F(2,30) = 15.99$, $P < .001$), whatever the consonant family ($F(1,15) = 1.98$, $P = 0.16$), speaker gender ($F < 1$) or asynchrony ($F<1$) (see S3 Table for detailed statistics).

We first used dynamic source modelling of MEG data to explore the global dynamics of AV *fusion* and *combination* relative to the congruent condition (Figure 4B). We analysed the evoked activity in those six regions of interest that had the strongest incongruence effect (incongruent > congruent): namely the PAC (Primary Auditory Cortex), MT (Middle temporal visual area), LSTS (left Superior Temporal Sulcus), STG (Superior Temporal Gyrus), IFG (Inferior Frontal Gyrus) and ATC (Anterior Temporal Cortex). The contrasts *Fusion* vs. *congruent (Figure 4B, in blue)* and *combination* vs. *congruent* (Figure 4B, in red) revealed a common statistical effect in the LSTS (from ~100 ms pre-auditory stimulus onset), reflecting

that the LSTS detects very quickly AV incongruence, and signals it by stronger response. This was the only time point and location where fusion and combination were signalled by a similar response pattern. We then directly contrasted the two conditions to characterize the specific processes occurring in *fusion* and *combination* (Figure 4B, yellow lines). For *fusion*, increased activity in left temporal areas that started from the auditory onset (0 ms) and dropped at about 250 ms post-auditory stimulus onset. Interestingly, in all selected areas the activity pattern observed in the *combination* vs. *congruent* contrast and in the *combination* vs. *fusion* contrast were very similar, suggesting that *fusion* and *congruent* conditions possibly involve related processes (Figures 4B, yellow and red lines). When AV conflict resulted in *combination*, increased activity in the IFG (~100 ms pre-auditory stimulus onset) was followed by a sustained period of increased activity (80-450 ms) in middle temporal visual cortex (MT), and by a more transient increase in left temporal areas (~350 ms post-auditory stimulus onset). These data suggest that activity in the LSTS was delayed for *combination* relative to *fusion* (by about 200ms), a delay that is comparable to the delay observed at the perceptual level (estimated in both behavioural experiments to about 200ms) when integrating incongruent AV inputs. These data also show that the IFG is mostly involved in combination, suggesting a specific contribution of this area in combination.

Figure 4 *(legend on next page)*

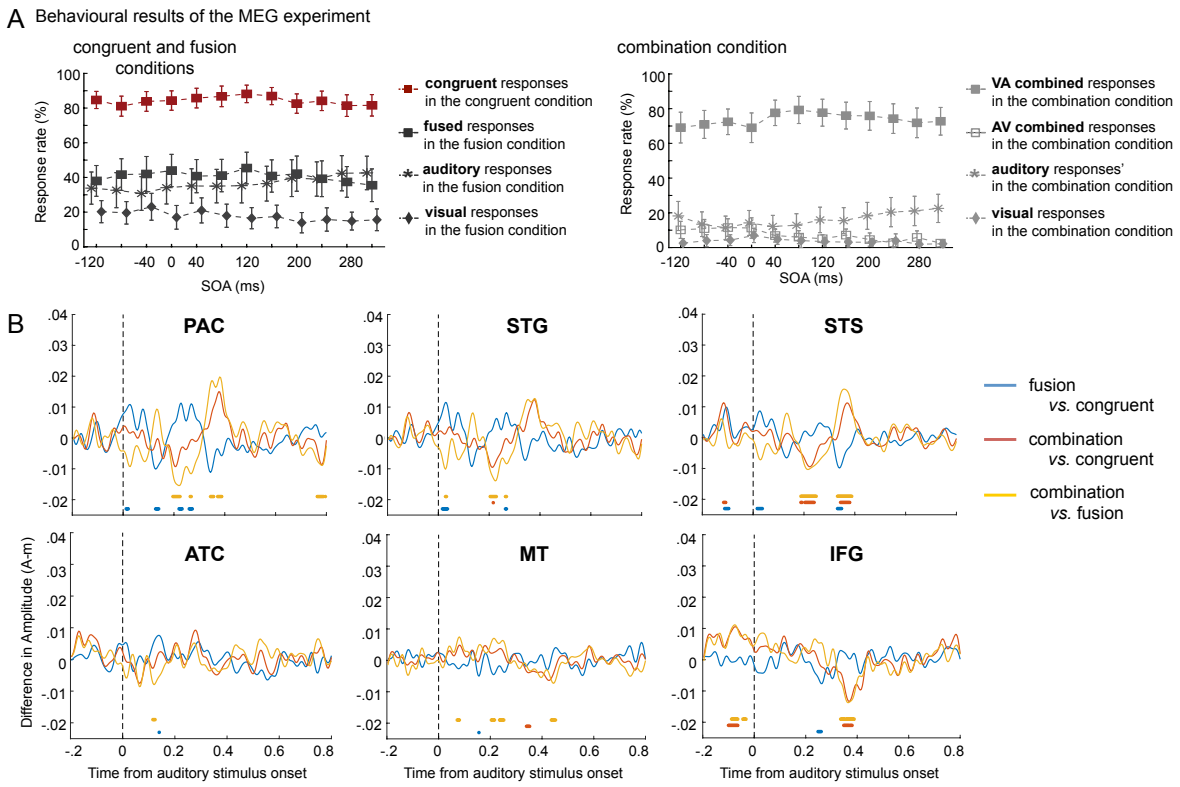

Figure 4. MEG experiment. (A). Behavioural results. Response Rate depending on the stimulus onset asynchrony (SOA) between visual and auditory stimuli. A negative stimulus onset asynchrony indicates that the auditory leads the visual input whereas a positive stimulus onset asynchrony shows that auditory lags visual. Error bars correspond to s.e.m. The left panel shows responses in *congruent* and *fusion* conditions: response of interest rates in *fusion* and *congruent* conditions (filled squares), auditory response rate in the *fusion* condition (dark grey stars), and visual response rate in the *fusion* condition (dark grey diamonds). The right panel shows responses in the *combination* condition only: response of interest rate, i.e., VA (visual-auditory) combined responses (light grey filled squares) and AV (audio-visual) combined responses (light grey squares) in the *combination* condition, auditory response rate in the *combination* condition (light grey stars), and visual response rate in the *combination* condition (light grey diamonds). (B) Differences in event-related activity between conditions, in each region of interest (fusion > congruent conditions in blue, combination > congruent conditions in red, combination > fusion conditions in yellow). Stars indicate significant Student *t-test values that* were estimated in each difference: fusion *vs.* congruent conditions in blue, combination *vs.* congruent conditions in red, combination *vs.* fusion conditions in yellow ($P < 0.05$, corrected for multiple comparisons using FDR).

Directional connectivity patterns for *Fusion* vs. *Combination*

The behavioural results and the dynamic source modelling MEG data both indicate that AV *combination* was more demanding than *fusion*, suggesting that combining discrepant stimuli might involve extra resources and perhaps a different neural network than fusing them. Based on previous modelling work we conjectured that the LSTS plays a pivotal function, notably that combination could be triggered by the impossibility to converge locally on a bimodal syllable solution. To further explore this hypothesis we first probed directional functional coupling across the 6 previously defined regions of interest (ATC, IFG, the LSTS, STG, MT, and PAC, see S1 Fig for a spatial location of the corresponding scouts) using dynamic causal modelling (DCM). This analysis was not time-resolved, thus only showed dominant connectivity patterns throughout the experimental trials. We found that *fusion* and *combination* had radically different neural dynamics, characterized by a dominant modulation of feed-forward and feedback connectivity from and to the LSTS, respectively (Figure 5). *Fusion* was associated with increased connectivity from the LSTS toward ATC and MT, and a decrease from MT to LSTS, consistent with the propagation of the fusion solution to higher-order regions and the update of visual motion representation as a function of the solution elaborated in the LSTS. In line with previous studies showing that visual speech prediction could directly influence the activity in auditory cortex[35,37], connectivity also increased from MT to PAC and decreased from PAC to ATC during *fusion.* In contrast, *combination* was associated with increased connectivity from IFG and PAC to LSTS. Finally, connectivity also increased from PAC to both STS and IFG, and decreased from LSTS to ATC, from ATC to IFG, from STG to PAC.

