'Hearing tongue and seeing voices': neural correlates of audio-visuo-lingual speech perception.

Avril Treille[1], Coriandre Vilain[1], Thomas Hueber[1], Jean-Luc Schwartz[1],
Laurent Lamalle[2], Marc Sato[1]

[1]*GIPSA-lab, Département Parole & Cognition, CNRS & Grenoble Université, Grenoble, France*
[2]*Inserm US 17 / UMS IRMaGE, CHU de Grenoble / CNRS UMS 3552, Unité IRM 3T Recherche, Grenoble, France*

avril.treille@gipsa-lab.grenoble-inp.fr, coriandre.vilain@gipsa-lab.grenoble-inp.fr, thomas.hueber@gipsa-lab.grenoble-inp.fr, jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr, laurent.Lamalle@ujf-grenoble.fr, marc.sato@gipsa-lab.inpg.fr

Abstract

The present fMRI study examined the neural substrates of auditory, visual and audio-visual speech perception in relation to either labial or lingual movements (acquired with a camera and an ultrasound system). Common overlapping activities between modalities were mainly observed in the posterior part of the left superior temporal gyrus/sulcus as well as in the premotor cortex and inferior frontal gyrus. Stronger activity of the premotor and somatosensory cortices was observed during the observation of lingual compared to labial speech movements. Conversely, greater activation of the visual and auditory cortices was observed for labial movements. Altogether these results suggest that audio-visuo-labial and audio-visuo-lingual speech perception recruit a common sensory-motor neural network and are partly driven by the listener's knowledge of speech production.

Keywords: audio-visual speech perception, ultrasound, fMRI.

1. Introduction

Audio-visual speech perception is a special case of multisensory processing that interfaces with the linguistic system. In face-to-face interaction, visual cues from the speaker's face can benefit the listener, notably by improving speech perception in noise or the understanding of a semantically complex statement or a foreign language (Sumby and Pollack, 1954; Reisberg et al, 1987; Navarra et al., 2005). Conversely, seeing incongruent articulatory gestures may also modify auditory speech perception (McGurk & McDonald, 1976).

At the brain level, audio-visual speech perception is known to rely on both primary and associative auditory and visual regions (Calvert et al., 1997; 2000). Because an enhancement of neural responses to audio-visual compared to unimodal speech inputs has been observed in the posterior part of the left superior temporal gyrus/sulcus, it has been proposed that the acoustic and visual speech signals are integrated in this multisensory region, and that modulation of activity within sensory-specific brain areas might partly be caused by backward projections and would represent the physiological correlates of the perceptual changes experienced after audio-visual speech integration (Calvert et al., 2000). In addition, audio-visual speech integration might partly be mediated not only by sensory-specific and multisensory brain regions but also by the speech motor system (including the posterior part of the inferior frontal gyrus and the adjacent ventral premotor cortex), with increased motor activity observed during audio-visual compared to unimodal auditory and visual speech perception (Skipper et al., 2005; 2007), as well as during audio-visual speech perception under adverse listening or viewing conditions (Callan et al., 2003; 2004).

From these studies, one unanswered issue is whether cross-modal speech interactions only depend on well-known auditory and visuo-facial modalities (in relation to labial movements of a speaker) or, rather, might also be triggered by less familiar visual modalities. In the present fMRI study, we examined the neural substrates of auditory, visual and audio-visual speech perception in relation to either labial or lingual movements (acquired with a camera and an ultrasound system, respectively). Since labial and lingual biological speech movements naturally exhibit temporal proximity with auditory speech inputs, evidence for cross-modal speech interactions in relation to both lip and tongue movements would strengthen the hypothesis that multisensory speech perception is partly driven by the listener's knowledge of speech production (Skipper et al., 2007; Schwartz et al., 2012).

2. Methods

2.1. Participants

Twelve healthy adults, native French speakers, participated in the study. All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders.

2.2. Stimuli

Multiple utterances of /pa/, /ta/ and /ka/ syllables were individually recorded by one male and one female speakers in a sound-proof room. Synchronous recordings of auditory, visual and ultrasound signals were acquired by a Terason T3000 ultrasound system (Hueber et al., 2008; see Figure 1) including a 140° microconvex transducer with 128 elements (tongue movements acquired with a sampling rate of 60 fps with a 640x480 pixel resolution), an industrial USB color camera (facial movements acquired with a sampling rate of 60 fps with a 640x480 pixel resolution) and an external microphone connected to the built-in soundcard of the T3000 ultrasound system (audio digitizing at 44.1 kHz).

