Different Neural Networks Are Involved in Audiovisual Speech Perception Depending on the Context

Nicole Malfait[1], Pierre Fonlupt[2,3], Laurie Centelles[1], Bruno Nazarian[1], Liana E. Brown[4], and Anne Caclin[2,3]

Abstract

■ How are we able to easily and accurately recognize speech sounds despite the lack of acoustic invariance? One proposed solution is the existence of a neural representation of speech syllable perception that transcends its sensory properties. In the present fMRI study, we used two different audiovisual speech contexts both intended to identify brain areas whose levels of activation would be conditioned by the speech percept independent from its sensory source information. We exploited McGurk audiovisual fusion to obtain short oddball sequences of syllables that were either (a) acoustically different but perceived as similar or (b) acoustically identical but perceived as different. We reasoned that, if there is a single network of brain areas representing abstract speech perception, this network would show a reduction of activity when presented with syllables that are acoustically different but perceived as similar and an increase in activity when presented with syllables that are acoustically similar but perceived as distinct. Consistent with the long-standing idea that speech production areas may be involved in speech perception, we found that frontal areas were part of the neural network that showed reduced activity for sequences of perceptually similar syllables. Another network was revealed, however, when focusing on areas that exhibited increased activity for perceptually different but acoustically identical syllables. This alternative network included auditory areas but no left frontal activations. In addition, our findings point to the importance of subcortical structures much less often considered when addressing issues pertaining to perceptual representations. ■

INTRODUCTION

How are we able to easily and accurately recognize speech sounds when the relationship between acoustic signals and perceived phonemes is so variable? For instance, because of coarticulation, the acoustic information for a given phoneme such as /d/ is highly variable depending on the vowel context. It has been proposed that the "lack of invariance problem" becomes tractable when contextual information that naturally accompanies speech is taken into account, observable facial gestures, for instance (e.g., Skipper, Nusbaum, & Small, 2006).

More than half a century ago, Sumby and Pollack (1954) demonstrated that under acoustically noisy conditions adding visible facial movements congruent with the acoustic signal enhances speech recognition. Symmetrically, recognition of perfectly audible speech is impaired when dubbed onto visible facial movements that are incongruent with the acoustic signal (Dodd, 1977). A classic illustration of this impairment is the fusion McGurk effect, in which an auditory /ba/ dubbed onto a visual /ga/ is heard as /da/ (McGurk & MacDonald, 1976). Many studies have explored the neural processes underlying audiovisual integration but the mechanism by which visual information modifies auditory speech perception remains uncertain.

According to the influential motor theory of speech perception formulated by Lieberman and colleagues in the 1950s, speech perception is a form of gesture perception and is therefore mediated by the involvement of the speech production system, in particular in audiovisual contexts (Liberman & Mattingly, 1985; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). More recently, the discovery of mirror neurons in the macaque monkey (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992) and the numerous studies that have since suggested the existence of a mirror neural system in humans (for reviews, e.g., Casile, Caggiano, & Ferrari, 2011; Rizzolatti & Craighero, 2004) have rejuvenated the idea that audiovisual integration for speech may involve the speech production system (e.g., Hasson, Skipper, Nusbaum, & Small, 2007; Skipper, Nusbaum, & Small, 2005). This position, however, has long been strongly contested by psycholinguists claiming that the integration of auditory and visual information is independent of any speech production motor processes (e.g., Massaro, 1972). Consistent with the latter view, neuroscientists studying multisensory integration using nonspeech audiovisual stimulations and

[1]CNRS/Aix Marseille Université, [2]Lyon Neuroscience Research Center, [3]University Lyon 1, [4]Trent University, Peterborough, Ontario, Canada

anatomical approaches have recently identified several different neural pathways by which visual information may modulate auditory processing, including feedback projections from polysensory structures, feedforward projections from nonspecific thalamic afferents, and direct lateral projections from the visual cortex (e.g., Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008).

Sams et al. (1991) were among the first to demonstrate indirectly that visual observation of speech can influence auditory perception very early in the time course of speech processing. In a magnetoencephalography (MEG) study, they presented oddball sequences of audiovisual speech stimuli in which the deviant syllables differed from the standards in their visual properties only. Specifically, the standard stimulus in the sequence was a congruent audiovisual /pa/ (auditory /pa/ paired with visual /pa/), whereas the deviant syllable consisted of an auditory /pa/ dubbed onto a visual /ka/ that was heard as "ta" by a fusion McGurk effect. The authors showed that, when this sequence was presented, the illusory auditory deviance induced by the visual deviant elicited an MMN response over auditory areas similar to those elicited by an acoustic deviant (for a review, see Näätänen, Paavilainen, Rinne, & Alho, 2007). Recently, the same group reported the symmetrical effect in which the MMN typically observed in presence of acoustic deviance was suppressed by a McGurk stimulus designed to preserve the auditory percept (Kislyuk, Mottonen, & Sams, 2008). In this case, the standard stimulus in the sequence was a congruent /va/ (auditory /va/ paired with visual /va/), whereas the deviant was an auditory /ba/ synchronized with a visual /va/ heard as "va." That is, although standard and deviant stimuli differed clearly acoustically, they were rendered indistinguishable by a visual capture effect, leading to the suppression of the MMN normally associated with acoustic differences.

In an fMRI experiment, using a repetition–suppression paradigm, Hasson et al. (2007) exploited a similar effect using pairs of syllables that differed both acoustically and visually but that were perceived as similar by way of McGurk audiovisual fusion. They identified regions that showed BOLD signal suppression (Grill-Spector, Henson, & Martin, 2006) for a pair of syllables in which an auditory /pa/ dubbed onto a visual /ka/ and perceived as "ta" was preceded by an audiovisual /ta/ (auditory /ta/ paired with visual /ta/). They reported suppression in response to the second stimulus in the pars opercularis of the left inferior frontal gyrus (IFG) and in the planum polare of the left superior temporal pole. They suggested that the speech production system, long associated with the left IFG, is involved in the abstract, modality-independent representation of the speech percept. This finding suggests that the speech production system is involved in abstract speech perception, perhaps by constraining perceptual options despite the variability in the acoustic signal. If this hypothesis is true, it would constitute a major advance toward solving a classic problem in speech perception. The hypothesis, however, also predicts that these same regions should demonstrate increased activity for pairs of syllables that have similar auditory qualities but that induce different perceptual experiences. If there is a brain region, which truly represents the perception of the stimulus independent from its sensory source information, then this region needs to pass both of these tests. This would be consistent with the McGurk-MMN results showing auditory MMN elicitation when the percept changes despite no acoustic change (Sams et al., 1991) and MMN suppression when the percept does not change despite an acoustic change (Kislyuk et al., 2008).

Our aim was to test this idea by creating two different audiovisual speech contexts both intended to identify brain areas whose levels of activation are conditioned by the percept. In the present fMRI study, we exploited McGurk audiovisual fusion to obtain short oddball sequences of syllables either (a) acoustically different but perceived as similar or (b) acoustically identical but perceived as different. We reasoned that if there is a single network of brain areas representing abstract speech perception, as suggested by Hasson et al. (2007), this network should show a reduction (suppression) of activity when presented with different acoustic syllables that are perceived as similar relative to when they are perceived as different (a) AND an increase of activity when presented with syllables that are acoustically similar but perceived as distinct relative to perceived as the same (b). To preview, in keeping with Hasson et al. (2007), the left inferior frontal cortex was part of the neural network that showed reduced activity for sequences of perceptually similar but acoustically different stimuli, consistent with the involvement of the speech production system in abstract speech perception. Another network was revealed, however, when focusing on the areas that exhibited increased activity for perceptually different but acoustically identical syllables. This alternative network included auditory areas but no left frontal activations. In addition, our findings point to the importance of subcortical structures less often considered when addressing issues pertaining to perceptual representations.