The functional connectivity results confirm the central role of the LSTS in both *fusion* and *combination*. They further show that in *fusion*, the LSTS dispatches information (presumably about the found syllable) to other brain regions for recognition and sensory representation updating, whereas in *combination*, the LSTS centralizes information from higher-order regions, possibly regions that are more sensitive to the precise timing of the AV events than the LSTS.

Figure 5

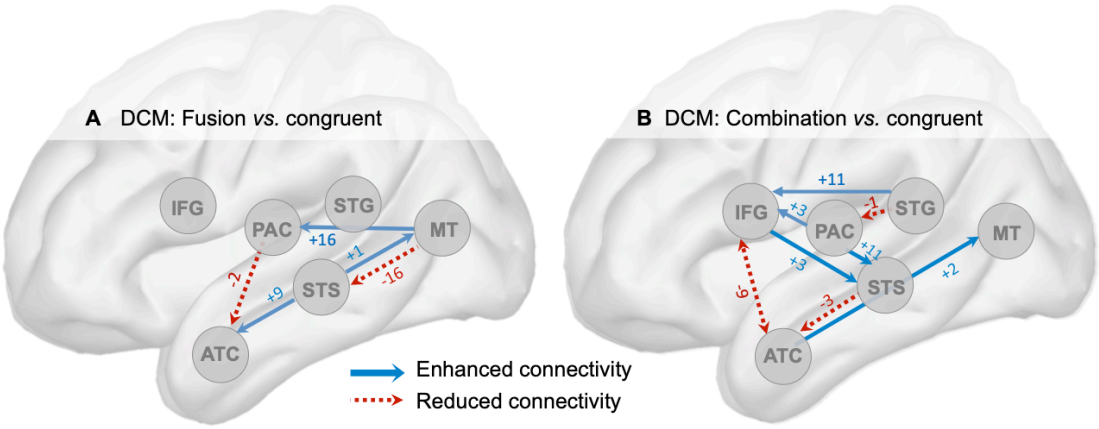

Figure 5. Dynamic causal modelling (DCM) of connectivity using event-related responses across the six main regions involved in audio-visual speech integration. The circles represent the sampled sources: primary auditory cortex (PAC), mediotemporal cortex (MT), the superior temporal gyrus (STG), the left superior temporal sulcus (LSTS), the inferior frontal gyrus (IFG) and the anterior temporal cortex (ATC). All connections and their values reflect enhanced or reduced connectivity of fusion (A) and combination (B) responses, relative to responses in the congruent condition. We tested the differences between conditions using Parametric Empirical Bayes (PEB) models. Rather than comparing different network architectures, we performed a post-hoc search by pruning away parameters that did not contribute to the model evidence ($p<0.05$). These results in a sparse graph where the connections shown are those that contributed significantly to the model evidence. Red dotted lines: reduced connectivity; Blue lines: enhanced connectivity.

Time-sensitivity and neural determinants of *Fusion* vs. *Combination*

To address time-sensitivity and enquire whether *fusion* or *combination* outcomes were signalled by one or more specific neuronal event(s), we used a general linear model (GLM) to regress auditory input-locked neuronal activity for each of the six regions of interest at each time point against three key trial-wise quantities: (i) *temporal asynchrony* values (from 0 ms to 320 ms) irrespective of whether auditory or visual signal came first, (ii) *physical AV incongruence* associated with the stimulus, i.e. congruent to incongruent lip motion/2nd formant patterns, and (iii) *perceptual output, i.e.* participant's responses (it had two levels; 'ada' or 'ata' for congruent and fusion conditions and 'abga', 'agba', 'apka', or 'akpa' for the combination condition). Importantly, the GLM was constructed using sequential orthogonalization to ensure that the regressor captured residual neuronal variance unaccounted for by the two other previous regressors.

Confirming recent literature, temporal asynchrony was reflected positively first in the IFG (20 ms pre- to 200 ms post-auditory stimulus onset), and then in PAC (around 300 ms post-auditory stimulus onset) (Figure 6). The sequence of positive coefficients in the IFG and then in PAC speaks to the role of the IFG in contextualising the sensory integration process[38,39].

Consistent with the DCM analysis, physical AV discrepancy was reflected in positive coefficients in the LSTS during a time period ranging from 150 ms to 50 ms pre-auditory stimulus onset. Auditory and visual inputs are asynchronous, and in our specific set of stimuli, the second input was easily predictable from the first one. Participants always received /ada/ + [ada] in the congruent condition, or /aba/ + [aga] in the fusion condition, or /aga/ + [aba] in the combination condition. This high predictability presumably explain cross-modal effects emerged very early in the LSTS (Figure 6 and S3 Fig), as previously shown in other studies[37]. Strong predictions likely resulted from a combination of overlearned AV associations and very short-term adaptations[40]. Sensitivity to physical AV incongruence was also observed within a broader network, including LSTS, PAC and the STG at later time points. A possible interpretation could be that prediction errors arising at auditory onset in the LSTS (the first region to detect AV physical features discrepancy) diffused to downstream regions.

Finally, the perceptual output was associated with positive coefficients (for combined percept) across the PAC, the STG, and the LSTS, all peaking around 400 ms post-auditory stimulus onset, showing that when the AV input did not match a known syllable (no possible *fusion*), the set of region previously involved in detecting AV physical discrepancy (0-100 ms) was reactivated for generating combined percepts. An even later *combination* effect occurred in the LSTS at ~600 ms, presumably signalling the elaboration of a double consonant combination by indexing two distinct positions within the multisensory feature space.

Negative coefficients, indicating positive association with responses of interest (i.e., 'ada' or 'ata') in *fusion* and *congruent* conditions (Figure 6, left panels), were visible in the IFG and ATC at ~400 ms post-auditory stimulus onset. These effects confirm quick stimulus recognition as observed for congruent stimuli (see Figure 4, patterns at 376ms). More surprisingly, however, *congruent* and *fusion* responses of interest were associated with an additional effect in MT (~150 ms).

Altogether, these results indicate that the LSTS is not involved in detecting AV asynchrony, but that it can predict auditory input on the basis of the visual stimulus and quickly detect AV discrepancy whatever the subsequent outcome (*fusion* or *combination*). AV physical incongruence is almost instantaneously registered by neighbouring left temporal regions, PAC and STG. In case of *fusion*, activity quickly drops in these regions and diffuses to higher-level areas (IFG and ATC) for conscious perception and response selection, as well as to MT where the lip-motion pattern is likely updated as a function of the solution onto which the LSTS has converged. In *combination*, activity in PAC, STG and LSTS is sustained until a complex (consonant sequence) solution is elaborated.

