AVSP 2003 - International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Early processing of visual speech information modulates the subsequent processing of auditory speech input at a pre-attentive level: Evidence from event-related brain potential data.

Riadh Lebib, David Papo, Abdel Douiri, Stella de Bode, et Pierre-Marie Baudonniere

Departamento de Psicología Cognitiva
Universidad de La Laguna – Tenerife, España
Neurosciences Cognitives et Imagerie Cerebrale
LENA CNRS UPR 640 – Paris, France

Riadh.Lebib@chups.jussieu.fr

1. Introduction

Among the five senses, vision and audition share very strong anatomical, functional, and cognitive links. This is especially true for speech perception. Visual speech (i.e. speech reading) influences auditory perception at the very early stages of audiovisual speech processing [1]. This could result in an enhancement of speech intelligibility, notably in noisy surroundings [2-4]. However, with incongruent auditory and visual stimuli, audiovisual integration can be a source of perceptual ambiguity and impaired speech comprehension [5], even with perfectly audible acoustic signals [6, 7].

Understanding how bimodal speech information is processed as a unitary and coherent percept is a central question. Classic accounts of the benefits of speechreading to speech recognition treat auditory and visual channels as independent sources of information that are integrated early in the speech perception process, most likely at a pre-categorical level, and possibly at the stage of phonetic categorization [6, 8-10]. However, little is known about the stimulus parameters, processing limitations and perceptual strategies that govern the intermodal perception of speech information.

Recent brain imaging studies have shown that silent speechreading can activate auditory cortex [11, 12]; this supports the assumption that seen speech influences the perception of heard speech at a prelexical stage. Furthermore, electrophysiological data revealed that neural responses elicited to observing lips movements occurred very early after the movement onset [13]. Taking into account that lip movements always precede the acoustic signal onset in normal speech, we questioned whether the brain can distinguish between congruent (i.e. redundant) or non-congruent bimodal sensory input, at a pre-attentive level. In other words, we questioned the possibility of visual-to-auditory cross-modal sensory gating responses, with audiovisual speech stimuli.

As an index of early sensory gating, we used the P50 paradigm, a neural response to auditory stimuli that appears between 35 and 85 ms following auditory stimulation. Sensory gating refers to the ability of the brain to modulate its sensitivity to incoming sensory stimuli [3]. This broad definition allows the concept of gating to include both the capacities of the brain to “gate out” incoming irrelevant sensory input, and to “gate in” significant or novel sensory input [14, 15].

In the present study, we examined the degree of attenuation/augmentation of the sensory response to congruent or non-congruent audiovisual speech stimuli as reflected by the P50 amplitudes in normal volunteers. We also analyse the early response to mouth movement onset, and dipole modeling was conducted to study the specific brain structures activated during early visual speech processing. The specific aims of the study were a) to show that speech processing starts as soon as lip movements occurred, b) to provide data that the brain can detect changes in incoming bimodal speech stimuli at either a pre-attentive or a very early attentive stage of information processing as reflected in the P50 component, and c) to confirm that this early detection is dependent upon the congruency status and the discriminability level of audiovisual speech input. Finally, the goal of this project was to generalize the concept of sensory gating with “real-life” stimuli, that is semantically-relevant and multimodal.

2. Procedure & Method

Stimuli consisted of 4 audiovisual French vowels: [i], [a], [ø], and [y] (phonetic symbols). Audiovisual dubbing was either congruent (C), i.e. vowel pronunciation corresponded to vowel sound, or incongruent (I), i.e. visual speech did not match acoustic signal. Moreover, according to visual speech cues, vowels were either easily distinguishable (E) or hard to distinguish (H). Thus, there were 4 categories of audiovisual speech stimuli (AVS): CE, CH, IE, and IH. Table 1 displays the different audiovisual dubbing we made.

	Visual stimulus			**Visual stimulus**	
Auditory stimulus	**[i]**	**[a]**	**Auditory stimulus**	**[ø]**	**[y]**
[i]	*Congruent*	*Incongruent*	**[ø]**	*Congruent*	*Incongruent*
[a]	*Incongruent*	*Congruent*	**[y]**	*Incongruent*	*Congruent*
	Easy			*Hard*	
	Discriminability level				

Table 1. Audiovisual dubbing for the selected French vowels (phonetic symbols)

The grouping was made on ease of speechreading rather than phonetic parameter, as a preliminary control EEG study run with the same vowels presented acoustically did not exhibit any significant difference (in peak amplitudes, peak latencies, and scalp voltage distribution) for the auditory-related ERP responses, i.e. the P50-N100-P200 complex, between the four selected vowels.