METHODS

Participants

Seventeen healthy volunteers (mean age = 24.9 years, nine women) with no record of neurological or psychiatric disorders participated in the study. All reported being right-handed, had normal or corrected-to-normal vision (magnet-compatible glasses), and were naive to the goal of the experiment. This group of participants was selected (from among 46 volunteers) on the basis of their results in a behavioral pretest assessing susceptibility to the fusion McGurk effect. In this pretest, participants were presented with 240 audiovisual stimuli: congruent /ba/, /ga/, /da/, and McGurk "da" (i.e., auditory /ba/ presented with a visual /ga/) presented each 60 times. For each audiovisual

syllable, three different utterances by the same speaker were used. A three-alternative forced-choice procedure was used; choices were [ba], [ga], or [da]. Participants were asked to determine what they heard while looking at the face. Those who exhibited the lowest percentages of [ba] responses to the McGurk "da" stimuli (with at least 90% of correct responses for the all congruent audiovisual stimuli) were selected to participate in the fMRI experiment. For the McGurk stimulus, the perceptual reports of the selected participants were as follows (mean ± *SD*): 13.5 ± 1.5%, 5 ± 0.6%, and 81.5 ± 1.5% for the choices [ba], [ga], and [da], respectively. Each participant provided informed written consent according to procedures approved by the *Comité de Protection des Personnes Sud Méditerranée I*.

Experimental Setup

Participants were presented with audiovisual clips of a woman speaker's lower face using a personal computer connected to a video projector. A digital video camera was used to make the audiovisual clips, and Avid ProTools 9 software (Avid Technology, Inc., Burlington, MA) was used to synchronize the sounds and images of the acoustically and visually incongruent McGurk syllables. The stimulus sequences were then edited using Final Cut Pro 7 software (Apple, Inc., Cupertino, CA). A custom-made LabVIEW (National Instruments, Inc., Austin, TX) program, triggered by the MR scanner control unit, was used to back-project the stimuli onto a frosted screen positioned at the end of the MRI tunnel and viewed by the participants through a mirror.

Audiovisual Stimuli and Task

The video recordings showed the mouth of a woman uttering syllables. A stationary fixation point (a white "pill") was superimposed on the image at the center of the mouth. Six different video clips (four audiovisual, one unimodal vision-alone, and one unimodal audio-alone), each corresponding to one of six conditions, were assembled. Each clip contained four syllables. The unimodal conditions were not used in the analyses of these data but were included to make the design consistent with a future analogue experiment with EEG. The four audiovisual conditions are presented in Table 1. The first three syllables in each clip were identical standards, and the fourth syllable was a deviant in all conditions (except one). Each four-syllable clip lasted 5400 msec; for each syllable, the acoustic signal duration was 360 msec, and the visual signal lasted about 1200 msec. We labeled the different sequences ≠A≠P, ≠A=P, =A≠P, and =A=P so as to indicate whether the acoustic signal of the fourth syllable changed (≠A) or not (=A) from the first three andwhether the percept was changed (≠P) or not (=P). The visual-alone condition showed the woman's lower face pronouncing the syllable sequence */ba ba ba ga/*. For the audio-alone condition, the sound track of */da da da ba/* was coupled with a fixed image of the closed woman's mouth. The imaging session comprised nine functional runs. In each functional run, all six conditions were presented six times each in a pseudorandom order. The four-syllable clips were separated by the presentation of a fixed image of the woman's lower face that lasted on average 3 sec (random duration chosen from an exponential distribution). Participants did not perform any auditory perceptual task. Instead, a visual detection task was used to ensure that they maintained attention on the woman's mouth. We used no explicit speech-sound perception task, as our purpose was to target perceptual mechanisms independently (as much as possible) for decision-making processes. Also, our idea was to use an experimental design that matched as closely as possible those most often used in EEG/MEG studies examining MMN responses (e.g., Näätänen, Kujala, & Winkler, 2011, for a review on the MMN; e.g., Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007, for a study on the MMN-McGurk) and the one used in the fMRI study by Hasson et al. (2007) in which participants did not perform any task. During each functional run, the superimposed white pill disappeared for 500 msec pseudorandomly once for each condition (six times per run). Participants were instructed to press a button as quickly as possible with their left thumb each time this happened.