Figure 6 *(legend on next page)*

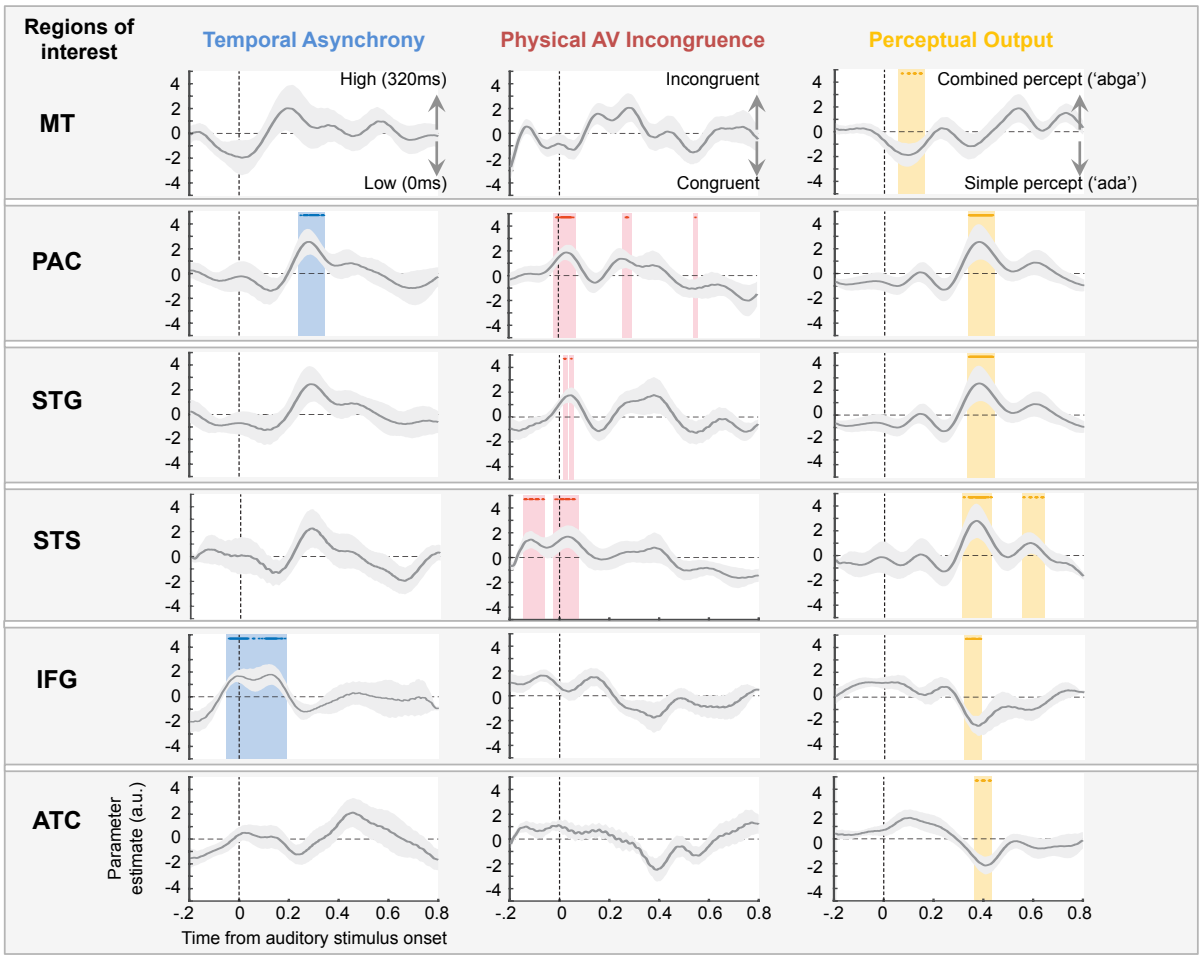

Figure 6. Results of the GLM analyses using Temporal Asynchrony (blue), *Physical AV Incongruence* (red), and Perceptual Output (yellow) showing the temporal dynamics of normalized beta in each area of interest. The correlation between the LSTS activity and the regressor *Physical AV Incongruence* emerges significantly before auditory onset (red) whereas a recursive activity in the LSTS is only visible for combination (yellow). Thick horizontal lines indicate time windows where parameter estimates diverge significantly from zero at a temporal cluster-wise corrected p-value of 0.05. The shaded error bounds indicate s.t.d. (light grey). The correlation between IFG and Temporal Asynchrony peaks around 0 ms, just before the correlation between MT and Temporal Asynchrony.
MT. mediotemporal area; PAC. primary auditory cortex; STG. superior temporal gyrus; LSTS. left superior temporal sulcus; IFG. inferior frontal gyrus; ATC. anterior temporal cortex.

Neural decoding of syllable identity for *Fusion* vs. *Combination*

Having established that the LSTS could quickly detect inconsistencies between auditory and visual physical features, and that the IFG was sensitive to AV asynchrony, we posited that the identity of *fusion* syllables (i.e., 'ada' or 'ata' from fusion condition) could be detected earlier in LSTS neural activity than that of *combination* syllables (i.e., 'abga', 'agba', 'apka' or 'akpa' from combination condition).

We hence directly probed whether neural activity expressed in the LSTS and IFG held reliable information about syllable identity using three decoding analyses on the trials with responses of interest, to classify (1) 'abga', 'agba', 'apka' and 'akpa' responses from combination condition vs. 'ada' and 'ata' responses from fusion condition, (2) 'abga', 'agba', 'apka' and 'akpa' responses from combination condition vs. 'ada' and 'ata' responses from congruent condition, and (3) 'ada' and 'ata' responses from fusion condition vs. 'ada' and 'ata' responses from congruent condition (Figure 7B and S4 Fig). To further probe that information propagation sequence, *time-resolved* decoding was performed. In line with previous findings[41], we observed that local evoked activity from one region was sufficiently discriminable to permit syllable categorization using a maximum correlation coefficient classifier (see Methods). We determined whether neural responses related to fusion or combination could be separated (Figure 7, left panel). As the two conditions use incongruent AV inputs, the decoding results only reflect the identity of the perceived syllable. We observed that the neural data could be clustered in the LSTS at two different time points (i.e., ~200 ms and ~400 ms), replicating part of the effect shown in the 2 other classifier analyses. Comparing this result with the classification of the responses of interest from fusion inputs vs. congruent inputs, we could determine that the neural activity peak at 200 ms reflected the identity of *fusion* syllables (Figure 7 right panel). Moreover, comparing with the classification of combination inputs vs. congruent inputs, we could determine that the neural activity peak at 400 ms reflected the identity of combined syllables (Figure 7 middle panel and S4 Fig). Interestingly, the combination percept could be also differentiated from congruent percept at the exact same time window (~400 ms) at 4 different locations (i.e., the IFG, LSTS, STG and PAC) (Figure 7 middle panel and S4 Fig), showing that the neural activity in the IFG also contains the identity of combined syllables.

These analyses confirmed two crucial points: (1) combination was the only AV integration

output that could be decoded from neuronal activity outside of the LSTS, confirming the implication of a broader network when combining discrepant AV inputs; (2) *fusion* output was only decodable from the LSTS activity before (about 200 ms) the combination responses, suggesting that the impossibility to fuse AV inputs in the LSTS is followed by a sequence of events leading to the combination process.

