



**HAL**  
open science

## Electrophysiological evidence for a self-processing advantage during audiovisual speech integration

Avril Treille, Coriandre Emmanuel Vilain, Sonia Kandel, Marc Sato

### ► To cite this version:

Avril Treille, Coriandre Emmanuel Vilain, Sonia Kandel, Marc Sato. Electrophysiological evidence for a self-processing advantage during audiovisual speech integration. *Experimental Brain Research*, 2017, 235 (9), pp.2867-2876. 10.1007/s00221-017-5018-0 . hal-01616078

**HAL Id: hal-01616078**

**<https://hal.science/hal-01616078>**

Submitted on 21 Nov 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Experimental Brain Research

## ELECTROPHYSIOLOGICAL EVIDENCE FOR A SELF PROCESSING ADVANTAGE DURING AUDIOVISUAL SPEECH INTEGRATION

--Manuscript Draft--

<b>Manuscript Number:</b>	EXBR-D-17-00084R2
<b>Full Title:</b>	ELECTROPHYSIOLOGICAL EVIDENCE FOR A SELF PROCESSING ADVANTAGE DURING AUDIOVISUAL SPEECH INTEGRATION
<b>Article Type:</b>	Research Article
<b>Funding Information:</b>	
<b>Abstract:</b>	<p>Previous electrophysiological studies have provided strong evidence for early multisensory integrative mechanisms during audiovisual speech perception. From these studies, one unanswered issue is whether hearing our own voice and seeing our own articulatory gestures facilitate speech perception, possibly through a better processing and integration of sensory inputs with our own sensory-motor knowledge. The present EEG study examined the impact of self-knowledge during the perception of auditory (A), visual (V) and audiovisual (AV) speech stimuli that were previously recorded from the participant or from a speaker he/she had never met. Audiovisual interactions were estimated by comparing N1 and P2 auditory evoked potentials during the bimodal condition (AV) with the sum of those observed in the unimodal conditions (A+V). In line with previous EEG studies, our results revealed an amplitude decrease of P2 auditory evoked potentials in AV compared to A+V conditions. Crucially, a temporal facilitation of N1 responses was observed during the visual perception of self speech movements compared to those of another speaker. This facilitation was negatively correlated with the saliency of visual stimuli. These results provide evidence for a temporal facilitation of the integration of auditory and visual speech signals when the visual situation involves our own speech gestures.</p>
<b>Corresponding Author:</b>	Avril Treille GIPSA-Lab Grenoble, FRANCE
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	GIPSA-Lab
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Avril Treille
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Avril Treille Coriandre Vilain Sonia Kandel Marc Sato
<b>Order of Authors Secondary Information:</b>	
<b>Author Comments:</b>	<p>Dear Editorial Office,</p> <p>Please, find the minor corrections (spaces were added and the reference to Baart et al. 2014 was moved as request by the Reviewer 2) directly into the manuscript.</p> <p>Regards</p> <p>Avril Treille</p>
<b>Response to Reviewers:</b>	Dear Editorial Office,

Please, find the minor corrections (spaces were added and the reference to Baart et al. 2014 was mooved as request by the Reviewer 2) directly into the manuscrit.

Regards

Avril Treille

[Click here to view linked References](#)

1 ELECTROPHYSIOLOGICAL EVIDENCE FOR A SELF PROCESSING  
2  
3 ADVANTAGE DURING AUDIOVISUAL SPEECH INTEGRATION  
4  
5  
6

7 Avril Treille<sup>1,CA</sup>, Coriandre Vilain<sup>1</sup>, Sonia Kandel<sup>1</sup> & Marc Sato<sup>2</sup>  
8  
9

10  
11  
12 <sup>1</sup>GIPSA-lab, Département Parole & Cognition, CNRS & Grenoble Université, France  
13

14 <sup>2</sup>Laboratoire Parole & Langage, CNRS & Aix-Marseille Université, France  
15  
16  
17  
18  
19  
20  
21