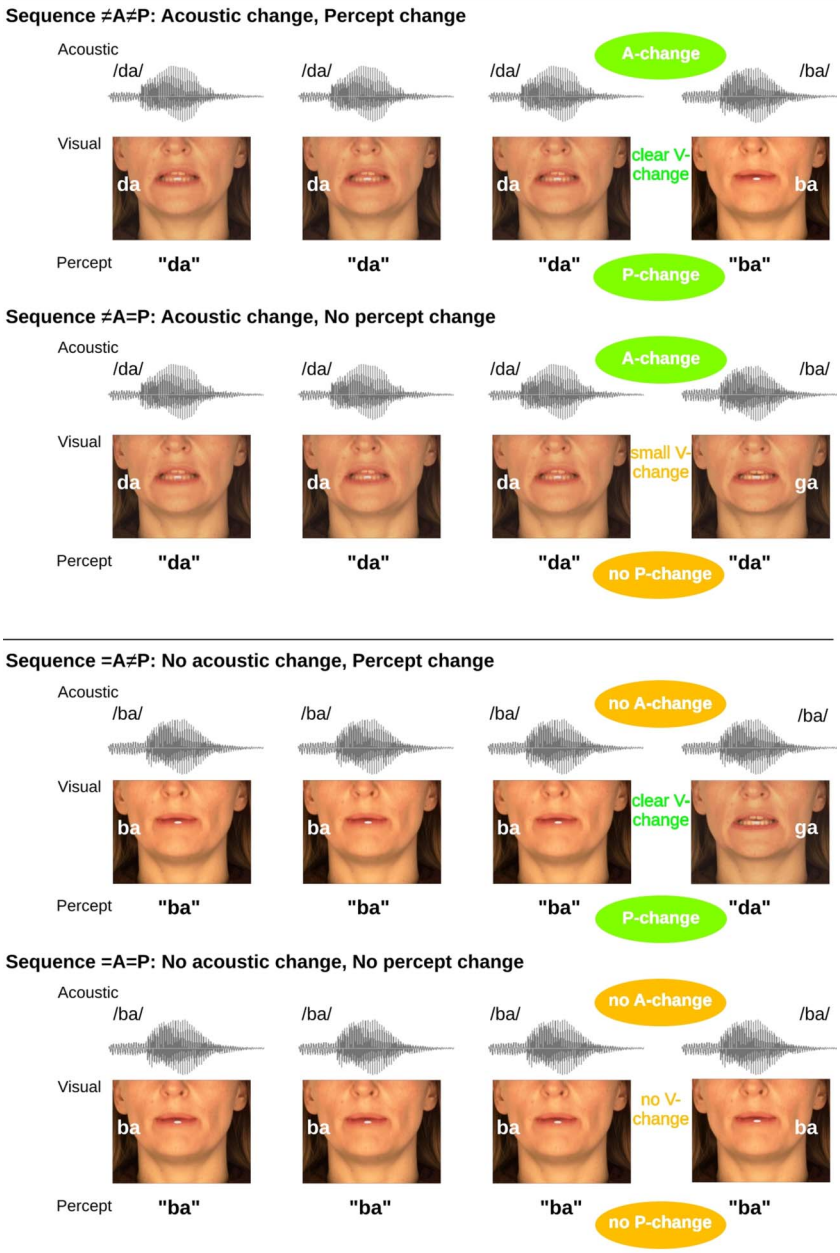
Imaging Procedure

Images were acquired on a 3-T MEDSPEC 30/80 AVANCE whole-body scanner (Bruker, Ettlingen, Germany) equipped with a circular polarized head coil. Participants were lying comfortably in the supine position in the MR scanner. An ergonomic MR-compatible response button was placed in the participant's left hand. Headphones were used to present the auditory stimuli and to communicate with the participant, as well as to dampen the scanner noise in conjunction with earplugs. Each imaging session comprised nine functional runs followed by a single high-resolution anatomical scan. Functional volumes were collected using a T2*-weighted echo-planar sequence covering the whole brain with 30 interleaved 3-mm-thick/1-mm gap axial slices (repetition time = 2000 msec, echo time = 30 msec, field of view = 192 × 192 mm^2, 64 × 64 matrix of 3 × 3 × 4 mm voxels). Each run lasted about 5 min, during which 158 functional volumes were acquired. Anatomical MRI data were acquired using high-resolution structural T1-weighted image (inversion recovery sequence, resolution 1 × 1 × 1 mm) in the sagittal plane, covering the whole brain.

fMRI Data Preprocessing

Data were processed and analyzed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London) along with custom-made MATLAB programs.

Table 1. Design of the Sequences of Syllables in the Video Clips



The pictures show the movie 80 msec after the acoustical onset of each four syllables. For each syllable, acoustical signal duration was 360 msec and visual signal lasted about 1200 msec.

The first six functional volumes acquired in each run were discarded to ensure that longitudinal relaxation time equilibration was achieved. The remaining 152 images were corrected for differences in slice acquisition time. The 14th slice acquired was chosen as a reference to correct for temporal differences between the first and last slices. All volumes were realigned to the first volume to correct for head movements between scans. The functional images were unwarped using fieldmap information before being coregistered to each individual anatomical T1-weighted image and spatially normalized using DARTEL-SPM8 procedure to the Montreal Neurological Institute (MNI) standard space. Data were then spatially smoothed using an 8-mm FWHM isotropic Gaussian kernel to accommodate for interparticipant differences in anatomy.

Statistical Analyses

Trials during which a manual response was produced (left thumb button-press) were not modeled. Event-related analyses were run according to the following general linear model: The BOLD response to the fourth

syllable (deviant) of each clip was modeled as a 1200-msec duration event (i.e., duration of the visual signal) convolved with the SPM8 canonical hemodynamic response function to create one regressor for each condition (≠A≠P, ≠A=P, =A≠P, and =A=P). We did not explicitly model the initial three syllables because they are identical in the conditions that we aim to compare directly and thus the contrasts of interest would not be affected (see below). The mean of the scans of each run as well as weights corresponding to a linear regression capturing possible signal drifts within the different runs were also included in the design matrix. Contrast images were generated for each participant and were then submitted to the group analysis using a random-effects model.

RESULTS

Our aim was to examine two contrasts, both intended to reveal the neural networks coding for the percept of speech sounds. We designed two audiovisual sequences of syllables instantiating two symmetrical perceptual effects: in sequence ≠A=P, acoustically different stimuli were heard as similar, whereas in sequence =A≠P, acoustically identical stimuli were perceived as different. Then these two sequences were each paired with another sequence that was composed of the same four acoustical tokens (sequence ≠A≠P and =A=P, respectively; see Table 1), but in which the fourth syllable (always an auditory /ba/) was heard differently because of different visual information (fourth syllable indeed perceived as "ba" in ≠A≠P and =A=P but perceived as "da" in ≠A=P and =A≠P).

For the contrast (≠A≠P vs. ≠A=P; see top panel of Table 1), in both conditions the fourth syllable differed acoustically from the three preceding ones, but in the sequence ≠A=P a fusion McGurk effect was elicited by the fourth stimulus so that speech perception remained unchanged throughout the whole four-syllable sequence. Thus, this first contrast was intended to reveal regions that exhibited less activity when all four syllables were heard as the same (≠A=P) than when a change was perceived (≠A≠P).

In the second contrast (=A≠P vs. =A=P; see bottom panel of Table 1), in both conditions the fourth syllable was acoustically identical to the first three, but in the sequence =A≠P the McGurk effect induced a change in the percept of the fourth syllable. This comparison was used to capture areas that showed increased activity for a change in percept, although no acoustic change occurred.

To determine the overlap between the two sets of areas revealed by each contrast, we ran an intersection analysis (i.e., we searched for voxels significantly activated in both contrasts), and to further distinguish the two networks, we also examined the interaction effect. To test the simple and the interaction effects, we set individual voxel threshold at $p < .001$, uncorrected (threshold value comparable to, e.g., McKenna Benoit et al., 2010; Hasson et al., 2007; Ojanen et al., 2005; Pekkola et al., 2005), and considered only clusters larger than 100 mm^3. For all areas identified using this criterion, we also tested significance at the cluster level (corrected $p < .05$) using a permutation test taking into account cluster size and level of activation. Using 5000 label condition permutations, significance of each cluster was tested as follows: for each permutation one "false positive" was counted if at least one cluster of activation exhibited greater extent (larger number of voxels) and higher level of activation (larger mean and maximum t value) than the cluster of interest. The percentage of "false positives" among the 5000 permutations corresponds to an adjusted p value (see Hénaff, Bayle, Krolak-Salmon, & Fonlupt, 2010, for details). For this purpose, we defined two ROIs: (1) a large region including areas known to be involved in speech processing: the inferior frontal cortex (pars opercularis, pars triangularis, and orbital part), the temporal areas (Heschl, middle and superior temporal, and temporal pole superior), the supramarginal (SMG) area, and the insula and (2) an ROI restricted to the two regions identified by Hasson et al. (2007): the pars opercularis and temporal pole. For the intersection analysis, we lowered the voxel-level threshold to $p < .01$, uncorrected, and discarded only clusters smaller than 36 mm^3. Regions are labeled according to Tzourio-Mazoyer et al. (2002).

Areas Showing Reduced Activity (Suppression) with Similar Percept Despite an Acoustic Change

First, to identify regions that showed less (suppressed) activity when all four syllables were heard as similar (despite an acoustical change), this condition was compared with one in which the acoustic change was heard (≠A≠P > ≠A=P). The results are presented in Table 2A and in Figure 1A. The largest effects were found subcortically, where a large area of activation extended bilaterally with its peak located in the left caudate nucleus and a cluster was located with its peak in the right thalamus extending into the right pallidum and putamen. Globally the modulation in activation was more pronounced for the right hemisphere with clusters in the insula, the frontal inferior orbital cortex, and the frontal superior cortex, as well as laterally in the STS. In the left hemisphere, activity modulation was observed in the opercular part of the frontal inferior cortex. When lowering the threshold to $p < .005$, uncorrected, this latter cluster extended precentrally with a peak in the precentral gyrus (MNI: −50, 0, 38). Medially, a cluster was observed in the SMA.