We shot the same person pronouncing these 4 vowels. Each pronunciation was preceded by still face period which duration ranged from 1800 to 2400 ms, and was followed by a varying 2880 to 5080 ms-long pause (black screen). The still face period was intended to avoid contamination of the ERPs of interest by face-specific encoding potentials. So that our dubbing seems naturalistic, the dynamic portion of the visual stimulus (from a resting face position) started before, and finished after the auditory stimulus. For every vowel, duration of the visible speech and the auditory tokens were both homogenized respectively at 880 ms and 240 ms. The vowel sounds always started 320 ms after the lips movement onset. These selected durations were computed after analyzing and averaging those of the 4 vowels.

There were 10 experimental blocks of 32 stimuli each (each category of AVS played 8 times). Within each block, the stimuli order was quasi-randomized. During the experiment, a training block was always played first. Using a double-button press device, subjects judged whether audiovisual stimuli were congruent or not in a delayed response task. This unusual procedure in P50 paradigms was used to keep our subjects alert during the processing of the bimodal speech stimuli. The left and right position of the response-buttons and the sequence of the experimental blocks were counterbalanced across subjects.

The EEG was recorded while subjects (7 females and 6 males, with a mean age of 25.6 years, native French speakers, right-handed, normal auditory function and normal or corrected-to-normal vision, no history of psychiatric or neurological problems) performed the judgement task. Data was recorded from 62 electrodes referenced to the nose, using an Electrocap International montage along the midline at FPz to Iz and homologous positions over the left and right hemispheres. Electrode impedance was kept below 3 kΩ. Blinks, vertical and horizontal eye movements (EOG) were recorded from four bipolar electrodes. The EEG was amplified and filtered (high band-pass: 0.16-100 Hz; 50 Hz notch filter), then digitized at a 500 Hz sampling rate and stored for off-line analyses. Eye movements and blinks were automatically corrected, and the visually detected artifacted EEG epoch were rejected (movement artifacts, amplifier saturation).

For the visual speech analysis, ERPs were time-locked to the onset of the lips' movement. Data were low-pass filtered at a 10 Hz cut-off frequency. ERP data were analyzed by computing mean peak amplitudes and latencies in specific time windows relative to a 200 ms pre-stimulus baseline.

ERP analysis was complemented by spatio-temporal source modeling using the Brain Electrical Sources Analyze software (BESA) [16]. We used a classical four-shell spherical head model for conductive volumes and equivalent current dipoles (ECDs) for generator (local activity of brain regions). The procedure consisted in identifying ECDs leading to the best fit between experimental and model distribution. A non-linear iterative procedure was applied to optimize location and orientation parameters, and a linear least-mean square algorithm determined the time-varying magnitude. The model adequacy was assessed by a goodness-of-fit criterion based on the percentage of residual variance, i.e. the experimental variance unexplained by the model. Then, a mathematical program registered the dipoles in a standard coordinate space [17]. Independent models were separately developed for the grand average data and for the individual subjects' data.

For the bimodal speech analysis, ERPs were time-locked to the onset of the vowel sound. Data were low-pass filtered at a 25 Hz cut-off frequency. Only trials with correct responses were taken into account. ERP data were analyzed by computing mean peak amplitudes and latencies in specific time windows relative to a 200 ms pre-stimulus baseline.

Repeated-measure analyses of variance (ANOVAs) were computed with congruency (congruent/incongruent), discriminability (easy/hard), and electrodes as within-subject factors.

3. Results

Figure 1 displays the ERPs associated with the mouth movement onset. Lips movement onset of the studied vowels elicited a wide spread temporo-parieto- occipital negative-going wave peaking at about 180 ms (N180) after the movement onset (Figure 2). In the time range of this N180 component (analysis conducted on a 150-200 ms time window), the infero-parietal leads P_7 – P_8 and the parieto-occipital leads PO_7 – PO_8 showed the most prominent negative ERP component (mean peak amplitude = - 4.2 µV ± 1.8 µV). The amplitude and the latency of the N180 did not differ in response to the different vowels, and the topographical differences on the scalp distribution across the different vowels were not significant ($P > 0.05$ in all comparisons). The N180 was, however, larger ($F_{1,12} = 17.08$, $MSE = 1.44$, $P = 0.0014$) over the left temporal leads