Two clearly articulated /pa/, /ta/ and /ka/ tokens were selected per speaker (with the speaker initiating each utterance from a neutral mid-open mouth position). Sixty stimuli were created

consisting of twelve /pa/, /ta/ and /ka/ syllables related to five conditions: an auditory condition (A), two visual and two audio-visual conditions related to either lip or tongue movements of a speaker (V_L, V_T, AV_L, AV_T).

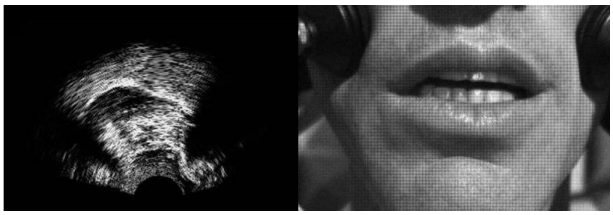

Figure 1: *An example of lip (left) and tongue (right) visual stimuli*.

2.3. Procedure

Before the fMRI session, participants were first presented with a subset of the recorded speech stimuli, with short explanations on the ultrasound system and on the tongue movements required for the production of /pa/, /ta/ and /ka/ syllables. Then participants underwent a three-alternative forced-choice identification task, being instructed to categorize as quickly as possible each perceived syllable with their right hand. The experiment consisted on 60 trials presented in a randomized sequence, with 12 trials related to each modality of presentation (A, V_L, V_T, AV_L, AV_T). The intertrial was of 3s and the response key designation was fully counterbalanced across participants.

The fMRI session consisted of one anatomical scan and one functional run. During the functional run, participants were instructed to passively listen-to and/or watch speech stimuli presented in five different modalities (A, V_L, V_T, AV_L, AV_T). There were 144 trials, with a 8s intertrial, consisting of 24 trials for each modality of presentation and to a resting condition without any sensory stimulation.

2.4. Data acquisition

Magnetic resonance images were acquired with a 3T whole-body MR scanner (Philips Achieva TX). Participants were laid in the scanner with head movements minimized with a standard birdcage 32 channel head coil and foam cushions. Visual stimuli were presented using Presentation software (Neurobehavioral Systems, Albany, USA) and displayed on a screen situated behind the scanner via a mirror placed above the subject's eyes. Auditory stimuli were presented through the MR-confon audio system (www.mr-confon.de).

A high-resolution T1-weighted whole-brain structural image was acquired for each participant before the functional run (MP-RAGE, sagittal volume of 256x224x176mm^3 with a 1mm isotropic resolution, inversion time = 900ms, two segments, segment repetition time = 2500ms, segment duration = 1795ms, TR/TE = 16/5 in ms with 35% partial echo, flip angle = 30°).

Functional images were obtained in a subsequent functional run using a T2*-weighted, echo-planar imaging (EPI) sequence with whole-brain coverage (TR = 8s, acquisition time = 3000ms, TE = 30ms, flip angle = 90°). Each functional scan comprised fifty-three axial slices parallel to the antero-posterior commissural plane acquired in a non-interleaved order (72x72 matrix; field of view: 216mm; 3x3mm^2 in plane resolution with a slice thickness of 3mm without gap). In order to reduce acoustic noise, a sparse sampling acquisition was used (Gracco et al., 2005). This acquisition technique is based on neurophysiological properties of the slowly rising hemodynamic response, which is estimated to occur with a 4–6s delay in case of speech perception and production (Grabski et al., 2013). In the present study, functional scanning therefore occured only during a fraction of the TR, alternating with silent interscanning periods, where stimuli were presented. The time interval between each stimulus onset and the midpoint of the following functional scan acquisition was set at 5s. All conditions were presented in a pseudorandom sequence. Altogether, 144 functional scans were therefore acquired ((5 perceptual conditions + 1 baseline) x 24 trials). In addition, three 'dummy' scans at the beginning of the functional run were added to allow for equilibration of the MRI signal and were removed from the analyses.

2.5. Data analyses

2.5.1. Behavioral analysis

For each participant and modality, the percentage of correct responses and mean reaction-times (RTs), from the onset of the acoustic syllables, were computed. For each dependent variable, a repeated-measures ANOVA was performed with the modality (A, V_L, V_T, AV_L, AV_T) as the within-subjects variable. For both analyses, the significance level was set at $p = .05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, posthoc analyses were conducted with Newman-Keuls tests.