Areas Showing Increased Activity (Release from Adaptation) with Different Percept in the Absence of an Acoustic Change

The relation (=A≠P > =A=P) was intended to reveal regions exhibiting more activity when, in the absence of any acoustic change, perceptual deviance was induced

Table 2. Percept Change (≠P) > No Percept Change (=P)

A **With acoustical change (≠A): (≠A≠P) > (≠A=P)**

Regions	Hemisphere	MNI max x	MNI max y	MNI max z	*T*-value	Cluster size (mm^3)	
Temporal							
temp mid (STS)	right	64	−38	−2	5.5856	688	**
					Total volume = 688		
Parietal							
SMG	right	60	−42	26	4.0186	104	
					Total volume = 104		
Frontal							
insula	right	38	6	−2	4.9961	672	**
frontal inf orb	right	42	24	−16	4.5361	536	**
SMA		0	−2	68	4.3226	296	
frontal inf oper	left	−60	10	24	4.0701	128	*
					Total volume = 1632		
Subcortical							
caudate	left	−10	−2	12	6.7626	3024	***
thalamus	right	10	−12	2	5.3141	1952	***
					Total volume = 4976		

B **Without acoustical change (=A): (=A≠P) > (=A=P)**

Regions	Hemisphere	MNI max x	MNI max y	MNI max z	*T*-value	Cluster size (mm^3)	
Temporal							
temp sup/Heschl	right	54	−18	8	5.5678	912	**
temp sup (STS)	right	54	−40	8	5.0540	216	
					Total volume = 1128		
Parietal							
SMG	left	−50	−26	16	4.2467	112	
					Total volume = 112		
Subcortical							
thalamus	right	6	−22	14	7.5036	1560	***
					Total volume = 1560		

C **With AND without acoustical change (≠A AND =A): (≠A≠P)>(≠A=P) AND (=A≠P)>(=A=P)**

Regions	Hemisphere	MNI max x	MNI max y	MNI max z	*T*-value	Cluster size (mm^3)
Temporal						
temp sup (STS)	right	54	−42	4	3.2632	224
Subcortical						
thalamus	right	10	−20	10	3.4093	424

(A) Areas showing reduced activity with similar percept (despite an acoustic change). Voxel-level threshold: $p < .001$ uncorrected and minimum cluster size of 100 mm^3 were used. (B) Areas showing increased activity with different percept (in the absence of any acoustic change). Voxel-level threshold: $p < .001$ uncorrected, and minimum cluster size of 100 mm^3 were used. (C) Areas showing reduced activity with similar percept AND increased activity with different percept. Voxel-level threshold: $p < .01$ uncorrected and minimum cluster size of 36 mm^3 were used. Presented t values correspond to the contrast (≠A≠P > ≠A=P) masked by (=A≠P > =A=P). In A and B, cluster-level significance (corrected $p < .05$) are indicated depending on the search space that was used (see text for details): (***) whole-brain, (**) large speech-related areas ROI, and (*) ROI defined on the basis of Hasson et al.'s (2007) report.

by the fusion McGurk effect. Table 2B and Figure 1B summarize the results. The largest activation was lateralized to the right hemisphere, with the highest peaks located in the thalamus. Activity was also found in the right Heschl's area as well as in the right STS.

Reduced Activity with Similar Percept AND Increased Activity with Different Percept

In the Introduction, we argued that brain regions associated with the abstract representation of speech sounds

should show both a reduction (suppression) of activity when presented with syllables that are acoustically different but perceived as similar AND an increase of activity (release from adaptation) when presented with syllables that are acoustically identical but perceived as distinct. To determine if there was a detectable network of regions that showed this pattern, we conducted an analysis intended to find the intersection of the two sets of regions identified above. Very little overlap was observed between the two networks revealed by the symmetric perceptual manipulations. Namely, with a threshold as low as $p < .01$ uncorrected, the only regions that showed differential levels of activation for both contrasts were two clusters located in the right STS and the right thalamus respectively (see Table 2C and Figure 1C).

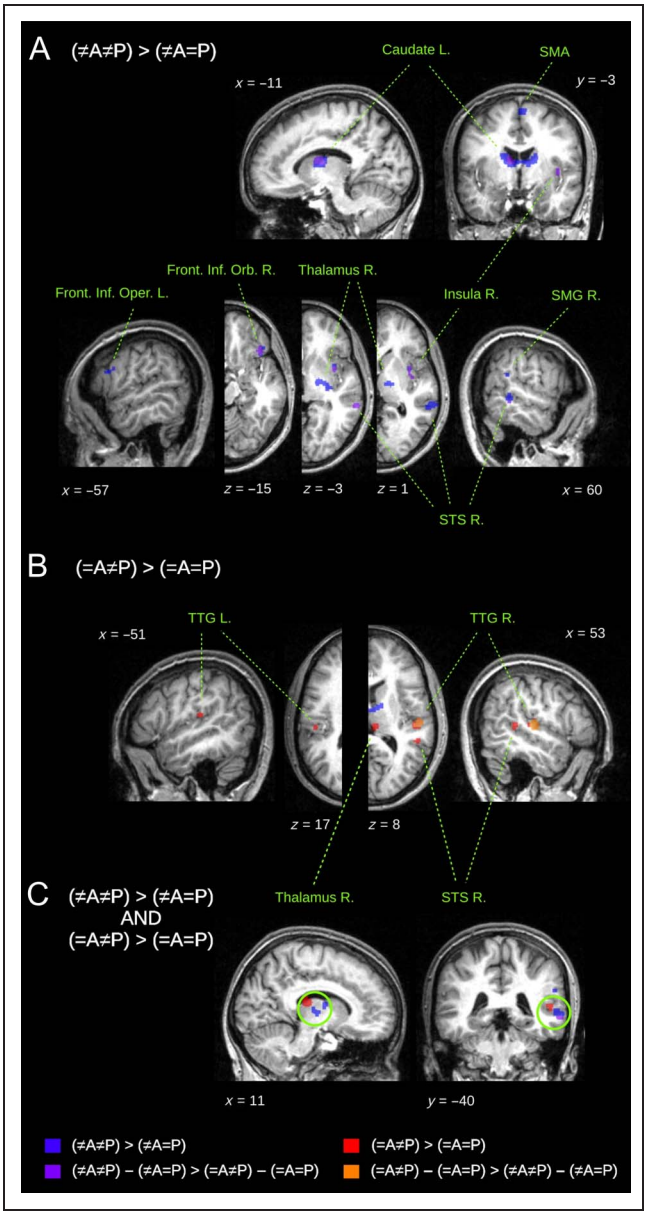


Figure 1. Activation clusters showing (A) reduced activity with similar percept (despite an acoustic change), (B) increased activity with different percept (in the absence of any acoustic change), (C) reduced activity with similar percept AND increased activity with different percept. (A, B, and C) Voxel-level threshold: $p < .001$ uncorrected.

Reduced Activity with Similar Percept OR (but Not AND) Increased Activity with Different Percept

To further dissociate the two networks, we analyzed the interaction (≠A≠P vs. ≠A=P) − (=A≠P vs. =A=P). We considered this contrast for all the areas that exhibited either reduced activity with perceptually similar stimuli OR increased activity with perceptually different stimuli (the union of the two sets reported above); that is, we defined an inclusive mask corresponding to the regions for which either (≠A≠P > ≠A=P) or (=A≠P > =A=P) hold, for a voxel-level threshold of $p < .001$ uncorrected.

First, we considered the areas for whom the relation (≠A≠P vs. ≠A=P) > (=A≠P vs. =A=P) holds, that is, areas that showed larger change in activation for a percept change (vs. no percept change) in the presence of an acoustical change than in the absence of acoustic deviance. The largest cluster was located in the left caudate. In the right hemisphere, clusters were found in the STS, the insula, as well as the right frontal orbital cortex. A region of the right thalamus was also revealed (see Table 3A and Figure 1A).