Figure 7

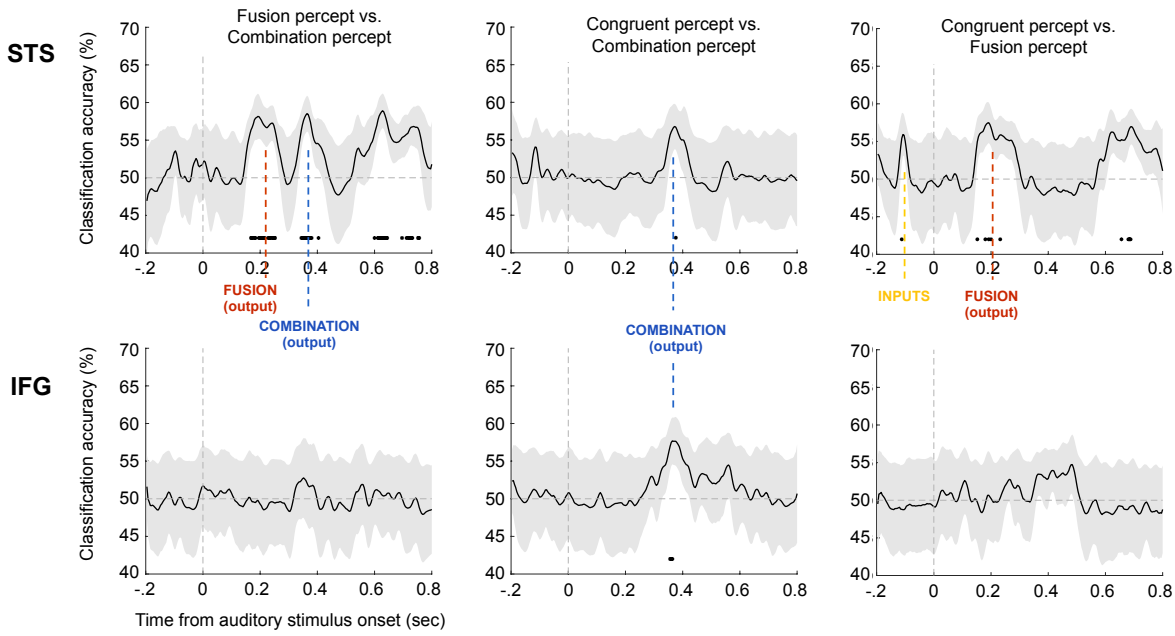

Figure 7. Decoding in the Left Superior Temporal Sulcus (LSTS) and the Inferior Frontal Gyrus (IFG). Time course of univariate classification (accuracy) for 'abga', 'agba', 'apka' and 'akpa' responses from combination vs. 'ada' and 'ata' responses from fusion (left panel), 'abga', 'agba', 'apka' and 'akpa' responses from combination vs. 'ada' and 'ata' responses from congruent (middle panel), and 'ada' and 'ata' responses from fusion vs. 'ada' and 'ata' responses from congruent (right panel). Left Panel. Univariate classification of responses of interest from *fusion and combination* conditions was possible in the LSTS at ~200 ms, ~400 ms and ~700 ms, but not in the IFG. Middle panel. Univariate classification between the responses of interest from congruent and combination was possible in the IFG and the LSTS at the same time point (~400 ms). Right panel. Univariate classification between responses of interest from *fusion and congruent* was possible in the LSTS at -100 ms, ~200 ms and ~700 ms, but not in the IFG.

Discussion

While the mechanisms leading to AV speech *fusion* are relatively well understood, those leading to AV stimulus *combination* are still unknown. Based on a previous computational model, we conjectured that AV *combination* follows from the difficulty to map the auditory and visual physical features in a multisensory space presumably located in the LSTS[28]. AV *combination* would hence result in a more demanding processing sequence than AV fusion, involving post-hoc temporal reordering of the auditory and visual input. The model was adapted to McGurk stimuli, and hence worked well with only lip and 2nd formant values (even though more features are presumably at play in AV speech integration). According to this simple 2-dimensional model, *fusion* occurs when the physical features of discordant AV stimuli fall in the vicinity of those corresponding to a lexically plausible and simple (single consonant transition) speech representation, whereas *combination* occurs when the physical features do not find AV matching features (Figure 1A). In this view, after *fusion* failure, *combination* demands additional processing, possibly by frontal areas, to covertly generate a plausible speech sound consistent with the AV input. The alternative scenario would be that *fusion* occurs when AV temporal asynchrony is non-detectable (e.g. because the visual stimulus is a weak predictor), whereas *combination* arises when A and V onsets can clearly be sequentially perceived (e.g. strong visual predictor). The two scenarios lead to distinct predictions regarding the delay with which a combination percept arises. In the first case, combining discrepant AV stimuli should take significantly longer than fusing them, i.e. the time needed for fusion failure and active generation of alternative operations such as ordering of A and V stimuli. In the second case, the time to fuse and combine AV discrepant stimuli should be about equal, because combination mostly depends on the on-line perception of the AV order (Figure 1B).

The results of both behavioural studies consistently show that combination was more time demanding than fusion (+200ms). The second behavioural study clarified that the effect was not imputable to added articulatory demands. When examining source-resolved MEG responses in the LSTS, i.e. the key region for integrating audio and visual speech information[15,42–44], we found that neural responses for combination were already delayed at the earliest stage of AV speech integration, and were hence unlikely to reflect mere

attention or effort mechanisms. The response delay for combination was hence partly associated with extra processing time in the LSTS.

To further understand the nature of the processing delay for AV *combination* relative to *fusion*, we explored how much the LSTS was sensitive to AV asynchrony, and whether other brain regions could be involved in on-line or retrospect ordering of A and V stimuli leading to *combination*. Our MEG experimental design hence involved different stimulus onset asynchronies. Although the behavioural effect of AV asynchrony was weak (Figure 4A), we found both left PAC and IFG to be sensitive to this parameter, to the point that a classification algorithm from the IFG activity could decode the contrast between combination and congruent percepts. This finding presumably reflects the implication of the IFG in perceptual speech tasks requiring precise temporal parsing or sequencing of speech signals[26,45,46].

Unlike the IFG, the LSTS was insensitive to asynchrony, but highly sensitive to the auditory and visual physical features incongruence[15,47], and in line with previous findings, it signalled very early whether the AV stimuli were discrepant, with a first effect driven by visual predictions (i.e., the auditory cortex response induced by visual input), and a second one following the auditory input driven by AV mismatch (Arnal et al., 2009).

Although congruent AV speech expected to produce faster and more accurate responses than mismatching AV speech[48,49], we did not confirm that AV *fusion* was more time demanding than responding to congruent AV syllables[2,10,50–54]. The difference between our results and previous ones is explained by the fact that contrary to previous studies[10,50–54], we only analysed trials where subjects effectively experienced *fusion* while discarding failed *fusion* trials (~45% of the trials).

In summary, our behavioural findings reveal longer delays for reporting AV *combination* than *congruent* and *fusion* percepts, a novel finding suggesting that reporting AV *combination* requires extra processing resources.

The role of the LSTS in the fusion/combination dynamic divergence

Although it is established that the LSTS integrates the information coming from A and V modalities, it is still not known whether it processes similarly or differently AV stimuli leading to *fusion* or *combination*. Using a GLM approach we found that the LSTS was the

first region to signal physical incongruence between the two sensory modalities. The incongruence effect was even anticipated by the LSTS when visual stimuli were strongly predictive of the audio (which in our specific experimental setting was the case for the McGurk fusion stimuli). A possible explanation for this anticipatory effect is that the LSTS quickly estimates whether to expect a precise auditory input (strong visual prediction) or rather a set of possible auditory inputs (weak visual prediction)[27]. When visual prediction is weak (e.g., visual /aga/), the LSTS could more easily fuse auditory and visual inputs (within ~100ms), whereas when visual prediction is strong (e.g., /aba/), perceived incongruence is potentially stronger, in some cases resulting in a combination percept. In other words, AV fusion could partly depend on the confidence associated with the expected input[34].

However, according to our predictive model of AV syllable integration[28], the most important factor determining *fusion* is whether the two stimuli meet close to an articulatory valid syllable representation within the 2nd acoustic formant/lip aperture 2D space. In the McGurk *fusion* case, visual /aga/ and auditory /aba/ fall in the vicinity of the 2D /ada/ representation, which quickly appears as a valid solution. This scenario is supported by the decoding analyse showing that neural activity in the LSTS signals the identity of fused syllables 200 ms before that of combined syllables. Crucially, by combining DCM and the "perceptual output" part of the GLM analyses, we show that the incongruence was both registered and solved extremely rapidly in the LSTS, and that the outcome was propagated forward (to the IFG and ATC). *Fusion* only differed from congruent processing in that the LSTS output also flowed backward to MT, presumably to update the visual motion model of the *fusion* syllable. The backward LSTS influence on lower sensory areas (MT)[55,56], provides an interesting illustration of how predictive coding could apply to AV integration[28,57,58].