2.5.2. fMRI analysis

fMRI data were analyzed using the SPM8 software package (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) running on Matlab (Mathworks, Natick, MA, USA). Brain activated regions were labeled using the SPM Anatomy toolbox (Eickhoff et al., 2005) and, if a brain region was not assigned or not specified in the SPM Anatomy toolbox, using the Talairach Daemon software (Lancaster et al., 2000).

For each participant, the functional series were first realigned by estimating the six movement parameters of a rigid-body transformation in order to control for head movements between scans. After segmentation of the T1 structural image and coregistration to the mean functional image, all functional images were spatially normalized into standard stereotaxic space of the Montreal Neurological Institute (MNI) using segmentation parameters of the T1 structural image. All functional images were then smoothed using a 6mm full-width at half maximum Gaussian kernel, in order to improve the signal-to-noise ratio and to compensate for the anatomical variability among individual brains.

For each participant, neural activations related to the perceptual conditions were analyzed using a General Linear Model, including five regressors of interest (A, V_L, V_T, AV_L, AV_T) and the six realignment parameters as covariates of no-interest, with the silent trials forming an implicit baseline. The BOLD response for each event was modeled using a single-bin finite impulse response (FIR) basis function spanning the time of acquisition (3s). Before estimation, a high-pass filtering with a cutoff period of 128s was applied. Beta weights associated with the modeled FIR responses were then computed to fit the observed BOLD signal time course in each voxel for each condition. Individual statistical maps were calculated for each perceptual condition with the related baseline and subsequently used for group statistics.

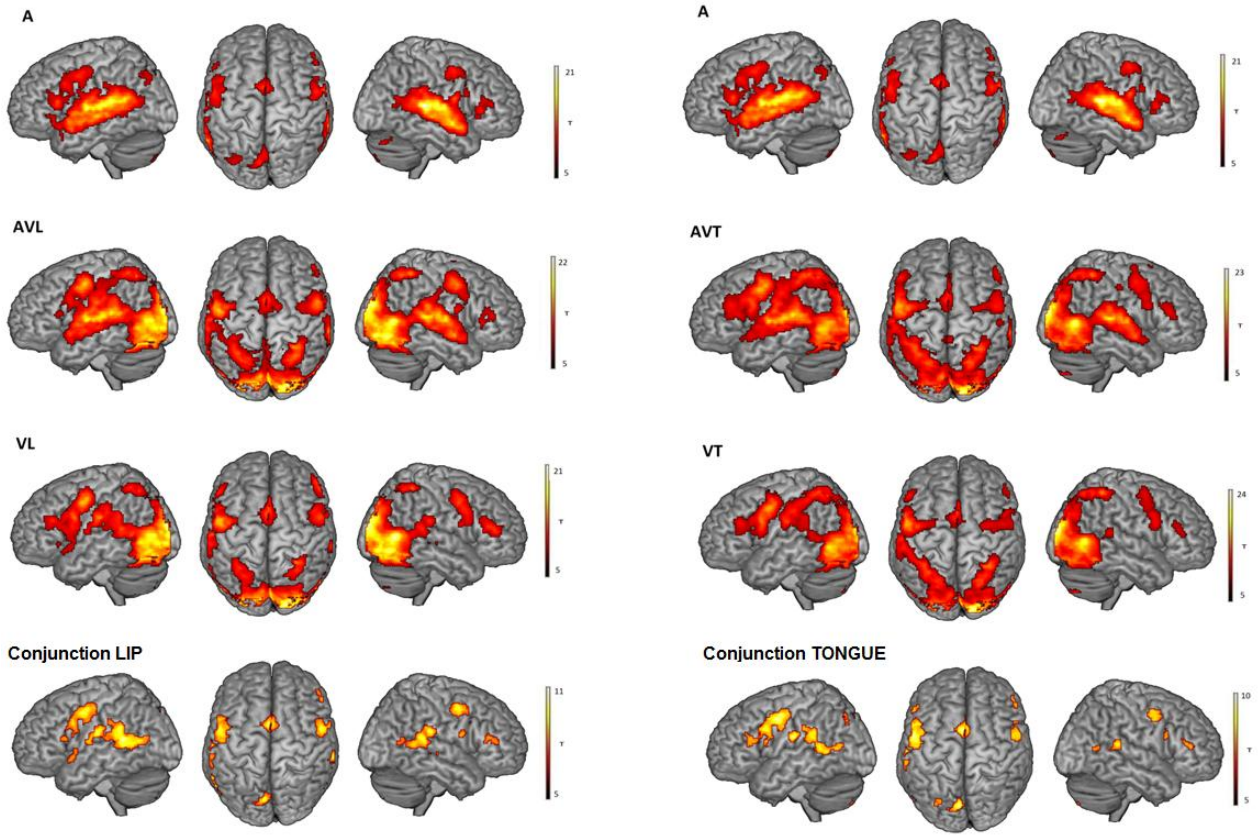

Figure 2: *Overlapping activity between auditory, visual and audio-visual modalities in relation to lip (L) and tongue (T) movements.*