The inverse relation (=A≠P vs. =A=P) > (≠A≠P vs. ≠A=P) was intended to reveal regions whose activities were more clearly modulated by perceptual change in the absence of any acoustic deviance than when acoustics changed concurrently. The only cluster for which this relation held was located in auditory areas in the right transverse temporal gyri (Heschl; see Table 3B and Figure 1B).

(See the Appendix, for a potential effect related to audiovisual (in)congruency, as well as for the relations (≠A=P > ≠A≠P) and (=A=P > =A≠P), which are the opposite of the two main contrasts of interest and exhibit in particular activations in visual motion processing areas.)

DISCUSSION

Speech perception is faced with the lack of invariance problem: There is variability in the relationship between the acoustic signal and the perceived phoneme. The motor theory of speech perception was originally proposed to solve this lack of invariance problem. According to this view, speech sounds are transduced into gestural codes consisting of invariant motor programs. The conscription of these motor programs to the perception process narrows the range of interpretations. However, this idea has been challenged by vocal tract imaging studies that demonstrate high variability in motor articulation too. In general, the hypothesis that speech perception is always mediated by the motor system is difficult to reconcile with neuropsychological and experimental data. As

Table 3. Areas Showing Reduced Activity with Similar Percept OR (but Not AND) Increased Activity with Different Percept

A Effect of percept change with acoustical change (≠A) > Without acoustical change (=A)

(≠A≠P)-(≠A=P) > (=A≠P)-(=A=P)

Regions	Hemisphere	MNI max			*T*-value	Cluster size	
		x	*y*	*z*		(mm^3)	
Temporal							
temp mid (STS)	right	64	−38	−6	5.2730	152	**
Frontal							
insula	right	38	6	−4	4.3435	352	
frontal inf orb	right	42	28	−18	5.2851	216	**
Subcortical							
caudate	left	−10	0	14	4.6305	576	**

B Effect of percept change without acoustical change (=A) > With acoustical change (≠A)

(=A≠P)-(=A=P) > (≠A≠P)-(≠A=P)

Regions	Hemisphere	MNI max			*T*-value	Cluster size	
		x	*y*	*z*		(mm^3)	
Temporal							
temp sup/Heschl	right	56	−18	8	6.1377	584	**

(A) Reduced activity despite an acoustic change. (B) Increased activity with different percept in the absence of any acoustic change. For A and B voxel-level threshold: $p < .001$ uncorrected and minimum cluster size of 100 mm^3 were used. (**) areas significant at the cluster level ($p < .05$); search space defined by the inclusive mask (see Methods).

examples, the destruction of the motor cortex does not preclude functional speech perception and language comprehension, and results of imaging studies have been inconsistent in demonstrating activation of the motor system when speech stimuli are presented in the auditory modality alone and only passive listening is involved (for a review, e.g., Skipper et al., 2006).

Nevertheless, the body of evidence suggesting the existence of a mirror neural system in humans provides support to the idea that speech perception may involve perception of vocal tract gestures. Also, although there have been few specific algorithms describing how motor processes could contribute to perception, new formulations indebted to Liberman's theory and Stevens and Halle's analysis-by-synthesis model (Stevens & Halle, 1967) have integrated advances in computational motor control modeling (Jordan & Rumelhart, 1992). According to these views (e.g., Skipper et al., 2006), during audiovisual speech perception, inferior frontal and premotor regions could play a role akin to generating an initial hypothesis about the phoneme produced by the speaker. At the level of the opercular part of the IFG (Broca's area), where mirror neurons putatively exist (Kilner, Neal, Weiskopf, Friston, & Frith, 2009), the abstract goal of observed actions would be activated and translated into motor commands through interaction with premotor and primary motor cortices. Subsequent efference copy (von Holst & Mittelstaedt, 1950), or corollary discharge (Sperry, 1950) or a forward model (Jordan & Rumelhart, 1992) would predict the acoustic and somatosensory consequences of the activated motor commands. This sensory prediction could then influence auditory perception.

In their fMRI study, Hasson et al. (2007) considered the existence of a neural representation of speech syllable perception that transcends its sensory properties. As a matter of fact, identifying regions that represent the abstract aspects of the speech percept would represent a major advance in determining how we deal with variability in the acoustic signal, a classic problem in speech perception. These authors reported two left-hemisphere regions (pars opercularis and planum polare) exhibiting reduced activity (suppression) when syllable sequences were perceptually similar despite being acoustically different. According to them, the identified brain regions fit into the theoretical framework sketched above in which Broca's area, an area long associated with speech production, plays a central role.

The goal of our study was to question whether a unique neural network encoding sublexical audiovisual speech at an abstract perceptual level could be identified. In particular, we assessed if the same network would be revealed by two symmetric perceptual manipulations: (a) presenting different acoustic signals that are heard as similar, for which a decrease in activity was expected in the unique network, as in Hasson et al. (2007) and (b) presenting identical acoustic signals that are heard as different, which should entail an increase in activity in the same network. Comparing these two situations within the same participants allowed us to control for interindividual variability in susceptibility to the audiovisual fusion

effect. This was necessary as we chose to avoid having participants perform an auditory perceptual task and instead used an unrelated visual detection task to ensure that visual information was encoded. Indeed, having neural activation induced by higher-level decision-making processes could have obscured some patterns of activation (e.g., Binder, Liebenthal, Possing, Medler, & Ward, 2004). Nonetheless, we found very little overlap between the set of areas revealed by the two perceptual manipulations.

Frontal Areas and the Left Caudate Showed Reduced Activity for Sequences of Perceptually Similar Syllables

Presenting syllables that have different acoustic components but are perceived as similar (i.e., using the contrast: ≠A≠P > ≠A=P; Table 2A) paralleled the strategy used by Hasson et al. (2007) to identify percept-sensitive regions. However, although these authors employed a brain parcellation restricted to the cerebral cortex (Fischl et al., 2004) and found activity related to the abstract representation of speech sounds in the cortex (left frontal pars opercularis and left temporal planum polare), we found the largest activity modulation subcortically with the highest peak in the left caudate.

The idea that the BG may be involved in speech is supported by neuropsychological work showing a relation between pathology of the putamen and caudate nuclei and aphasia (e.g., Lieberman et al., 1992). Several studies have suggested that the left caudate in particular, in relation with the thalamus, is implicated in word comprehension and articulation in bilingual speakers (e.g., Crinion et al., 2006). Moreover, models have been proposed in which articulatory–auditory and articulatory–orosensory mappings involving the left caudate would be used to facilitate perceptual identification when a phonetic contrast is ambiguous, as in the case of second-language speakers (Callan et al., 2004).

At the cortical level, temporal and parietal regions showing differential activation were right-hemispher e STS and the SMG. STS involvement in audiovisual integration processes in general (for a review, see Campbell, 2008) and in the McGurk fusion effect in particular (Beauchamp, Nath, & Pasalar, 2010) has been well established. Note, however, that because the stimuli contrasted here were audiovisual and because activation was reduced for the condition involving the McGurk stimuli, the activation we found in the posterior STS likely corresponds to a region known to be activated by sight of speech mouth movements (Puce, Allison, Bentin, Gore, & McCathy, 1998). Together with posterior STG/STS, planum temporal, and superior temporal parietal junction, the SMG bilaterally has been implicated in both speech perception and production (e.g., Hickok & Poeppel, 2000; Paus, Perry, Zatorre, Worsley, & Evans, 1996) and has been suggested to serve as an interface between articulatory, auditory, and orosensory mappings (Callan et al., 2004). Activation of right-hemisphere STS is consistent with lateralization observed in functional imaging studies of (full body) biological motion (e.g., Grossman et al., 2000).

