The sight of your tongue: neural correlates of audio-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] UJF, UMS IRMaGE, CHU de Grenoble, Unité IRM 3T Recherche, INSERM, Grenoble, France

avril.treille@gipsa-lab.inpg.fr, marc.sato@gipsa-lab.inpg.fr

Abstract

While functional neuroimaging studies demonstrate that multiple cortical regions play a key role in audio-visual integration of speech, whether cross-modal speech interactions only depend on well-known auditory and visuo-facial modalities or, rather, might also be triggered by other sensory sources remains unexplored. The present functional magnetic resonance imaging (fMRI) study examined the neural substrates of cross-modal binding during audio-visual speech perception in response to either seeing the facial/lip or tongue (tongue movement inside the mouth acquired by means of ultrason) movements of a speaker. To this aim, participants were exposed to auditory and/or visual speech stimuli in five different conditions: an auditory-only condition, and two visual-only and two audiovisual conditions that showed either the facial/lip or tongue movements of a speaker. Common overlapping activity between conditions were mainly observed in the posterior part of the superior temporal gyrus/sulcus, extending ventrally to the posterior middle temporal gyrus and dorsally to the parietal operculum, the supramarginal and angular gyri, as well as in the premotor cortex and in the inferior frontal gyrus. In addition, sub-additive neural responses were observed in the left posterior superior temporal gyrus/sulcus during audio-visual perception of both facial and tongue speech movements compared to unimodal auditory and visual speech perception. Altogether these results suggest that the left posterior superior temporal gyrus/sulcus is involved in multisensory processing of auditory speech signals and their accompanying facial/lip and tongue speech movements, and that multisensory speech perception is partly driven by listener's knowledge of speech production.

Index Terms: audio-visual speech perception, ultrasound, fMRI.

1. Introduction

Although humans are proficient at extracting phonetic features from the acoustic signal alone, interactions between auditory and visual modalities are beneficial in daily conversation. Notably, visual information is known to effectively improve speech perception in noise, the understanding of a semantically complex statement or a foreign language [1-3], and may even change our auditory experience in case of a mismatch between the auditory and visual speech signals [4].

At the brain level, functional neuroimaging studies demonstrate that multiple cortical regions play a key role in audio-visual integration of speech. Activity within sensory-specific and multisensory brain regions (including the primary/secondary auditory cortex, the visual motion-sensitive cortex and the posterior part of the left superior temporal gyrus/sulcus) is modulated during audio-visual speech perception, compared to unimodal auditory and visual speech perception [5-10]. Because an enhancement of neural responses (supra-additivity) to audio-visual speech inputs has been observed in the posterior part of the left superior temporal sulcus, it has been proposed that 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 [6]. However, several magneto-encephalographic and electroencephalographic studies challenge this hypothesis by demonstrating that visual speech input modulates activity in the primary and secondary auditory cortices at an early stage in the cortical speech processing hierarchy [11-16]. Based on these results, both early non-phonetic activation of auditory areas depending upon visual motion cues and a later speech-specific left-lateralized response mediated by backward projections from the multisensory integrative brain areas have also been suggested [12,14]. In addition, apart from sensory-specific and multisensory brain regions, other studies suggest that audio-visual speech integration might partly be mediated 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 [9-10], as well as during audio-visual speech perception under adverse listening or viewing conditions [7-8].

Based on these studies, one fundamental issue is whether cross-modal speech interactions only depend on well-known auditory and visuo-facial modalities (i.e., facial movements of a speaker) or, rather, might also be triggered by other sensory sources. The present fMRI study aimed at determining the neural correlates of cross-modal speech interactions in relation to facial but also to tongue movements (tongue movement inside the mouth acquired by means of ultrasound acquisition) of a speaker. Since facial/lip and tongue biological speech movements naturally exhibit temporal proximity with auditory speech inputs, evidence for cross-modal speech interactions in relation to both facial and tongue movements would strength the hypothesis that multisensory speech perception is partly driven by listener's knowledge of speech production [10,17].

2. Method

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. The protocol was approved by the Grenoble University Ethical Committee with all participants

screened for neurological, psychiatric, other possible medical problems and contraindications to MRI.