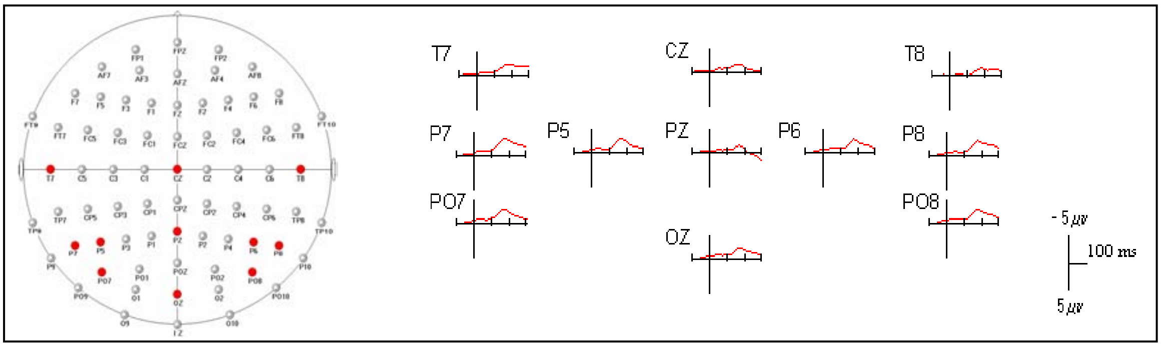
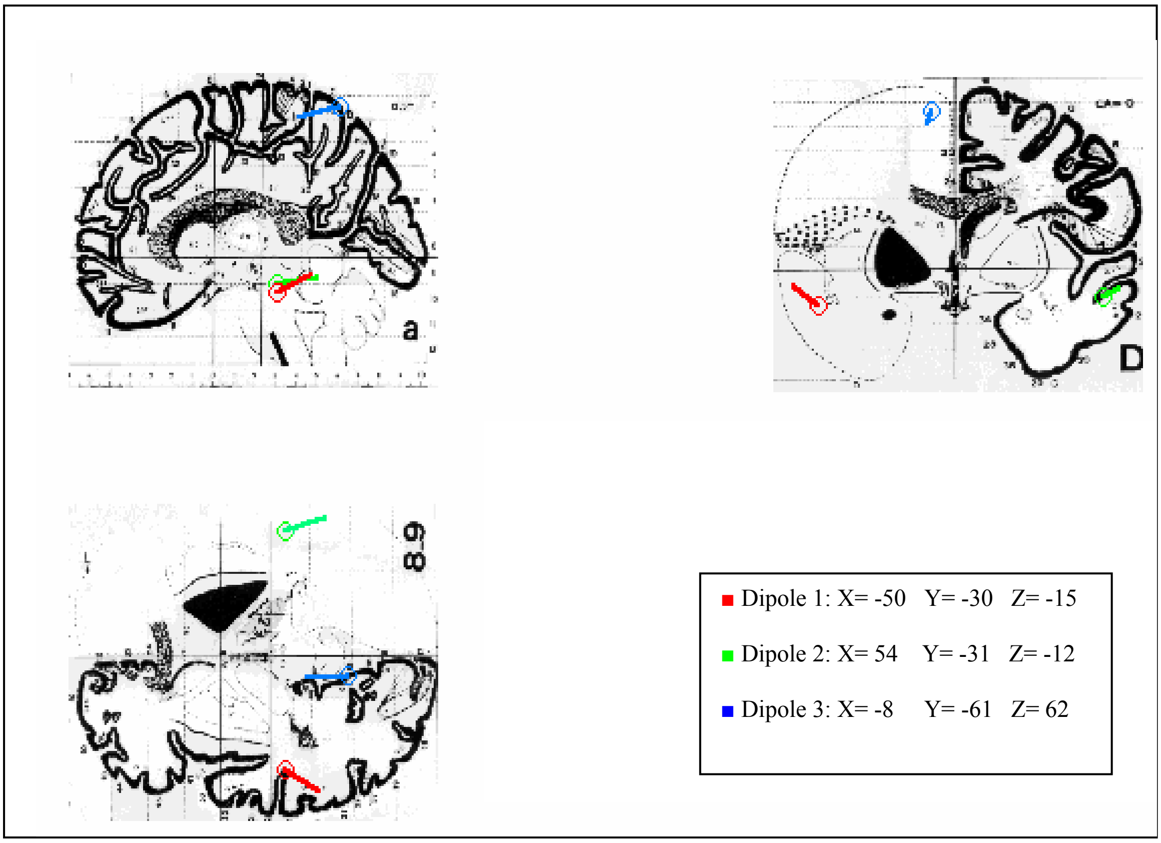