Frontally, right-dominant modulation of activity was observed in the insula, an area implicated in multisensory integration (e.g., Bushara, Grafman, & Hallett, 2001), and the inferior orbital cortex. A left-hemisphere cluster was located in the opercular part of the IFG. Interestingly, this activation extended (when lowering the threshold) into a cluster with a second peak in the precentral gyrus; that is, a region of the premotor cortex (PMC) implicated in both speech perception and production (Wilson, Saygin, Sereno, & Iacoboni, 2004). Broca's area, the PMC, and the anterior insula are interconnected regions known to be involved in articulatory processing (e.g., Dronkers, 1996; Habib et al., 1995; Paulesu, Frith, Bench, & Bottini, 1993) and that are functionally connected to auditory and somatosensory areas. Here, one should note that our results are consistent with Hasson et al.'s (2007) finding as for the pattern of activation found in the opercular part of the IFG. This, despite differences in the tasks used in the two studies: In their experiment, participants passively listened to and observed the stimuli, whereas here they performed a simple visual detection task. Medially, activation in the SMA, an area typically involved in attentional control, movement selection, as well as in speech production (Wilson et al., 2004), was also correlated with the abstract perceptual dimension of the stimuli.

One should also notice that no change in the activity in the transverse temporal gyrus (TTG or Heschl's gyrus) was revealed. This is important as it suggests that, for this perceptual manipulation (making acoustically different syllables be heard as similar), processing of the acoustical signal in the primary/secondary auditory areas was not altered by the visual input. This finding stands at odds with the results by Kislyuk et al. (2008). Indeed, these authors reported a suppression of the auditory MMN induced by visual information, which suggested that under the influence of the visual input the auditory cortex failed to discriminate between the acoustic stimuli. One possible explanation for this discrepancy might be found in the nature of the stimuli that were used. Kislyuk et al. (2008) exploited a visual-capture effect (auditory /ba/ paired with visual /va/ and perceived "va"), whereas we used a fusion effect in which a new percept emerges that was not present in the auditory and visual channels in isolation (auditory /ba/ paired with visual /ga/ and perceived "da").

Acoustically Identical but Perceptually Different Syllables Increase Activity in Temporal Areas

Although the network that showed reduced activity for similar percept (despite an acoustic change) was consistent with the idea that frontal regions part of the speech production system might be involved in the modality-independent representation of speech percept, the areas that exhibited

increased activity when the speech percept changed in the absence of any acoustic change (contrast: =A≠P > =A=P; Table 2B) were almost entirely different. Indeed, we found very little overlap between the areas revealed by the two perceptual manipulations in the intersection analysis intended to assess this overlap. Only small regions in the right STS and right thalamus exhibited activation in both cases (using a threshold lowered to $p < .01$ uncorrected; Table 2C). In both contexts, the strongest activations were observed subcortically but located differently. For the contrast (=A≠P > =A=P), the highest peak was found posterior in the right thalamus, whereas the strongest activation was located in the left caudate for the contrast (≠A≠P > ≠A=P). Also, although in both cases activation was elicited in the STS, this activity was more lateral for (≠A≠P > ≠A=P).

Interestingly, for the relation revealing release from adaptation with a percept change (=A≠P > =A=P), the cortical area that showed the most activation was the right TTG (Heschl) as well as the left SMG. This finding stands in striking contrast to the one found for suppression (≠A≠P > ≠A=P), in which there was no change in auditory TTG activity. It is, however, consistent with previous results reporting activation of primary and secondary auditory cortex during lipreading (e.g., Besle et al., 2008; Pekkola et al., 2004; Calvert et al., 1997), as well as with studies in MEG/EEG that used similar perceptual manipulation to elicit a so-called McGurk-MMN over the auditory areas (Saint-Amour et al., 2007; Colin, Radeau, Soquet, & Deltenre, 2004; Colin et al., 2002; Möttönen, Krause, Tiippana, & Sams, 2002; Sams et al., 1991). Importantly, no frontal activation was found, which suggests that the influence of the visual input on the auditory processing is not mediated by frontal speech production structures.

Electrophysiological recordings in human and nonhuman primates suggest that audiovisual interaction can occur as early as the first stage of cortical auditory processing in primary auditory area (A1). In particular, according to Schroeder et al. (2008), multisensory convergence does not begin in the STS, but in other areas near and in the primary auditory cortex. These authors propose a mechanism by which ongoing oscillatory activity of local neuronal ensembles in the primary auditory cortex is predictively modulated by visual inputs so that acoustic signals arriving during a high excitability phase are amplified. To be effective, visual inputs need to reach A1 slightly before auditory inputs, which is true in the case of visible articulatory gestures as they typically precede audible vocalizations. Visual information could reach A1 by several anatomical pathways, including a direct ascending (i.e., a nonspecific thalamic) input, a direct lateral connection from the visual cortex, or an indirect feedback input from the multisensory areas of the STS. However, one needs to acknowledge that, although this framework offers a plausible neural mechanism by which vision could enhance audition, how a new auditory percept emerges from the fusion of two incongruent sensory inputs, as in the McGurk effect, remains uncertain.

No Unique Network Represents Speech Percept Independent from Its Sensory Source

The two main contrasts we used to test the generality of the Hasson et al. (2007) results, (≠A≠P > ≠A=P) and (=A≠P > =A=P), both contrasted sequences that differed in terms of audiovisual congruency, that is, including or not the McGurk stimulus. In order, to rule out an interpretation in those terms, we ran two additional intersection analyses (Table A1). Among the areas that were identified by the relation (≠A≠P > ≠A=P), the regions whose activation could be explained by an audiovisual-congruency effect were the SMA, the right SMG, insula, and inferior orbital cortex. The second intersection analyses indicated that only activation in the right thalamus could have been because of this confound in (=A≠P > =A=P).

Thus, our results show that the neural network identified by presenting syllables that are acoustically different but heard as similar fits well with the model of audiovisual speech perception proposed by Skipper et al. (2006) in which articulatory motor representations play a central role. Visual speech information would, by a predictive mechanism carried out by a sensorimotor internal forward model, influence auditory perception through a network involving the pST, opercular part of IFG, PMC, and SMG. Nevertheless, the possible involvement of subcortical regions should not be overlooked. Indeed, the largest changes in activity we observed were in the BG and the thalamus. Consistent with this observation, Callan et al. (2004) in a model of unimodal auditory speech perception also inspired by motor control theories, proposed the cerebellum as a site of putative internal model instantiations (Jordan & Rumelhart, 1992), and suggested that the BG may play a role (that remains undetermined) in the selection of these internal models.

However, a different pattern of cortical activation was observed when using syllables having identical acoustical components but that were heard as different. In this case, the location of the activity changes pointed to early audiovisual interaction and more specifically appeared to be compatible with the scenarios proposed by Schroeder et al. in which visual information reaches (and predictively modulates the activity of) the auditory areas in the TTG through thalamic connections or inputs from the visual cortex. Although we appreciate how valuable and polymorphous the concept of prediction can be, we briefly speculate on the role "predictive coding" may play in the explanation of our findings (e.g., Wacongne, Changeux, & Dehaene, 2012; Friston, 2005; Lee & Mumford, 2003; Rao & Ballard, 1999). According to this theoretical framework, the nervous system does not respond passively to sensory inputs but instead actively predicts the incoming

stimuli. The brain learns the regularities in sensory inputs and develops sensory predictions on the basis of internal models. Any discrepancy between the predictions and the actual inputs would result in a mismatch signal.