2.2. Stimuli

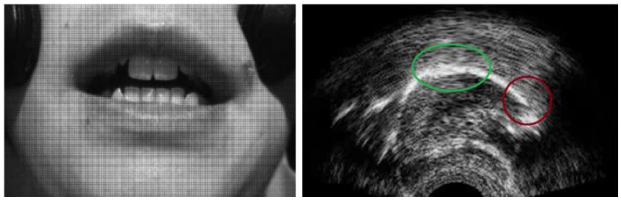

Figure 1: *Examples of facial and tongue visual stimuli (tongue tip and dorsum are highlighted respectively in red and green circles).*

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 including a 140° microconvex transducer with 128 elements (tongue movements acquired with a sampling rate of 60 fps with a 640x480 pixels resolution), an industrial USB color camera (facial movements acquired with a sampling rate of 60 fps with a 640x480 pixels resolution) and an external microphone connected to the built-in soundcard of the T3000 ultrasound system (audio digitizing at 44.1 kHz) [18].

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 5 conditions: an auditory condition (A), two visual conditions related to either facial or tongue movements of a speaker (V_L, V_T), two audio-visual conditions including either facial or tongue movements of a speaker (AV_L, AV_T).

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 during the production of /pa/, /ta/ and /ka/ syllables. They then underwent a three-alternative forced-choice identification task, with participants 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 5 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. fMRI 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 anteroposterior commissural plane acquired in 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 [19]. 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 [20]. In the present study, functional scanning therefore occurs 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 analyses

For each participant and modality, the percentage of correct responses and mean 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 analyses

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 [21] and, if a brain region was not assigned or not specified in the SPM Anatomy toolbox, using the Talairach Daemon software [22]. For visualization, activation maps were superimposed on a standard brain template using the MRICRON software (http://www.sph.sc.edu/comd/rorden/mricron/).

The first three volumes ('dummy' scans) were discarded. For each participant, the functional series were first realigned by