In the combination case, since there is no 2D syllable representation at visual /aba/ and auditory /aga/ coordinates, the LSTS cannot readily converge on a viable articulatory solution. Activity in the LSTS rises 200 ms post-auditory stimulus onset like fusion but remains sustained until 400 ms; during this time lapse, the LSTS gets recursively involved in a more complex integration process involving PAC, STG and IFG. These findings suggest that *combination* required a tight coordination between the temporal network and the IFG[35], presumably to organize the temporal serialization of AV inputs.

The Inferior Frontal Gyrus tracks AV temporal asynchrony

Several previous studies have tied the IFG to AV *fusion* but its specific contribution is still unclear[15,26,59,60]. The current results suggest that the early IFG involvement in AV integration does not relate to feature identification[61], but are specific to AV timing. We found that the IFG activity tracked temporal AV asynchrony, at least within the range used in the experiment (from -120 ms auditory lead to 320 ms auditory lag). Interestingly, this sensitivity did not translate into a misalignment sensation, confirming that this range of AV asynchrony is perceptually well tolerated[4]. These findings are consistent with previous studies[23,24,62] showing that the IFG is involved in implicit timing[22], a sensitivity that allows listeners to predict word order from syntactic information[63]. Even though our brain is not aware of precise event temporality, the latter is essential to monitor e.g., articulatory and syntactic correctness, and to prepare for correct production.

Importantly, we observed that the temporal asynchrony effect in the IFG was followed by a similar effect in PAC, a finding that fits well with the observation that regularity in tone sequences modulates IFG activity, and that the estimated intervals propagate from the IFG to PAC[38,39]. Overall, our study confirm that the IFG originates descending signals about estimated precision or predictability of the sensory stimuli[64].

On-line AV temporal tracking, versus post-hoc temporal ordering in combination

When auditory and visual signals are well aligned, integrating AV syllables is an easy and relatively low-level process. Yet, when signals are asynchronous or physically discrepant, integrating them involves a more complex dynamics. Our findings show that combination results both from the identification of auditory and visual features in the LSTS, and from the detection of fine AV temporal asynchronies in the left IFG. The global information flow directionality (DCM) and the event sequence (GLM) indicate that *fusion* and *combination* arises from a visual pre-activation of the LSTS, which is either matched by a compatible audio (*fusion*) and further processed as "congruent" AV stimuli (*fusion*), or discarded by an incompatible one and further processed as temporally asynchronous AV sequence. Our results hence support that *combination* is triggered by a failure of AV fusion in the LSTS, and reflects *a posteriori* reconstruction of the most plausible AV sequence.

We found no behavioural or neurophysiological arguments for the alternative scenario in which temporal ordering in combination occurs on-line. The combination of visual [aba] and auditory /aga/ can take two forms[31]: one ('abga', 72% of the responses) that respects the real AV temporal order, and another one ('agba', 8% of the responses) that does not (for similar results on the two combination percepts formed with /aba/ and [ada], see[32]). The production of an "erroneous" /abga/ percept strongly supports a post-hoc process resulting from recursive activity across the LSTS, PAC, STG and IFG, rather than from on-line sequencing, which offers less possibilities of mixing the AV stimuli.

Conclusion

The present findings contribute to delineate the hitherto unknown mechanisms of AV *combination.* By showing that *combination* percepts arise from a two-step process consisting in first registering AV physical incompatibility in the LSTS, and then reordering the stimuli by AV asynchrony sensitive regions, these results illustrate the primordial role of recursive/predictive mechanisms in AV speech integration, and unravel the dynamics leading to the elaboration of novel constructs that best explain the cause of AV stimuli.

Methods

Subjects. Twenty healthy subjects participated in the first behavioural experiment (9 males - age range: 20-28 years), 16 subjects in the second behavioural experiment (9 males – age range: 21-31 years), and 15 took part in the MEG study (10 males – age range: 21-24 years). All participants were right-handed, French-native speakers, and had no history of auditory or language disorders. Each behavioural experiment consisted of 1 hour-long session performed in a quiet room while the MEG experiment consisted of 2 sessions lasting 2 hours each. All participants were paid for their participation. Ethics permission was granted by the University Hospital of Geneva in Switzerland for the behavioural experiments (CEREH 13-117), and by the Inserm ethics committee in France (biomedical protocol C07-28) for the MEG experiment. All participants provided written informed consent prior to the experiment.

Stimuli. We recorded natural speech consisting of a man's and a woman's face articulating syllables. These two native French-speaker pronounced the syllables /apa/, /ata/, and /aka/ or the syllables /aba/, /ada/, and /aga/. The two syllable continua vary according to the place of articulation; furthermore syllables are voiceless in one continuum, i.e., /apa/, /ata/ and /aka/, and voiced in the other continuum, i.e., /aba/, /ada/, and /aga/. To preserve the natural variability of speech, we used 10 exemplars of each syllable pronounced. Movies were recorded in a soundproof room into a 720 x 480-pixel movie with a digitization rate of 29.97 frames per s (1 frame = 33.33 ms). Stereo soundtracks were digitized at 44.1 kHz with 16-bits resolution.

We created 3 movie categories, which corresponded to the 3 stimulation conditions. *Congruent* videos corresponded to the initial recorded movie of the syllables /ada/ or /ata/. All videos had the same length and lasted 1000 ms. Using the soundtrack, we homogenized the duration of the stimuli: the vocal burst of the first /a/ and the consonantal burst were aligned across videos. The length of the second vocalic part was slightly variable across stimuli. Incongruent *fusion* pairs were created by dubbing an audio /apa/ or /aba/ onto a video [aka] or [aga], respectively. Audio and video were merged from the same speaker. The new soundtrack (/apa/ or /aba/), was systematically aligned to the initial

soundtrack (/aka/ or /aga/) based on the vocalic burst of the first /a/ and on the consonantal burst. Incongruent *combination* pairs were created by dubbing an audio /aka/ or /aga/ onto respectively a video [apa] or [aba], using the same alignment procedure. Auditory and Visual parameters of each condition are shown in Figure 1 and in Table 1.

Tasks design. Auditory-visual stimuli were presented using Psychophysics-3 Toolbox and additional custom scripts written for Matlab (The Mathworks, Natick, Massachussetts, version 8.2.0.701). Sounds were presented binaurally at a sampling rate of 44100 Hz and at an auditory level individually set before the task via earphones using an adaptive staircase procedure. For each participant, we determined prior the experiments their auditory perceptual threshold corresponding to 80% categorization accuracy. The estimated sound level was used to transmit the stimuli (mean 30 dB sensation level) during the behavioral experiments (Experiment 1 and experiment 2) and MEG experiment.

Experiment 1, Repetition task. Participants were individually tested and were instructed to watch each movie and repeat what they heard as fast as possible. We used the same instruction provided by McGurk and MacDonald (1976). Participants were asked to repeat as fast as they can what they heard. We did not limit the possible answers to a limited set of syllables. Nevertheless, note that for the three conditions (*congruent*, *fusion* and *combination*), different responses were expected according to our hypotheses (see S1 Fig).

Experiment 2, Pairing task. Participants were individually tested and were instructed to read the syllable written on the screen, then to watch the movie and to press the space bar as fast as possible when the written syllable matched what they heard. In the b-d-g sessions, participants could read 'aba', 'ada', 'aga', 'abga' or 'agba'. In the p-t-k sessions, participants could read 'apa', 'ata', 'aka', 'apka' or 'akpa'.

In the two behavioural experiments, participants were presented with four blocks, each one containing one speaker gender (female or male voice) and one continuum (b-d-g or p-t-k). Each block presented 90 AV stimuli corresponding to 30 *congruent* stimuli (A[d]V[d] or A[t]V[t]), 30 *fusion* stimuli (A[b]V[g] or A[p]V[k]), and 30 *combination* stimuli (A[g]V[b] or A[k]V[p]), for a total of 360 stimuli per subject. Trials were randomly presented in each block, and blocks were randomly presented across participants.