22 <sup>CA</sup>Corresponding author:  
23

24 Avril Treille  
25

26 Gipsa-lab, Université Stendhal  
27

28 1180, avenue Centrale BP25  
29

30 38031 GRENOBLE CEDEX 9  
31

32 Email: [avril.treille@gipsa-lab.grenoble-inp.fr](mailto:avril.treille@gipsa-lab.grenoble-inp.fr)  
33

34 Phone: +33 (0)4 76 82 41 28  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## ABSTRACT

1  
2  
3 Previous electrophysiological studies have provided strong evidence for early multisensory  
4 integrative mechanisms during audiovisual speech perception. From these studies, one unanswered  
5 issue is whether hearing our own voice and seeing our own articulatory gestures facilitate speech  
6 perception, possibly through a better processing and integration of sensory inputs with our own  
7 sensory-motor knowledge. The present EEG study examined the impact of self-knowledge during the  
8 perception of auditory (A), visual (V) and audiovisual (AV) speech stimuli that were previously recorded  
9 from the participant or from a speaker he/she had never met. Audiovisual interactions were estimated  
10 by comparing N1 and P2 auditory evoked potentials during the bimodal condition (AV) with the sum  
11 of those observed in the unimodal conditions (A+V). In line with previous EEG studies, our results  
12 revealed an amplitude decrease of P2 auditory evoked potentials in AV compared to A+V conditions.  
13 Crucially, a temporal facilitation of N1 responses was observed during the visual perception of self  
14 speech movements compared to those of another speaker. This facilitation was negatively correlated  
15 with the saliency of visual stimuli. These results provide evidence for a temporal facilitation of the  
16 integration of auditory and visual speech signals when the visual situation involves our own speech  
17 gestures.

18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
Keywords: Self recognition, speech perception, audiovisual integration, EEG.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## INTRODUCTION

Lip-reading alone is not enough to understand an utterance. However, information from the speaker’s face (e.g., lip movements) is known to improve auditory speech perception. Several studies indicate that visual speech information enhances auditory speech intelligibility in noisy environments (Sumbly & Pollack, 1954; Benoît, Mohamadi & Kandel, 1994), facilitates phoneme identification of non-native phonemes (Navarra & Soto-Faraco, 2005; Burfin et al., 2014) or even contributes to the comprehension of complex content (Reisberg, McLean & Goldfield, 1987). In addition, in laboratory experimental situations, visual incongruent information (/ga/) when added to an auditory syllable (/ba/) can generate a new percept (/da/) different from both the auditory and visual syllables. This perceptual illusion was first displayed by McGurk and MacDonald in 1976 and strikingly underlines the complementarity and intimate interaction between auditory and visual speech information. Interestingly, visual information is not the only way to facilitate auditory speech decoding. Behavioral studies on tactile and audio-tactile speech perception also demonstrate that perceiving orofacial gestures of the speaker through the hand (via the TADOMA method; see Alcorn, 1932) can facilitate syllable discrimination (Reed et al., 1985, 1992; Reed et al., 1982; Fowler & Dekle, 1991; Gick et al., 2008; Sato et al., 2010; Treille et al., 2014a, 2014b).

At the brain level, electro-encephalographic (EEG) and magneto-encephalographic (MEG) studies demonstrate that N1/M1 and P2 auditory evoked potentials are attenuated and speeded up when an auditory syllable is combined with visual or tactile information from the speaker’s face (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg & Vroomen, 2007; Arnal et al., 2009; Pilling, 2010; Vroomen & Stekelenburg, 2010; Frtusova, Winneke & Phillips, 2013; Kaganovich & Schumaker, 2014; Treille et al., 2014a, 2014b; Baart et al., 2014; Baart & Samuel, 2015). This temporal facilitation of latency (onset of neural processing) and amplitude suppression (size of neural population and activation synchrony during the component generation) of N1/M1 and P2 auditory evoked potentials is thought to reflect early multisensory integrative mechanisms through visual predictions of the incoming auditory events. However, the speech specific nature of these effects remains controversial. Indeed, Stekelenburg and Vroomen (2007) and Vroomen and Stekelenburg (2010) observed similar N1 latency and amplitude decreases during the observation of biological transitive (spoon hitting a cup, handclapping) and intransitive (Tearing of paper) non-speech actions, and even during the observation of non-biological actions (a pure tone synchronized with a deformation of a rectangle, or a collision of moving disks). These studies suggested that N1 and P2 modulations would reflect different aspects of audiovisual integration mechanisms (van Wassenhove et al., 2005; Arnal et al., 2009; Baart et al., 2014). There would be a non speech-specific stage in audiovisual integration that processes the early arrival of visual information. This would be reflected by N1 latency and amplitude

1 modulations. A subsequent speech-specific featural phonetic stage would be reflected in P2  
2 modulations (see Baart et al., 2014 for a review).  
3

4 Neuroimaging studies further demonstrate the existence of specific brain areas playing a key role  
5 in the audiovisual integration of speech. In particular, audiovisual speech perception has an impact on  
6 the activity of unisensory visual and auditory regions (the visual motion-sensitive cortex, V5/MT, and  
7 the Heschl's gyrus) as well as multisensory regions (the posterior part of the left superior temporal  
8 gyrus/sulcus, pSTS/pSTG), when compared to auditory and visual unimodal conditions (Calvert,  
9 Campbell and Brammer, 2000; Callan et al., 2003, 2004; Skipper et al., 2005, 2007). Interestingly, the  
10 premotor cortex-that is involved in speech production and is part of the dorsal stream (Hickok &  
11 Poeppel, 2007) might also play a role in audiovisual speech integration mechanisms. Indeed, previous  
12 studies on audiovisual speech perception demonstrated stronger activation of this premotor region  
13 during the presentation of bimodal speech stimuli compared to auditory and visual only conditions  
14 (Campbell et al., 2001; Calvert & Campbell, 2003; Watkins, Strafella & Paus, 2003; Watkins & Paus,  
15 2004; Skipper et al., 2005, 2007; Sato et al., 2010). This occurred during the presentation of  
16 incongruent stimuli compared to congruent ones (Jones & Callan, 2003; Ojanen et al., 2005; Pekkola et  
17 al., 2006) and also in the case of degraded visual or auditory speech signals (Callan et al., 2003, 2004).  
18 Taken together – and although the debate is still open- these studies, support the idea that motor  
19 knowledge used to produce speech sounds might constrain phonetic decoding of the sensory inputs.  
20 This comforts, to a certain extent, the motor and sensorimotor theories of speech perception and  
21 language comprehension (Lieberman & Mattingly, 1985; Skipper et al., 2007; Schwartz et al., 2012;  
22 Pickering & Garrod, 2013) and supports the long-standing proposal that perception and action are two  
23 closely linked processes.  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39