Figure 1. Grand mean ERP wave forms elicited after the lips movement onset (t_0). A negative-going wave peaks at about 180 ms over the temporal, the infero-parietal, and the occipital scalp regions. On the left is a schematic view of the head from above (nose upward), with the position of the selected leads in red.

(T_7, C_5, C_3; mean amplitude = -3.16 µV ± 1.6 µV) than over the right temporal ones (T_8, C_6, C_4; mean amplitude = -2.03 µV ± 1.1 µV). Moreover, the N180 peaked 15 ms later ($F_{1,12}$ = 7.98, MSE = 352, P = 0.0153) over the right occipital leads (PO_8, O_2; mean latency = 189.9 ms ± 5.2 ms) than over the left occipital ones (PO_7, O_1; mean latency = 175.2 ms ± 4.6 ms).

According to further analyzes relative to electrical source localization (BESA software), the best dipolar model fitting the surface potential distribution over the scalp for the grand average data was a 3-dipole model. Figure 2 display the dipoles' location with their respective coordinates in the Talairach space [17]. Dipoles were numbered according to their decreasing dipolar moment, that is respectively 0.12 nA.m, 0.10 nA.m, and 0.04 nA.m. No symmetrical constraint was used in our modeling, and the chosen time window was 60 ms long around the maximum of the N180 signal strength. The fitting model accounts for 95.7 % of the total experimental variance. Two dipoles were bilaterally distributed over the left and right inferior temporal gyrus (Brodmann area 20, or BA20). The third dipole was located in the left superior parietal lobule (BA 7). Based on individual models, intersubject variability of dipole locations was less than 6 mm (+/- SD).

The figure 3 displays the ERP waveforms after the vowel sound onset. As expected, every AVS elicited a P50



🟥 Dipole 1:	X= -50	Y= -30	Z= -15
🟩 Dipole 2:	X= 54	Y= -31	Z= -12
🟦 Dipole 3:	X= -8	Y= -61	Z= 62

Figure 2. Localization of the computed dipoles for the grand average data, transposed in a human encephalon atlas. The circle schematizes the origin of the dipole, while the segment of line schematize its direction. Coordinates are those of the origins, in mm.

component with maximum amplitude over a centro-parietal scalp area. Table 2 shows the means and standard deviations of the P50 peak amplitude at the Cz electrode.

When analysing P50 peak amplitudes over the centro-parietal scalp area (where maximum amplitude was observed), the P50 component was significantly different only for CE *vs*. IE comparison ($F_{1,12} = 9.87$, $MSE = 3.02$, $P = 0.009$), that is peak amplitude was more important for IE stimuli than for CE ones, and for IE *vs*. IH comparison ($F_{1,12} = 6.68$, $MSE = 3.33$, $P = 0.02$), where IE stimuli were associated with a larger peak amplitude than IH stimuli.

Similar analyses were conducted both on the N100 and P200 peak amplitudes latencies, but they resulted in non-significant differences for any comparison. Moreover, to assess possible "inter-peaks" significant differences in potential, we conducted a spatio-temporal analysis (STA). This analysis consisted in comparing the (Condition 1 – Condition 2) mean amplitude waves in the 80-220 ms time-window, to highlight a possible Congruency effect (i.e. IE – CE, or IH – CH), or a Discriminability effect (i.e. CE – CH, or IE – IH). Significant effects were assessed by Student's *t* tests comparing the amplitude of the (Condition 1 – Condition 2) difference to zero for each time sample (i.e. 2 ms, sampling frequency = 500 Hz), and each electrode. Student's *t* maps were obtained for each latency. Were considered as significant differences those spatio-temporal patterns that had a stable topography (that is over at least two adjacent electrodes) with a significant amplitude ($P<0.05$) for 10 consecutive samples (i.e. 20 ms) [18, 19]. STA revealed no significant difference for any comparison in this specific time-window.