Participants were sat 1 m from the monitor, and videos were displayed centered on a 17-inch Apple MacBookPro laptop on a black background. Sounds were presented through earphones (sennheiser CX 275).

MEG experiment. Each continuum (/bdg/ and /ptk/) was delivered to participants in two independent sessions of 360 trials each. Participants were asked to perform an identification task. Each trial comprised one video (randomly chosen among the 3 conditions), followed by a 1s silent gap; then, a response screen with 'aba', 'ada', 'aga', 'abga' and 'agba' in the b-d-g sessions, and 'apa', 'ata', 'aka', 'apka' and 'akpa' in the p-t-k sessions, were displayed. Syllables were randomly displayed from right to left on the screen to prevent motor preparation and perseverative responses. During MEG recording, the appearance of the response screen was randomly jittered 100, 300 or 500ms after the silent gap. Participants indicated their response by moving a cursor under the syllables and pressing a key to select the chosen syllable as quickly as possible. Subject's responses were purposely delayed to avoid temporal overlap between perceptual processes and motor effects due to button press. Response times hence do not constitute relevant data. To limit eye movements, subjects were asked to blink only after giving their motor response. After the response, a jittered delay varying from 3 to 5 s led to the next trial.

MEG recording and preprocessing. Brain signals were recorded using Neuromag Elekta with a total of 306 channels composed of 204 axial gradiometers and 102 magnetometers. Recordings were first preprocessed using signal-space separation through Neuromag software MaxFilter. This allows removing signal coming from outside of the electrode sphere which allows removal of EOG (electroocculography) and ECG (electrocardiography) interference among other sources of noise. Originally, signals were sampled at a rate of 1000Hz and were re-sampled at 250Hz for further preprocessing stages. Before MEG recording, headshape was acquired for each participant using Polhemus. After the MEG session, an individual anatomical MRI was recorded (Tim-Trio, Siemens; 9 min anatomical T1-weighted MP-RAGE, 176 slices, field of view = 256, voxel size = 1 x 1 x 1 mm^3). MEG data were preprocessed, analyzed and visualized using dataHandler software (wiki.cenir.org/doku.php), the Brainstorm toolbox [65] and custom Matlab scripts.

Analysis.

Experiment 1, Repetition task. We recorded the participant's vocal response using a microphone. No feedback was provided after each response. The response time was measured as the interval between video onset and start of the syllable repetition from the audio recording on each trial. We also assessed the identification choice made by participants, i.e., the syllable repeated, on each trial.

Experiment 2, Pairing task. The number of matching between the written syllable and the video that led to a response of interest served as the measure of syllable identification. The response time was measured as the interval between video onset and the button press on each trial.

In the two behavioural experiments, percentage of responses of interest and response onset latency were calculated for each condition. Percentage of responses of interest ([ada-ata] for *congruent* and *fusion* conditions, or [abga-apka-agba-akpa] for the *combination* condition) was averaged separately across Consonant Family ('bdg' and 'ptk'), Speaker Gender (male and female), and Conditions (*congruent*, *combination* and *fusion*) factors. Response onset latency was calculated and averaged based on responses of interest across each condition). We reported the percentage of congruent responses in the *congruent* condition (i.e. /ada/ or /ata/ responses), the percentage of visual (i.e., /aga/ or /aka/), auditory (i.e., /aba/ or /apa/) and fused (i.e., /ada/ or /ata/) responses in the *fusion* condition, the percentage of visual (i.e., /aba/ or /apa/), auditory (i.e., /aga/ or /aka/), VA combined (i.e., /abga/ or /apka/) and AV combined (i.e., /agba/ or /akpa/) responses in the *combination* condition.

Behavioural analyses: Analysis of variance. Percentage of responses was analysed within each experiment (experiments 1 and 2) using a 3 X 2 repeated-measures ANOVAs with Conditions (*congruent*, *fusion*, *combination*) and responses of interest ([ada-ata] responses for the *congruent* and *fusion* conditions, and [abga-apka-agba-akpa] responses for the *combination* condition) as within-subjects factors. For the mean response latency, we measured the interval between video and vocal response onsets, for each type of response of interest. A 3 X 1 repeated measures ANOVA was performed on response times (RTs) with Conditions (*congruent*, *combination* and *fusion*) as a within-subjects factor. All

ANOVAs modelled the variables Speaker Gender (female and male), and Consonant Family ('bdg' and 'ptk') as fixed-factors so as to generalize the results obtained to each speaker and each consonant family tested.

MEG processing: Using structural data, brain models for each subject were build using Brain Visa Software [66]. Individual brain models were mapped to the ICBM-112 brain model template for group-level analysis. Data analysis was performed with Brainstorm [65], which is documented and freely available for download online under the GNU general public license.

The data considered (trials) started 0.2s before the auditory event and until 0.8s after the auditory input (i.e., the first vowel /a/ in /aXa/). We only analysed trials without eye artefacts or jumps in the signal. We did not conduct the analyses at sensor level as the contribution from different sources is mixed and our aim was to describe the brain network involved in AV speech perception. We computed forward models using overlapping-sphere method, and source imaging using weighted minimum norm estimates (wMNEs) onto preprocessed data, all with using default Brainstorm parameters. The wMNEs included an empirical estimate of the variance of the noise at each MEG sensor, which brings both magnetometers and gradiometers into the same basic range of units, allowing the source estimation to be proceed with a combined array of 306 sensors (204 planar gradiometers and 102 magnetometers). We also downsampled the MEG data to 250 Hz.

Regions of Interest (ROI). We defined six regions of interest in the left hemisphere, using functional localizers at a group-level (S1 Fig). We selected all the regions that had the largest M100 (~110 ms) auditory evoked responses in all conditions across subjects. ROI analyses were carried out by performing ANOVA across subjects in order to ensure that the selected regions are the ones where more activity was found independently of the conditions.

Evoked Responses. Signals from each region of interest were extracted and analysed. Evoked responses were computed by averaging MEG signals after source reconstruction across trials for each time sample around stimuli, for each subject and each condition (i.e., fusion, combination and congruent) (S2 Fig). We then contrasted the conditions by subtracting their respective ERP response, which allowed testing 3 contrasts: fusion minus

congruent, combination minus congruent, and combination minus fusion. Differences in the ERP response were detected across conditions by performing *t* tests against 0. FDR corrections for multiple comparisons were applied over the dimensions of interest (i.e., time samples, regions-of-interest and conditions), using the Benjamini–Hochberg procedure.

Dynamic Causal Modelling (DCM) analysis procedure. Different Analysis of functional connectivity through Dynamic Causal Modeling was performed to determine significant changes in connectivity strength across conditions. No prior information was included in the model specification, which implies that connections between every pair of nodes were established. Significance test for connectivity strength changes were implemented using Parametrical Empirical Bayes method through SPM, creating nested models with different connections switched off each time, and comparing the model evidence obtained. This allows determining which connections affect significantly the predictive power of the model.

General Linear Model procedure. We then created one GLM and applied this GLM to each time point and each ROI separately. This analysis was performed on single-trial event-related activity, on the trials with responses of interest. The GLM included the following parametric modulators: (i) the *temporal asynchrony* values that vary from 0 ms to 320 ms, without taking into account whether the auditory or the visual signal came first, (ii) the *physical AV incongruence* associated with the stimulus that is either congruent or incongruent, and (iii) the *output choice* associated with each participant's response, which indicates whether the response given by the participant corresponded to a simple syllabic percept (i.e., [ada-ata] for *congruent* and *fusion* conditions) or a combined syllabic percept (i.e., [abga-apka-agba-akpa] for the *combination* condition) (note that the response of noninterest are coded with the value NaN). We regressed single-trial MEG signals against these 3 parametric quantities at successive time points from -200 ms to 800 ms following auditory stimulus onset. Obtained time courses for parametric modulators in the GLM were smoothed using bandpass filtering (1-40 Hz) and then averaged across subjects. We next determined the time window where parametric modulators for temporal asynchrony, p*hysical AV incongruence* and output choice value were significantly different from zero. FDR corrections for multiple comparisons were applied over the dimensions of interest (i.e.,

time samples, regions-of-interest and number of regressors), using the Benjamini–Hochberg step-up procedure.