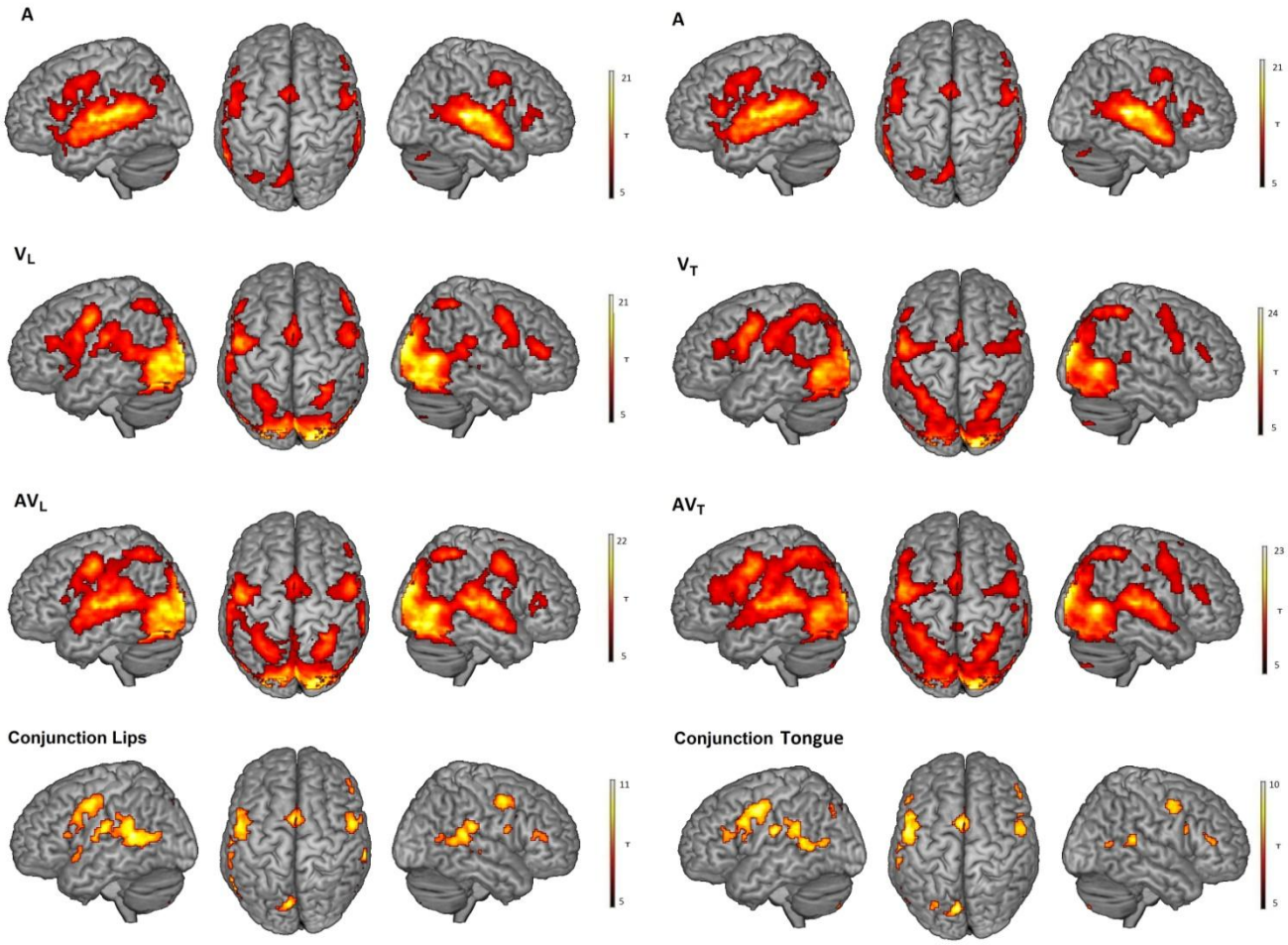

Figure 2: *Surface rendering of brain regions activated in A, V_L, V_T, AV_L and AV_T conditions and showing overlapping activity between A, V_L and AV_L and between A, V_T and AV_T condition (conjunction analyses). All contrasts are computed from the random-effect group analysis ($p < .05$, FWE corrected, cluster extent threshold of 20 voxels).*

estimating the 6 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 5 regressors of interest (A, V_L, V_T, AV_L, AV_T) and the six realignment parameters, 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.

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 facial and tongue movements, two conjunction analyses were separately performed on A, V_L and AV_L conditions (i.e., $A \cap V_L \cap AV_L$) and on A, V_T and AV_T conditions (i.e., $A \cap V_T \cap AV_T$). In addition, two further analyses were separately performed on A, V_L and AV_L conditions (i.e., $AV_L \neq A + V_L$) and on A, V_T and AV_T conditions (i.e., $AV_T \neq A + V_T$) to determine brain regions showing supra-additive or sub-additive audio-visual responses compared to unimodal auditory and visual responses. 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. Behavorial results

Overall, the mean proportion of correct responses was of 81%. The main effect of modality was significant ($F(4,44) = 38.09$, $p < .001$), with more correct responses in the A, AV_L, AV_T conditions than to the V_L condition, and in the V_L condition than in the V_T condition (on average, A: 99%, AV_L: 98%, AV_T: 94%, V_L: 69%, V_T: 47%).

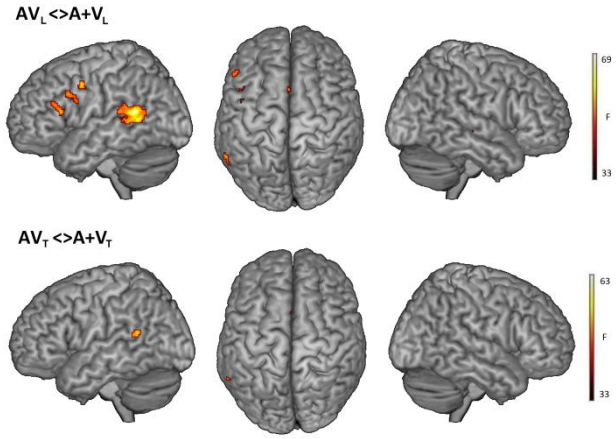

Figure 3: *Surface rendering of brain regions showing activity differences between AV and the sum of A and V conditions. All contrasts are computed from the random-effect group analysis ($p < .05$, FWE corrected, cluster extent threshold of 20 voxels).*

The ANOVA on RTs demonstrate a significant effect of the modality ($F(4,44) = 18.16$, $p < .001$), with faster RTs in the AV_L conditions 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 (i.e., $A \cap V_L \cap AV_L$) 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 (i.e., $A \cap V_T \cap AV_T$) 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. Sub-additive responses (see Figure 3)

Sub-additive AV_L responses compared to the sum of unimodal A and V_L conditions (i.e., $AV_L < A + V_L$) were observed in the left pSTG/STS, extending dorsally to the SMG. Sub-additive AV responses were also found in the left premotor cortex and inferior frontal gyrus (left pars opercularis and triangularis) and in the SMA and adjacent anterior cingulate cortex.

Sub-additive AV_T responses compared to the sum of unimodal A and V_T conditions (i.e., $AV_T < A + V_T$) were also observed in the left pSTG/STS, as well as in the anterior cingulate cortex and in the entorhinal cortex.

Importantly, no supra-additive responses (i.e., $AV > A + V$) were observed in these two analyses.

4. Discussion and Conclusions

The present fMRI study aimed at determining the neural substrates of cross-modal speech interactions in relation to facial and tongue movements of a speaker.

In relation to both facial and tongue movements, our results first demonstrate common overlapping activity during auditory, visual and audio-visual speech perception in the pSTS and in adjacent regions on the surface of STG/MTG and SMG/AG. These results appear exquisitely in line with previous studies indicating a key role of pSTS in biological motion perception (including face perception), speech processing and audio-visual integration.