In order to draw population-based inferences, a second-level random effect group analysis was carried-out with the modality (A, V_L, V_T, AV_L, AV_T) as the within-subjects variable and the subjects treated as a random factor. In order to determine common neural activity related to auditory, visual and audio-visual speech perception, in relation to lip and tongue movements, two conjunction analyses were separately performed (i.e., $A \cap V_L \cap AV_L$ and $A \cap V_T \cap AV_T$). Then, an analysis by modality was conducted in order to determine which regions were more activated during lip compared to tongue movements, and vice-versa (i.e., $V_L \neq V_T$ and $AV_L \neq AV_T$).

All contrasts were calculated with a significance level set at p = .05, family-wise-error (FWE) corrected at the voxel level with a cluster extent of at least 30 voxels.

3. Results

3.1. Behavioral results

Overall, the mean proportion of correct responses was of 81%. The main effect of modality was significant ($F(4,44) = 38.1$, $p < .001$), with more correct responses in the A, AV_L, AV_T conditions than in the V_L condition, and in the V_L condition than in the V_T condition (on average, A: 99%, AVL: 98%, AVT: 94%, VL: 69%, VT: 47%).

For RTs, a significant effect of the modality was also observed ($F(4,44) = 18.2$, $p < .001$), with faster RTs in the AV_L condition than in the AV_T and V_L conditions, and in the AV_T and V_L conditions than in the V_T condition (on average, A: 837ms, AV_L: 732ms, AV_T: 926ms, V_L: 984ms, V_T: 1187ms).

3.2. fMRI results

3.2.1. Conjunction analyses - see Figure 2

The conjunction analysis on A, V_L and AV_L conditions demonstrates common activity in the posterior part of the superior temporal gyrus/sulcus (pSTG/STS), extending rostrally to the Heschl's gyrus and insular cortex, ventrally to the posterior middle temporal gyrus (MTG) and dorsally to the parietal operculum and the ventral part of the supramarginal (SMG) and angular gyri (AG). Common neural responses were also observed in the premotor cortex, the inferior frontal gyrus (pars opercularis and right pars triangularis), the middle frontal gyrus and the left primary sensorimotor cortex. Additional activity was found in the cerebellum, the supplementary motor area (SMA) and adjacent anterior cingulate cortex, and the precuneus.

Similarly, the conjunction analysis on A, V_T and AV_T conditions demonstrates common activity in pSTG/STS, extending ventrally to the left posterior MTG and dorsally to SMG, AG and the left parietal operculum. Common neural responses were also observed in the premotor cortex, the inferior frontal gyrus (pars opercularis and right pars triangularis), the middle frontal gyrus, the insular cortex and the left primary sensorimotor cortex. Additional activity was found in the cerebellum, the SMA and adjacent anterior cingulate cortex, the precuneus, and the associative extrastriate visual cortex.

3.2.2. Analyses by modality - see Figure 3

$V_L \neq V_T$: Compared to tongue movements, audio-visuo-labial speech perception induced a greater activation of the auditory

(including the Heschl's gyrus, the temporopolar area, pSTG/STS, MTG) and visual cortices (from the primary visual cortex extending to the extrastriate visual cortex and to the dorsal part of the cerebellum). Greater activity of frontal regions was also evident (middle frontal and dorsolateral prefrontal cortices), especially in the right hemisphere. Conversely, the audio-visuo-lingual speech perception entailed greater activity in motor and premotor cortices as well as in parietal regions (including parts of the sensorimotor cortex, intraparietal sulcus, inferior and superior parietal cortices).

$AV_L \neq AV_T$: Compared to tongue movements, audio-visuo-labial speech perception induced a greater activation of the visual cortex (from the primary visual cortex extending to the right extrastriate visual cortex and to the dorsal part of the cerebellum). Conversely, the audio-visuo-lingual speech perception entailed greater activity in left frontal (premotor and prefrontal cortices), parietal (including parts of the intraparietal sulcus and the surrounding inferior and superior parietal cortices) and auditory areas (pSTG/STS) as well as in the bilateral ventral cerebellum.