Peak latency analysis did not exhibit any significant difference for any comparison (not reported here).

	CE	CH	IE	IH
Amplitude	5.8 ± 0.7	5.1 ± 0.8	7.3 ± 1.1	5.4 ± 0.8

Table 2. Means and Standard Deviations of the P50 peak amplitude (in µV), in the four experimental conditions.

4. Discussion

In our experiment, early lip movements played a preparatory role in that they always preceded vowel sound onset. Then, by analogy with classical S1-S2 P50 paradigms, lip movements onset would stand for S1, whereas the following vowel sound stands for S2. In accordance with previous findings [13, 20], our results showed that the processing of early lip movement occurred as soon as 180 ms after the movement onset, that is before the vowel sound onset. This N180 response, however, is also elicited by viewing other body movement.

A dipolar modelization using BESA at the N180 peak latencies showed that the current field activity could be

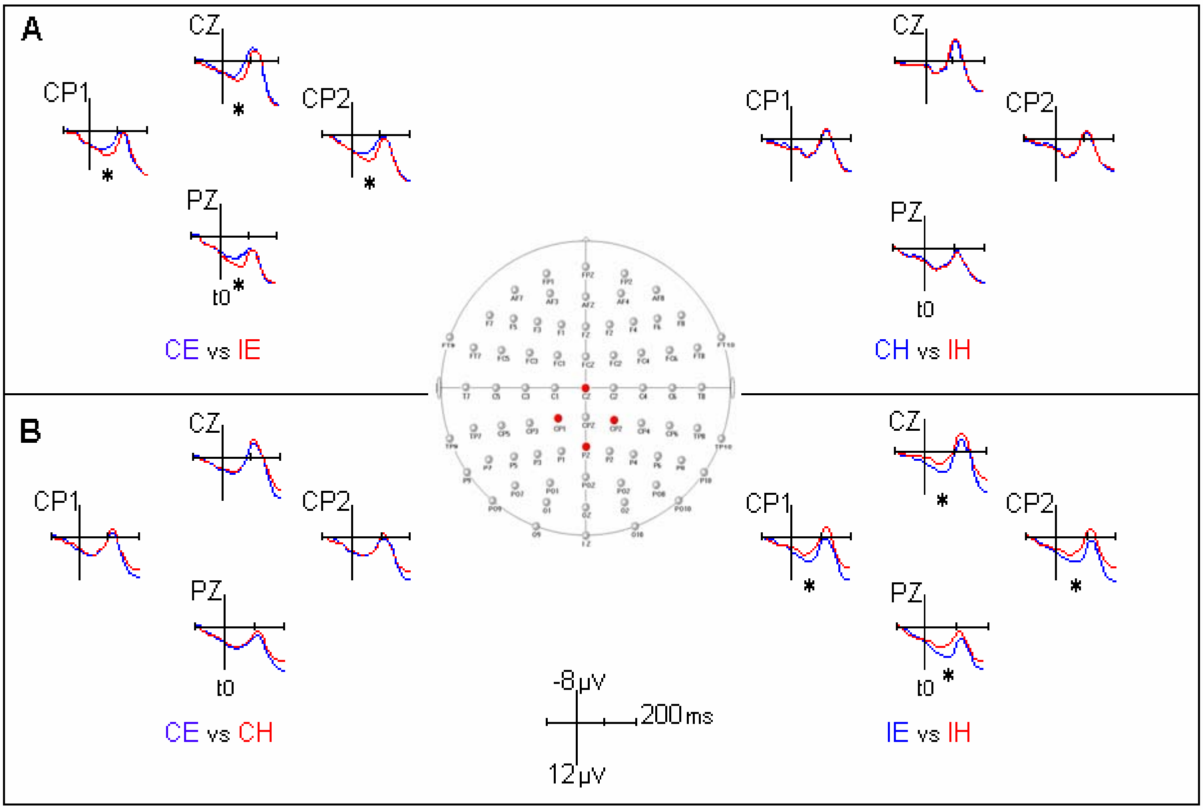


Figure 3. Grand mean ERP wave forms elicited after the vowel sound onset (i.e. origin of axis, or t_0), for each experimental condition, i.e. congruent easy (CE), congruent hard (CH), incongruent easy (IE), and incongruent hard (IH). Conditions were compared along the "congruency" dimension (upper panel, **A**), or along the "discriminability" dimension (lower panel, **B**). Experimental conditions and their corresponding ERP wave forms shared the same color. Significant differences in peak amplitude for the P50 ERP component are illustrated by asterisks. In the center is a schematic view of the head from above (nose upward), with the position of the selected leads in red.

accounted for by the activity of three dipoles. Two dipoles of the fitting model were located in the inferior temporal gyri bilaterally (BA 20). These areas were previously shown to be involved in shape and face detection [21], and mental imagery tasks using linguistic cues [22, 23]. The third dipole was located in the left superior parietal lobule (BA 7), an area involved in visual localization, visuo-motor control, and particularly in spatial attention shift [24].