Classification of syllable identity. Decoding analyses were performed with the Neural Decoding Toolbox[67], using a maximum correlation coefficient classifier on evoked responses in each region of interest. Analysis was constrained to trials with the responses of interest. Three different pattern classifiers were built: one classifier was used to detect the neural activity capable of distinguishing between the response of interest in the fusion (i.e., 'ada' and 'ata' responses) and combination conditions (i.e., 'abga', 'agba', 'akpa' and 'apka' responses). As the two conditions use incongruent AV inputs, the decoding results reflect only the identity of the outputs. Another classifier was used to detect where neural activity allowed the distinction between the response of interest in the combination (i.e., 'abga', 'agba', 'akpa' and 'apka' responses) and congruent conditions (i.e., 'ada' and 'ata' responses). A third classifier was used to detect where neural activity differed between an 'ada' percept from *fusion* stimuli compared to an 'ada' percept from *congruent* stimuli. Comparing the results of the classifier between fusion and combination with the two other classifiers (i.e., fusion vs. congruent and combination vs. congruent) allowed us to assess profile similarity between the classifiers. By matching the curves of the classifiers, we can define the elements that correspond to the output of the fusion vs. the output of the combination.

In the decoding procedure, each classifier was trained to associate MEG data patterns with corresponding stimulus conditions (for each trial, the identity of the syllable perceived). The amount of relevant information in the MEG signal was evaluated by testing the accuracy of the classifier on a separate set of test data. We performed the analyses at each time point, within 1-ms non-overlapping bins.

Decoding analyses were performed on each ROI with a cross-validation procedure where the classifier is trained on a subset of the data, and then the classifier's performance is evaluated on the held-out test data. For each decoding run, data from the selected trials were divided into sets of 10 trials, and the data from each set of 10 trials were averaged together (see[68] for a similar procedure). Each decoding run was performed at the group level, pooling all subjects together (see[69] for a similar procedure). For example, in the first decoding procedure, a pattern classifier was trained to associate MEG patterns with the

participants' responses (the identified syllable, i.e., 'ada' vs. 'abga') in fusion and combination conditions.

For each decoding analysis, the pattern classifier was trained on the participant's response. It computed the correlation between MEG data and the syllable identified at each time point, and was trained on 80% of the data, while its performance was assessed on the withheld 20% of the test data. The splitting procedure between training and test data was performed 100 times to reduce the variance in the performance estimate. The classifier computed the correlation between test vectors (i.e., randomly selected mean values of 10 trials in the ROI at each time point) and a vector created from the mean of the training vectors. Each test point took the label of the class of the training data with which it maximally correlated. The reported final classification accuracy is reported as the percentage of correct trials classified in the test set averaged over all cross-validation splits. We then assessed the time window where decision values between the two categories were significantly different from zero (*t* test against zero). FDR corrections for multiple comparisons were applied over the dimensions of interest (i.e., time samples, regions-of-interest and number of classifiers), using the Benjamini–Hochberg step-up procedure.

Acknowledgments

We thank Jean-Luc Schwartz for support and inspiration. We are grateful to Jean-Luc Schwartz, Alexis Hervais-Adelman, and Valérian Chambon for comments and useful discussions about earlier versions of this manuscript, and Christophe Savariaux for technical support of the video editing. This work was funded by the Swiss National Science Foundation (SNF 320030_149319 and SNF 320030_163040 to A-L.G. and SNF P300P1_167591 to S.B.), and by the Fondation pour l'Audition (RD-2016-5 to S.B.).

Author contributions

Conceived and designed the experiments: S.B., I.O., A-L.G.; performed the experiments: S.B.; analysed the data: S.B., J.D., I.O., A-L.G.; wrote the paper: S.B., J.D., I.O., A-L.G.

Competing interests

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264**, 746–8 (1976).
2. Alsius, A., Paré, M. & Munhall, K. G. Forty Years after Hearing Lips and Seeing Voices: The McGurk Effect Revisited. *Multisens. Res.* **31**, 111–144 (2017).
3. Matchin, W., Groulx, K. & Hickok, G. Audiovisual Speech Integration Does Not Rely on the Motor System: Evidence from Articulatory Suppression, the McGurk Effect, and fMRI. *J. Cogn. Neurosci.* **26**, 606–620 (2014).
4. van Wassenhove, V., Grant, K. W. & Poeppel, D. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* **45**, 598–607 (2007).
5. Baart, M., Lindborg, A. & Andersen, T. S. Electrophysiological evidence for differences between fusion and combination illusions in audiovisual speech perception. *Eur. J. Neurosci.* **46**, 2578–2583 (2017).
6. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
7. Bernstein, L. E. & Liebenthal, E. Neural pathways for visual speech perception. *Front. Neurosci.* **8**, 386 (2014).
8. Flinker, A. *et al.* Redefining the role of Broca's area in speech. *Proc. Natl. Acad. Sci.* **112**, 2871–2875 (2015).
9. Miozzo, M., Williams, A. C., McKhann, G. M. & Hamberger, M. J. Topographical gradients of semantics and phonology revealed by temporal lobe stimulation. *Hum. Brain Mapp.* **38**, 688–703 (2017).
10. Beauchamp, M. S., Nath, A. R. & Pasalar, S. fMRI-Guided Transcranial Magnetic Stimulation Reveals That the Superior Temporal Sulcus Is a Cortical Locus of the McGurk Effect. *J. Neurosci.* **30**, 2414–2417 (2010).
11. Szycik, G. R., Stadler, J., Tempelmann, C. & Münte, T. F. Examining the McGurk illusion using high-field 7 Tesla functional MRI. *Front. Hum. Neurosci.* **6**, 95 (2012).
12. Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H. & Martin, A. Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nat. Neurosci.* **7**, 1190–1192 (2004).
13. Nath, A. R. & Beauchamp, M. S. A neural basis for interindividual differences in the

McGurk effect, a multisensory speech illusion. *Neuroimage* **59**, 781–787 (2012).

14. Venezia, J. H. *et al.* Auditory, Visual and Audiovisual Speech Processing Streams in Superior Temporal Sulcus. *Front. Hum. Neurosci.* **11**, 174 (2017).
15. Hickok, G. *et al.* Neural Networks Supporting Audiovisual Integration for Speech: A Large-Scale Lesion Study. *Cortex* **103**, 360–371 (2018).
16. Bhat, J., Miller, L. M., Pitt, M. A. & Shahin, A. J. Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. *J. Neurophysiol.* **113**, 1437–1450 (2015).
17. Schwartz, J.-L. & Savariaux, C. No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLoS Comput. Biol.* **10**, e1003743 (2014).
18. Macaluso, E., George, N., Dolan, R., Spence, C. & Driver, J. Spatial and temporal factors during processing of audiovisual speech: A PET study. *Neuroimage* **21**, 725–732 (2004).
19. Olson, I. R., Gatenby, J. C. & Gore, J. C. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cogn. Brain Res.* **14**, 129–138 (2002).
20. Simon, D. M., Nidiffer, A. R. & Wallace, M. T. Single Trial Plasticity in Evidence Accumulation Underlies Rapid Recalibration to Asynchronous Audiovisual Speech. *Sci. Rep.* **8**, 12499 (2018).
21. Foss-feig, J. H. *et al.* An extended multisensory temporal binding window in autism spectrum disorders. *Exp. brain Res.* **203**, 381–389 (2010).
22. Coull, J. T. & Nobre, A. C. Dissociating explicit timing from temporal expectation with fMRI. *Curr. Opin. Neurobiol.* **18**, 137–144 (2008).
23. Hironaga, N. *et al.* Spatiotemporal brain dynamics of auditory temporal assimilation. *Sci. Rep.* **7**, 4–9 (2017).
24. Baumann, O. *et al.* Neural Correlates of Temporal Complexity and Synchrony during Audiovisual Correspondence Detection. *Eneuro* **5**, a0294-17.2018 (2018).
25. Hagoort, P. Nodes and networks in the neural architecture for language: Broca's region and beyond. *Curr. Opin. Neurobiol.* **28**, 136–141 (2014).
26. Miller, L. M. & D'Esposito, M. Perceptual Fusion and Stimulus Coincidence in the

Cross-Modal Integration of Speech. *J. Neurosci.* **25**, 5884–5893 (2005).