Importantly, several criteria have been proposed to determine brain areas involved in the audio-visual integration of speech [6]: they should respond to both unimodal auditory and visual speech stimuli, and show either supra-additive AV responses in case of congruent audio-visual speech inputs or, conversely, sub-additive AV responses in case of incongruent audio-visual speech inputs. In our study, although facial and tongue visual stimuli were presented in natural synchrony with auditory speech stimuli, we however observed sub-additive AV responses in this region. One tentative explanation might come from the strong activity observed in several motor brain areas, notably in the inferior frontal gyrus and the premotor cortex. This later result likely indicates that participants mentally simulate the motor consequence of the perceived actions. Backward projections from speech motor regions to the left STS (in the form of efference copy) might have then influenced audio-visual speech binding thought to occur in the left pSTS. This hypothesis, although highly speculative, would suggest that multisensory speech perception in relation to facial and tongue movements is partly driven by listener's knowledge of speech production [10,17].

5. References

[1] Sumby, W. H., and Pollack, I. "Visual contribution to speech intelligibility in noise". Journal of Acoustical Society of America 26:212-215, 1954.

[2] Reisberg, D., McLean, J. and Goldfield, A. "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, 1987.

[3] Navarra, J. and Soto-Faraco, S. “Hearing lips in a second language: visual articulatory information enables the perception of second language sounds”. Psychological research 71(1):4-12, 2005.

[4] McGurk, H. and MacDonald, J. “Hearing lips and seeing voices”. Nature 264:746-748, 1976.

[5] 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. “Activation of auditory cortex during silent lipreading”. Science 276:593-596, 1997.

[6] Calvert, G.A., Campbell, R. and Brammer, M.J. “Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex”. Current Biology 10(11): 649-657, 2000.

[7] Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. and Vatikiotis-Bateson, E. “Neural processes underlying perceptual enhancement by visual speech gestures”. NeuroReport, 14:2213-2217, 2003.

[8] Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. and Vatikiotis-Bateson, E. “Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information”. Journal of Cognitive Neuroscience 16:805-816, 2004.

[9] Skipper, J.I., Nusbaum, H.C. and Small, S.L. “Listening to talking faces: motor cortical activation during speech perception”. NeuroImage 25:76–89, 2005.

[10] Skipper, J.I., van Wassenhove, V., Nusbaum, H.C. and Small, S.L. “Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception”. Cerebral Cortex 17(10):2387-2399, 2007.

[11] Sams, M., Aulanko, R., Ha¨ma¨la¨inen, M., Hari, R., Lounasmaa, O.V., Lu, S.T. and Simola, J. “Seeing speech: visual information from lip movements modifies activity in the human auditory cortex”. Neuroscience Letters 127:141-145, 1991.

[12] Klucharev, V., Möttönen, R. and Sams, M. “Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception”. Brain Res. Cogn. Brain Res. 18:65–75, 2003.

[13] van Wassenhove, V., Grant, K.W. and Poeppel, D. “Visual speech speeds up the neural processing of auditory speech”. Proceedings of the National Academy of Sciences U.S.A., 102:1181-1186, 2005.

[14] Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H. and Ackermann, H. “Sequential audiovisual interactions during speech perception: a whole-head MEG study”. Neuropsychologia 45(6):1342–1354, 2007.

[15] Stekelenburg, J.J. and Vroomen, J. “Neural correlates of multisensory integration of ecologically valid audiovisual events”. Journal of Cognitive Neuroscience, 19:1964–1973, 2007.

[16] Arnal, L.H. and Giraud, A.L. “Dual neural routing of visual facilitation in speech processing”. The Journal of Neuroscience, 29(43):13445-13453, 2009.

[17] Schwartz, J.L., Ménard, L., Basirat, A. and Sato, M. “The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception”. Journal of Neurolinguistics, 25(5):336-354, 2012.

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

[19] Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., et al. “Sparse” temporal sampling in auditory fMRI”. Human Brain Mapping, 7(3):213–223, 1999.

[20] Grabski, K., Schwartz, J.-L, Lamalle, L., Vilain, C., Vallée, N., Baciu, M. Le Bas, J.-F and Sato, M. “Shared and distinct neural correlates of vowel perception and production”. Journal of Neurolinguistics, 26(3): 384-408, 2013.

[21] Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., et al. “A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data”. NeuroImage 25:1325–1335, 2005.

[22] Lancaster, J.L.,Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., et al. “Automated Talairach atlas labels for functional brain mapping. Human Brain Mapping”, 10(3):120-131, 2000.