These data suggest that the dual visual stream activation we found could reflect a visuo-spatial attention shift toward the labial zone [25], and a pre-activation of the acoustical representation corresponding to the same articulatory event.

As compared to [ø] and [y], [a] and [i] vowels are characterized by very distinctive visual properties in their pronunciation. Thus, subjects could easily distinguish the irrelevant audiovisual dubbing with IE stimuli. Taking into account the existing literature on P50 suppression reflecting an electrophysiological index of early sensory gating, we suggest that the larger P50 elicited with IE stimuli could result in an early cerebral detection of non-redundant audiovisual information. The absence of significant P50 amplitude difference in the CE *vs*. CH comparison corroborates this hypothesis, in that visual and auditory information is always redundant in a congruent audiovisual dubbing. Moreover, as “hard” stimuli share very close visual speech cues, the incongruent audiovisual dubbing elicited a sensory gating-like response. This is consistent, notably, with the hardly identical P50 amplitude elicited with CH and IH stimuli. Observation of P50 decrement with particular audiovisual speech stimuli shows that early visual-to-auditory cross-modal effects can also be a source of sensory gating phenomenon. To date, these results provide the first evidence of intersensory gating not only in the perception of audiovisual events, but also in the perception of speech information [26].

5. Conclusion

Our findings support the hypothesis of an existing sensory gating-like response with relatively complex, yet realistic stimuli. Moreover, our results strongly suggest that audiovisual integration takes place very early during the perceptual processes [27, 28], even for speech information [1, 29]. Visual speech, then, modulates auditory pre-attentive detection of irrelevant bimodal stimuli. Boutros and Belger (1999) postulated that sensory gating is a multistage operation, involving sensory gating activity as early as the P50, and as late as the N100 [30] and the mismatch negativity (MMN). These authors concluded to the existence of a multistage and multicomponent sensory gating system, including both early (P50) and late (N100 and MMN) gating indices as well as habituation (i.e. inhibitory) and dishabituation (i.e. excitatory) mechanisms. Our results suggest that this sensory gating system must also include a cross-modal dimension. However, cerebral regions governing sensory gating mechanisms have not been identified yet and further studies should be conducted to identify the brain structures involved in such processing.

6. Acknowledgement

This study was supported by the French Medical Research Foundation (FRM code: FDT20010629055/2). The authors thank Florence Bouchet, Jean-Claude Bourzeix and Laurent Hugueville for technical assistance.

7. References

[1] R. Möttönen, C. M. Krause, K. Tiippana, and M. Sams, "Processing of changes in visual speech in the human auditory cortex," *Brain Res Cogn Brain Res*, vol. 13, pp. 417-25, 2002.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J Acoust Soc Am*, vol. 26, pp. 212-215, 1954.

[3] D. L. Braff and M. A. Geyer, "Sensorimotor gating and schizophrenia. Human and animal model studies," *Arch Gen Psychiatry*, vol. 47, pp. 181-188, 1990.

[4] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones, "Effects of training on the visual recognition of consonants," *J Speech Hear Res*, vol. 20, pp. 130-145, 1977.

[5] B. Dodd, "The role of vision in the perception of speech," *Perception*, vol. 6, pp. 31-40, 1977.

[6] D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception," *J Exp Psychol Hum Percept Perform*, vol. 9, pp. 753-71, 1983.

[7] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-8, 1976.

[8] K. P. Green and P. K. Kuhl, "Integral processing of visual place and auditory voicing information during phonetic perception," *J Exp Psychol Hum Percept Perform*, vol. 17, pp. 278-88, 1991.

[9] D. W. Massaro, "Ambiguity in perception and experimentation," *J Exp Psychol Gen*, vol. 117, pp. 417-21, 1988.