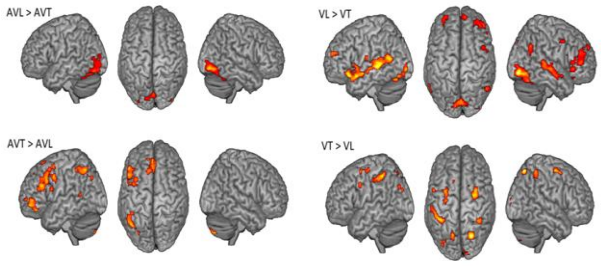

Figure 3: *Activity differences between lip (L) and tongue (T) visual and audio-visual conditions.*

4. Discussion

The present fMRI study examined the neural substrates of cross-modal binding during audio-visual speech perception in relation to both facial and tongue movements. Our results first demonstrate for both labial and lingual stimuli, common overlapping activity between modalities in the posterior part of the superior temporal gyrus/sulcus. These results appear in line with previous studies indicating a key role of this region in biological motion perception (including face perception), speech processing and audio-visual integration (Calvert et al., 1997, 2000; Beauchamps et al., 2004). In addition, while more activity in visual and auditory cortices was observed during the perception of lip movements, more activity was observed in motor and somatosensory areas during the perception of tongue movements. This latter result likely indicates that participants simulated, covertly or even overtly, the motor consequence of the perceived actions, particularly for less familiar visual information as in the case of lingual movements. Altogether, these results suggest that audio-visuo-labial and audio-visuo-lingual speech perception recruit a common sensory-motor neural network and are partly driven by the listener's knowledge of speech production (Skipper et al., 2007; Schwartz et al., 2012).

5. References

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H and Martin, A. (2004). "Unraveling multisensory integration: patchy organization within human STS multisensory cortex". In: Nature Neuroscience 7.11, pp. 1190-1192.

Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. and Vatikiotis-Bateson, E. (2003). "Neural processes underlying perceptual enhancement by visual speech gestures". In: *NeuroReport* 14, pp. 2213-2217.

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

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D. and David, A.S. (1997). "Activation of auditory cortex during silent lipreading".In: *Science* 276, pp. 593-596.

Calvert, G.A., Campbell, R. and Brammer, M.J. (2000). "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex". In: *Current Biology* 10.11, pp. 649-657.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., et al. (2005). "A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data". In: *NeuroImage* 25, pp. 1325–1335.

Grabski, K., Schwartz, J.-L, Lamalle, L., Vilain, C., Vallée, N., Baciu, M. Le Bas, J.-F and Sato, M. (2013). "Shared and distinct neural correlates of vowel perception and production". In: *Journal of Neurolinguistics* 26.3, pp. 384-408.

Gracco, V.L., Tremblay, P. and Pike, G.B. (2005). "Imaging speech production using fMRI". In: *NeuroImage* 26, pp. 294–301.

Hueber, T., Chollet, G., Denby, B., and Stone, M. (2008). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application,". In: *Proceedings of International Seminar on Speech Production (Strasbourg, France),* pp. 365-369.

Lancaster, J.L.,Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., et al. (2000). "Automated Talairach atlas labels for functional brain mapping." In: *Human Brain Mapping* 10.3, pp. 120-131.

McGurk, H. and MacDonald, J. (1 976)."Hearing lips and seeing voices".In: *Nature* 264, pp.746-748.

Navarra, J. and Soto-Faraco, S. (2005). "Hearing lips in a second language: visual articulatory information enables the perception of second language sounds". In: *Psychological research* 71.1, pp. 4-12.

Reisberg, D., McLean, J. and Goldfield, A. (1987). "Easy to hear but hard tounderstand: a lipreading advantage with intact auditory stimuli". In: *Campbell, R., Dodd, B. (Eds.), Hearing by Eye: The Psychology of Lipreading. Lawrence Erlbaum Associates,* London (UK), pp. 97–113.

Schwartz, J.L., Ménard, L., Basirat, A. and Sato, M. (2012). "The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception". In: *Journal of Neurolinguistics* 25.5, pp. 336-354.

Skipper, J.I., Nusbaum, H.C. and Small, S.L. (2005) "Listening to talking faces: motor cortical activation during speech perception". In: *NeuroImage* 25, pp.76–89.

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C. and Small, S.L. (2007)."Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception". In: *Cerebral Cortex* 17.10, pp. 2387-2399.

Sumby, W.H. and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise". In: *Journal of Acoustical Society of America* 26, pp. 212–215.