27. Arnal, L. H., Wyart, V. & Giraud, A.-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* **14**, 797–801 (2011).
28. Olasagasti, I., Bouton, S. & Giraud, A.-L. Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex* **68**, 61–75 (2015).
29. Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* **15**, 839–843 (2005).
30. Colin, C., Radeau, M., Deltenre, P., Demolin, D. & Soquet, A. The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *Eur. J. Cogn. Psychol.* **14**, 475–491 (2002).
31. Cathiard, M.-A., Schwartz, J.-L. & Abry, C. Asking a naive question to the McGurk effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]? in *AVSP 2001 International Conference on Auditory-Visual Speech Processing* 138–142 (2001).
32. Soto-Faraco, S. & Alsius, A. Deconstructing the McGurk-MacDonald Illusion. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 580–587 (2009).
33. Morís Fernández, L., Torralba, M. & Soto-Faraco, S. Theta oscillations reflect conflict processing in the perception of the McGurk illusion. *Eur. J. Neurosci.* 1–12 (2018). doi:10.1111/ejn.13804
34. Giordano, B. L. *et al.* Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *Elife* **6**, e24763 (2017).
35. Park, H., Kayser, C., Thut, G. & Gross, J. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* **5**, e14521 (2016).
36. Kayser, S. J., Ince, R. A. A., Gross, J. & Kayser, C. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J. Neurosci.* **35**, 14691–14701 (2015).
37. Arnal, L. H., Morillon, B., Kell, C. a & Giraud, A.-L. Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* **29**, 13445–53 (2009).
38. Auksztulewicz, R. *et al.* The Cumulative Effects of Predictability on Synaptic Gain in the Auditory Processing Stream. *J. Neurosci.* **37**, 6751–6760 (2017).

39. Phillips, H. N. *et al.* Convergent evidence for hierarchical prediction networks from human electrocorticography and magnetoencephalography. *Cortex* **82**, 192–205 (2016).

40. Olasagasti, I. & Giraud, A.-L. Integrating prediction errors at two time scales permits rapid recalibration of speech sound categories. *BioRxiv* 1–19 (2018).

41. Kocagoncu, E., Clarke, A., Devereux, B. J. & Tyler, L. K. Decoding the Cortical Dynamics of Sound-Meaning Mapping. *J. Neurosci.* **37**, 1312–1319 (2017).

42. Zhu, L. L. & Beauchamp, M. S. Mouth and Voice : A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *J. Neurosci.* **37**, 2697–2708 (2017).

43. Stein, B. E. & Stanford, T. R. Multisensory integration : current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* **9**, 255–66 (2008).

44. Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W. & Perrett, D. I. Integration of Visual and Auditory Information by Superior Temporal Sulcus Neurons Responsive to the Sight of Actions. *J. Cogn. Neurosci.* **17**, 377–391 (2005).

45. Willems, R. M., Özyürek, A. & Hagoort, P. NeuroImage Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *Neuroimage* **47**, 1992–2004 (2009).

46. Özyürek, A. Hearing and seeing meaning in speech and gesture : insights from brain and behaviour. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130296 (2014).

47. Pratt, H., Bleich, N. & Mittelman, N. Spatio-temporal distribution of brain activity associated with audio-visually congruent and incongruent speech and the McGurk Effect. *Brain Behav.* **5**, 1–25 (2015).

48. Klucharev, V., Möttönen, R. & Sams, M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cogn. Brain Res.* **18**, 65–75 (2003).

49. Adank, P., Nuttall, H., Bekkering, H. & Maegherman, G. Effects of stimulus response compatibility on covert imitation of vowels. *Attention, Perception, Psychophys.* 1–10 (2018). doi:10.3758/s13414-018-1501-3

50. Green, K. P. & Kuhl, P. K. The interaction of visual place and auditory voicing information during phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**,

278–288 (1991).

51. Hessler, D., Jonkers, R., Stowe, L. & Bastiaanse, R. The whole is more than the sum of its parts - Audiovisual processing of phonemes investigated with ERPs. *Brain Lang.* **124**, 213–224 (2013).

52. Keane, B. P., Rosenthal, O., Chun, N. H. & Shams, L. Audiovisual integration in high functioning adults with autism. *Res. Autism Spectr. Disord.* **4**, 276–289 (2010).

53. Nahorna, O., Berthommier, F. & Schwartz, J.-L. Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* **132**, 1061–1077 (2012).

54. Norrix, L. W., Plante, E. & Vance, R. Auditory-visual speech integration by adults with and without language-learning disabilities. *J. Commun. Disord.* **39**, 22–36 (2006).

55. Friston, K. Does predictive coding have a future? *Nat. Neurosci.* **21**, 1019–1021 (2018).

56. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

57. Lüttke, C. S., Ekman, M., van Gerven, M. A. J. & de Lange, F. P. McGurk illusion recalibrates subsequent auditory perception. *Sci. Rep.* **6**, 32891 (2016).

58. Lüttke, C. S., Pérez-Bellido, A. & de Lange, F. P. Rapid recalibration of speech perception after experiencing the McGurk illusion. *R. Soc. Open Sci.* **5**, (2018).

59. Hertrich, I., Dietrich, S. & Ackermann, H. Cross-modal interactions during perception of audiovisual speech and nonspeech signals: An fMRI study. *J. Cogn. Neurosci.* **23**, 221–237 (2011).

60. Hasson, U., Skipper, J. I., Nusbaum, H. C. & Small, S. L. Abstract Coding of Audiovisual Speech: Beyond Sensory Representation. *Neuron* **56**, 1116–1126 (2007).

61. Murakami, T. *et al.* The motor network reduces multisensory illusory perception. *J. Neurosci.* **38**, 9679–9688 (2018).

62. Gau, R. & Noppeney, U. How prior expectations shape multisensory perception. *Neuroimage* **124**, 876–886 (2016).

63. Kristensen, L. B., Engberg-Pedersen, E. & Wallentin, M. Context Predicts Word Order Processing in Broca's Region. *J. Cogn. Neurosci.* **26**, 2762–2777 (2016).

64. Di Liberto, G. M., Lalor, E. C. & Millman, R. E. Causal cortical dynamics of a

predictive enhancement of speech intelligibility. *Neuroimage* **166**, 247–258 (2018).

65. Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. Brainstorm: A user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* 879716 (2011). doi:10.1155/2011/879716

66. Cointepas, Y., Geffroy, D., Souedet, N. & Denghien, I. The BrainVISA project: a shared software development infrastructure for biomedical imaging research. in *Proceedings 16th HBM* (2010).

67. Meyers, E. M. The Neural Decoding Toolbox. *Front. Neuroinform.* **7**, 8 (2013).

68. Isik, L., Meyers, E. M., Leibo, J. Z. & Poggio, T. The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* **111**, 91–102 (2014).

69. Boran, E. *et al.* Persistent hippocampal neural firing and hippocampal-cortical coupling predict verbal working memory load. *Sci. Adv.* **5**, eaav3687 (2019).