[10] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by eye: The psychology of lip-reading*, B. Dodd and R. Campbell, Eds. London: Erlbaum, 1987, pp. 3-51.

[11] M. MacSweeney, E. Amaro, G. A. Calvert, R. Campbell, A. S. David, P. McGuire, S. C. Williams, B. Woll, and M. J. Brammer, "Silent speechreading in the absence of scanner noise: an event-related fMRI study," *Neuroreport*, vol. 11, pp. 1729-33, 2000.

[12] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen, and A. S. David, "Activation of auditory cortex during silent lipreading," *Science*, vol. 276, pp. 593-6, 1997.

[13] K. J. Wheaton, A. Pipingas, R. B. Silberstein, and A. Puce, "Human neural responses elicited to observing the actions of others," *Vis Neurosci*, vol. 18, pp. 401-6, 2001.

[14] N. N. Boutros and A. Belger, "Midlatency evoked potentials attenuation and augmentation reflect different aspects of sensory gating," *Biol Psychiatry*, vol. 45, pp. 917-22, 1999.

[15] N. N. Boutros, M. W. Torello, B. A. Barker, P. A. Tueting, S. C. Wu, and H. A. Nasrallah, "The P50 evoked potential component and mismatch detection in normal volunteers: implications for the study of sensory gating," *Psychiatry Res*, vol. 57, pp. 83-8, 1995.

[16] M. Scherg and J. S. Ebersole, "Models of brain sources," *Brain Topogr*, vol. 5, pp. 419-23, 1993.

[17] J. Talairach and P. Tournoux, *A coplanar stereotactic atlas of the human brain*. Stuttgart, Germany: Thieme Verlag, 1988.

[18] M. D. Rugg, Doyle M. C., and T. Wells. "Word and nonword repetition within- and across-modality: an event-related potential study,".*J Neurosci*, vol. 7, pp. 209-227, 1995.

[19] S. Thorpe, D. Fize, and C. Marlot. "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520-522, 1996.

[20] A. Puce, A. Smith, and T. Allison, "ERPs evoked by viewing facial movements," *Cognitive Neuropsychology*, vol. 17, pp. 221-239, 2000.

[21] E. Mellet, L. Petit, B. Mazoyer, M. Denis, and N. Tzourio, "Reopening the mental imagery debate: lessons from functional anatomy," *Neuroimage*, vol. 8, pp. 129-39, 1998.

[22] S. M. Kosslyn, N. M. Alpert, W. L. Thompson, V. Maljkovic, S. B. Weise, C. F. Chabris, S. E. Hamilton, S. L. Rauch, and F. S. Buonanno, "Visual mental imagery activates topographically organized visual cortex: PET investigations," *Journal of Cognitive Neuroscience*, vol. 5, pp. 263-287, 1993.

[23] E. Mellet, N. Tzourio, M. Denis, and B. Mazoyer, "Cortical anatomy of mental imagery of concrete nouns based on their dictionary definition," *Neuroreport*, vol. 9, pp. 803-8, 1998.

[24] J. C. Culham, S. A. Brandt, P. Cavanagh, N. G. Kanwisher, A. M. Dale, and R. B. Tootell, "Cortical fMRI activation produced by attentive tracking of moving targets," *J Neurophysiol*, vol. 80, pp. 2657-70, 1998.

[25] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Percept Psychophys*, vol. 60, pp. 926-40, 1998.

[26] R. Lebib, D. Papo, S. de Bode, and P.-M. Baudonniere, "Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the P50 event-related brain potential modulation," *Neuroscience Letters*, vol. 341, pp. 185-188, 2003.

[27] A. Fort, C. Delpuech, J. Pernier, and M. H. Giard, "Early auditory-visual interactions in human cortex during nonredundant target identification," *Brain Res Cogn Brain Res*, vol. 14, pp. 20-30, 2002.

[28] M. H. Giard and F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study," *J Cogn Neurosci*, vol. 11, pp. 473-90, 1999.

[29] C. Colin, M. Radeau, A. Soquet, D. Demolin, F. Colin, and P. Deltenre, "Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory," *Clin Neurophysiol*, vol. 113, pp. 495-506, 2002.

[30] A. S. Smith, N. N. Boutros, and S. B. Schwartzkopf, "Reliability of P50 auditory event-related potential indices of sensory gating," *Psychophysiology*, vol. 31, pp. 495-502, 1994.