In the sequence =A≠P, the visual transition from the standards /ba/ to /ga/ clearly violates the prediction based on the regularities in the preceding audiovisual stimuli (Table 1, bottom). Thus, the response to the deviant input, preceding the acoustical input, may have influenced the auditory processing through a mechanism akin to that one proposed by Schroeder et al. (2008). In contrast, in the sequence ≠A=P, only a subtle change occurs between the visual /da/ and /ga/ (Table 1, top). Thus, one may speculate that in this case the visual /ga/ does not produce any clear sensory prediction violation, and thus, no substantial change occurred in the interaction between the visual and auditory system that may have altered the processing of the acoustic signal. This could explain why no change in the activation in the TTG auditory areas was revealed by the contrast (≠A≠P > ≠A=P).

Thus overall, the two symmetrical perceptual manipulations tested here using the McGurk fusion effect did not allow to uncover a unique network involved in abstract audiovisual speech representation. Instead, frontal areas (overlapping the speech motor system), temporal auditory areas, and subcortical structures appear differentially involved in audiovisual speech perception depending on the context of perception.

APPENDIX

Controlling for Audiovisual Congruence Effect

We used the interaction contrast [(=A≠P) − (=A=P)] − [(≠A≠P) − (≠A=P)] (Table 3) to dissociate the network showing increased activity with different percept in the absence of any acoustic change (revealed by (=A≠P) > (=A=P)) from the regions showing reduced activity with similar percept despite the presence of an acoustic change (identified by (≠A≠P > ≠A=P)). However, this interaction contrast is equivalent to [(=A≠P) + (≠A=P)] − [(≠A≠P) + (=A=P)], comparing the two sequences including the incongruent McGurk stimulus to those composed of congruent syllables (see Table 1).

To address this potential confound, we considered the two additional contrasts: (=A≠P > ≠A=P) and (≠A≠P >

Table A1. Intersection Analyses

A (≠A≠P)>(≠A=P) AND (=A≠P)>(≠A=P)

Regions	Hemisphere	MNI max x	y	z	*T*-value	Cluster size (mm^3)
Temporal						
temp mid (STS)	right	70	–38	0	3.7751	160
					Total volume = 160	
Frontal						
inf oper / prec	left	–58	4	28	3.4745	160
precentral gyrus	left	–54	2	38	3.3977	168
					Total volume = 328	
Limbic						
cingulum post	right	12	–40	8	3.1717	104
					Total volume = 104	
Subcortical						
caudate	left	–10	0	10	6.1626	4120
putamen	right	30	–14	–8	4.8006	1192
					Total volume = 5312	

B (=A≠P)>(=A=P) AND (≠A≠P)>(=A=P)

Regions	Hemisphere	MNI max x	y	z	*T*-value	Cluster size (mm^3)
Temporal						
temp mid (STS)	right	54	–40	8	5.0540	456
temp sup/Heschl	right	50	–26	8	3.4984	104
					Total volume = 560	
Parietal						
SMG	left	–50	–24	16	3.6781	184
					Total volume = 184	

(A) Areas identified using the intersection of the two contrasts (≠A≠P > ≠A=P) AND (=A≠P > ≠A=P); presented *t* values correspond to the contrast (≠A≠P > ≠A=P) Masked by (=A≠P > ≠A=P). (B) Regions showing activation for (=A≠P > =A=P) AND (≠A≠P > =A=P); presented *t* values correspond to (=A≠P > =A=P) masked by (≠A≠P > =A=P).

For both intersection analyzes, voxel-level threshold: $p < .005$ uncorrected.

Table A2. No Percept Change (=P) > Percept Change (≠P)

A **With acoustical change (≠A): (≠A=P) > (≠A≠P)**

Regions	Hemisphere	MNI max x	MNI max y	MNI max z	*T*-value	Cluster size (mm^3)
Occipital						
occipital sup	left	–22	–84	26	–5.6947	368
cuneus	left	–6	–70	22	–4.2986	360
					Total volume = 728	
Temporal and parietal						
parietal sup	left	–22	–64	62	–5.1836	440
angular	right	52	–68	34	–4.6905	152
angular	left	–52	–68	26	–4.6962	144
					Total volume = 736	
Frontal						
frontal sup	left	–24	18	60	–3.9907	112
					Total volume = 112	
Cerebellum						
6/Crus1	right	18	–64	–36	–5.1702	144
					Total volume = 144	

B **Without acoustical change (=A): (=A=P) > (=A≠P)**

Regions	Hemisphere	MNI max x	MNI max y	MNI max z	*T*-value	Cluster size (mm^3)
Occipito-temporal						
occ/temp mid	left	–42	–68	10	–5.1016	808
lingual	left	–14	–86	–16	–4.9904	1088
					Total volume = 1896	
Parietal						
parietal sup	left	–28	–48	56	–7.4050	664
					Total volume = 664	
Frontal						
frontal mid	right	38	48	32	–5.8686	472
front inf tri	left	–50	20	4	–4.5297	200
frontal inf orb	left	–22	26	–16	–5.4656	152
					Total volume = 824	

For both contrasts, voxel-level threshold: $p < .001$ uncorrected and minimum cluster size of 100 mm^3 were used.

=A=P). Like our main contrasts of interest, these contrasts both compare a sequence including a perceptual change (≠P) with another one without change of percept (=P). But, in opposition to the main contrasts, they are equivalent in terms of audiovisual congruence: (=A≠P > ≠A=P) compares sequences including the McGurk stimulus, whereas in (≠A≠P > =A=P) the two contrasted sequences are both composed of congruent syllables only. On one side, for (=A≠P > ≠A=P), we found regions similar to those identified by the main contrast (≠A≠P > ≠A=P). On the other side, for (≠A≠P > =A=P) activations were similar to those revealed by our second main contrast (=A≠P > =A=P).

Thus to exclude, from the two networks that we wanted to dissociate, any activation that could be attributed to differences in audiovisual congruence, we proceeded as follows. First, we considered the intersection (≠A≠P > ≠A=P) AND (=A≠P > ≠A=P), whose first part is the first main contrast of interest (reduced activity with similar percept despite an acoustic change), and the second half serves as a control for audiovisual incongruence. The areas identified using a threshold set at $p < .005$ are listed in Table A1A. Left hemisphere frontal regions known to be involved in speech production (opercular part of the frontal inferior cortex and left precentral gyrus) could be identified. Also, a large cluster extending from the left caudate to the right thalamus was identified. Altogether, the areas thus revealed are very similar to those revealed by (≠A≠P > ≠A=P; see Table 2A).

In an analog way, the second intersection analysis was intended to determine the areas that exhibited modulated activity for (=A≠P > =A=P) AND (≠A≠P > =A=P). Again, as in the second part of the intersection the two contrasted sequences are equivalent as for congruence (composed of congruent syllables only), any activity modulation that could have been related to changes in congruence in the main contrast of interest (=A≠P >

=A=P) was excluded. Activations in right STS, right Heschl's gyrus, and left SMG were found (Table A1B), similar to what has been observed for the contrast of interest (Table 2B).

Pattern of Activation in the Visual Motion-Sensitive Areas

As the auditory perception of the syllables was modified by manipulating their visual component (see Table 1), we expected to find modulation of neural activation in visual motion sensitive areas. As neither of the contrasts (≠A≠P > ≠A=P) and (=A≠P > =A=P) revealed activity modulations in these regions (see Table 2), we considered the inverse relations (≠A=P > ≠A≠P) and (=A=P > =A≠P). Results are listed in Table A2 (voxel-level threshold: $p < .001$ uncorrected, and minimum cluster size of 100 mm^3 were used). In both cases, we found that motion-sensitive areas exhibited less activity when a visual change occurred than when the visual input remained unchanged throughout the syllable sequence. [One will notice that these results do not fit with the widely accepted conception that repeated processing of a stimulus is associated with decreased neural activity in regions involved in the processing of the stimulus as suggested by the repetition–suppression effects in fMRI (Grill-Spector et al., 2006).] More precisely, the first comparison (≠A=P > ≠A≠P; Table A2A), contrasting two conditions both involving an acoustic change, further revealed a modulation of activity in the left superior occipital cortex centered on the transverse occipital sulci (likely homolog of V3A or V7; Medendorp, Goltz, Crawford, & Vilis, 2005; Schluppeck, Glimcher, & Heeger, 2005; Tootell et al., 1998) as well as a cluster in the left cuneus. More anterior, modulated activity was also observed in the left superior parietal cortex and the angular gyrus bilaterally. Finally, the left superior frontal area and Crus1 of the cerebellum also showed different levels of activity. In general, the areas indicated in this analysis are consistent with those involved in the processing of visual motion and biological motion. Note how (see top panel of Table 1), in condition ≠A≠P, /da/ and /ba/ are visually clearly different, whereas /da/ and /ga/, in ≠A=P, look similar because the change in the articulation of the tongue is not visible in the video. Indeed, the fusion McGurk effect depends critically on this similarity.

In the second comparison in which concurrent acoustic changes were absent (=A=P > =A≠P; Table A2B), we found two large clusters in the left hemisphere: one located in the lingual cortex and another one in the middle temporal cortex corresponding to the motion sensitive areas V5/MT. The larger effect observed for the latter relation might be explained by the fact that in the first contrast (≠A=P > ≠A≠P), the baseline condition is visually similar but not identical (in ≠A≠P, /da/ and /ga/ are visually similar, but not identical), whereas in the latter comparison (=A=P > =A≠P), in condition =A=P all stimuli are visually identical. The remaining changes in activation were predominantly located in the left hemisphere, with a relatively large cluster in the superior parietal cortex, and smaller clusters in the left frontal cortex. Right hemisphere responses to changes in visual information were focused in the middle frontal gyrus. In general, the areas indicated in this analysis are consistent with those involved in the processing of visual motion and in the detection of intention and emotion from facial movements (e.g., Craig, 2009).

Acknowledgments

This research was supported by grants from the CNRS (Centre National de la Recherche Scientifique). We thank L. Pézard, F.-X. Alario, J.-L. Anton, B. Burle, and L. Casini for helpful discussions. We are also especially grateful to M. Roth for her help with scanning the participants.

Reprint requests should be sent to Nicole Malfait, Institut de Neurosciences de la Timone, UMR 7289, CNRS-Aix Marseille Université, Campus Santé Timone, 27, Boulevard Jean Moulin, 13385 Marseille Cedex 5, France, or via e-mail: nmalfait@gmail.com.

REFERENCES

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30,* 2414–2417.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience, 8,* 14301–14310.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience, 7,* 295–301.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21,* 300–304.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16,* 805–816.

Calvert, G. A., Bullmore, E., Brammer, M. J., Campbell, R., Woodruff, P., McGuire, P., et al. (1997). Activation of auditory cortex during silent speechreading. *Science, 276,* 593–596.

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society, 363,* 1001–1010.

Casile, A., Caggiano, V., & Ferrari, P. F. (2011). The mirror neuron system: A fresh view. *Neuroscientist, 17,* 524–538.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115,* 1989–2000.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113,* 495–506.

Craig, A. D. (2009). How do you feel-Now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10,* 59–70.
Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., et al. (2006). Language control in the bilingual brain. *Science, 312,* 1537–1540.
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91,* 176–180.
Dodd, B. (1977). The role of vision in the perception of speech. *Perception, 6,* 31–40.
Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature, 384,* 159–161.
Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., et al. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex, 14,* 11–22.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 360,* 815–836.
Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10,* 14–23.
Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12,* 711–720.
Habib, M., Daquin, G., Milandre, L., Royere, M., Rey, M., Lanteri, A., et al. (1995). Mutism and auditory agnosia due to bilateral insular damage: Role of the insula in human communication. *Neuropsychologia, 33,* 327–339.
Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: Beyond sensory representation. *Neuron, 56,* 1116–1126.
Hénaff, M. A., Bayle, D., Krolak-Salmon, P., & Fonlupt, P. (2010). Cortical dynamics of a self driven choice: A MEG study during a card sorting task. *Clinical Neurophysiology, 121,* 508–515.
Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4,* 131–138.
Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16,* 307–354.
Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *Journal of Neuroscience, 29,* 10153–10159.
Kislyuk, D. S., Mottonen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience, 20,* 2175–2184.
Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, 20,* 1434–1448.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.
Liberman, A., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.
Lieberman, P., Kako, E., Friedman, J., Tajchman, G., Feldman, L. S., & Jiminez, E. B. (1992). Speech production, syntax comprehension, and cognitive deficits in Parkinson's disease. *Brain and Language, 43,* 169–189.
Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79,* 124–145.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.
McKenna Benoit, M., Raij, T., Lin, F.-H., Iiro, P., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping, 31,* 526–538.
Medendorp, W. P., Goltz, H. C., Crawford, J. D., & Vilis, T. (2005). Integration of target and effector information in human posterior parietal cortex for the planning of action. *Journal of Neurophysiology, 93,* 954–962.
Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13,* 417–425.
Näätänen, R., Kujala, T., & Winkler, I. (2011). Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. *Psychophysiology, 48,* 4–22.
Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118,* 2544–2590.
Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25,* 333–338.
Paulesu, P., Frith, C. D., Bench, C. J., & Bottini, G. (1993). Functional anatomy of working memory: The articulatory loop. *Journal of Cerebral Blood Flow and Metabolism, 13,* 551.
Paus, T., Perry, D. W., Zatorre, R. J., Worsley, K., & Evans, A. C. (1996). Modulation of cerebral blood-flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. *European Journal of Neuroscience, 8,* 2236–2246.
Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, I. P., Kujala, T., et al. (2005). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: An fMRI study at 3 T. *Neuroimage, 29,* 797–807.
Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Möttönen, R., Tarkiainen, A., et al. (2004). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *NeuroReport, 16,* 125–128.
Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCathy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18,* 2188–2199.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2,* 79–87.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.
Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45,* 587–597.
Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Louassmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movement modifies activity in the auditory cortex. *Neuroscience Letters, 127,* 141–145.
Schluppeck, D., Glimcher, P. W., & Heeger, D. J. (2005). Topographic organization for delayed saccades in human posterior parietal cortex. *Journal of Neurophysiology, 94,* 1372–1384.
Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12,* 106–113.
Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25,* 76–89.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2006). Lending a helping hand to hearing: Another motor theory of speech perception. In M. A. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 250–285). Cambridge, MA: Cambridge University Press.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology, 43,* 482–489.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Walthen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88–102). Cambridge, MA: MIT Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Tootell, R. B., Hadjikhani, N., Hall, E. K., Marrett, S., Vanduffel, W., Vaughan, J. T., et al. (1998). The retinotopy of visual spatial attention. *Neuron, 21,* 1409–1422.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.

von Holst, E., & Mittelstaedt, H. (1950). The reafference principle. Interaction between the central nervous system and the periphery. In *Selected papers of Erich von Holst: The behavioural physiology of animals and man* (pp. 39–73). London: Methuen.

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A Neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience, 32,* 3665–3678.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.