40 From these studies on audiovisual speech perception, one intriguing question is whether hearing  
41 our own voice and seeing our own articulatory gestures facilitate speech perception, possibly through  
42 a better processing and integration of sensory inputs with our own sensory-motor knowledge. From  
43 this question, a few behavioral studies have provided contrasted results. Tye-Murray and colleagues  
44 (2013, 2014) demonstrated that, during sentence lip-reading, participants recognize better their visual  
45 productions than those of others. In contrast, Aruffo and Shore (2012) found a self-auditory but not a  
46 self-visual advantage during the presentation of incongruent audiovisual speech stimuli. Other  
47 behavioral studies attempted to show a self-processing effect during audiovisual syllable perception,  
48 but the results were not concluding (Schwartz and Savariaux, 2001).  
49  
50  
51  
52  
53  
54  
55  
56

57 The present study examined whether self-information processing constitutes an advantage during  
58 audiovisual speech integration. We used EEG to examine N1 and P2 auditory evoked potentials during  
59  
60  
61  
62  
63  
64  
65

1 the perception of auditory and/or visual speech stimuli that were previously recorded from the  
2 participant (self) and a speaker he/she had never met (other). For each participant, eight conditions  
3 were tested, consisting on four distinct modalities: an auditory modality ( $A_{self}$ ,  $A_{other}$ ), a visual modality  
4 ( $V_{self}$ ,  $V_{other}$ ), an audiovisual modality ( $A_{self}V_{self}$ ,  $A_{other}V_{other}$ ) and an audiovisual modality with incongruent  
5 speakers in which the acoustic and visual signals were produced by the participant and the other  
6 speaker respectively ( $A_{self}V_{other}$ ,  $A_{other}V_{self}$ ). The audiovisual modality with incongruent speakers was  
7 designed to determine whether a possible self-effect comes from auditory or visual information. Using  
8 an additive model, we tested whether N1/P2 auditory evoked potentials were attenuated and speeded  
9 up during audiovisual conditions compared to the sum of those observed in unimodal conditions, and  
10 whether these effects were modulated by a self-processing advantage.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



## METHOD

### PARTICIPANTS

Eighteen healthy adults participated in the study (12 females; mean age 23, SD  $\pm$  5 years). All the participants were right-handed native French speakers, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. They gave written consent for their participation in the study. They were compensated for the time spent in the study. The study received approval by the Grenoble Alpes University Ethical Committee (CERNI, N°2013-12-24-33).

### STIMULI

*Recording* – We recorded 10 utterances of /apa/, /ata/ and /aka/ sequences of each participant in a soundproof room. Previous research on audiovisual speech perception has shown that these sequences correspond to a gradient of visuo-labial saliency: the unvoiced bilabial /p/ stop consonant is more salient visually than unvoiced alveolar stop consonant /t/ and in turn stop consonant velar /k/ unvoiced (e.g., van Wassenhove et al., 2005 for an EEG study). Moreover, these stop consonants have precise acoustics onsets, which is crucial for the EEG analyses we intended to carry out (see below). Then, we selected four utterances of each sequence for each participant on the basis of visual and acoustical durations (using Adobe Premiere, Adobe Systems, and Praat software; Boersma & Weenink, 2013).

*Stimulus preparation*—The movies were created on the basis of 30 frames (1200 ms) before the acoustic burst and 5 frames (200 ms) after it, for a total duration of 1400 ms for all the stimuli. Prior to generating movies, we extracted the acoustic signal and erased the first vowel /a/ so that all the audio signals began with a 1200 ms silence. This procedure allows building stimuli with the same duration of an initial neutral mid-open mouth position of each participant (for examples on AV stimuli see supplementary material). Then, we merged the audio and video signals in four different types of movies:

-Auditory modality (A): the movie consisted of a fixed image of the last frame before the acoustic onset during the initial vowel /a/ dubbed on the acoustic signal.

-Visual modality (V): The movie consisted of the visual input without the sound.

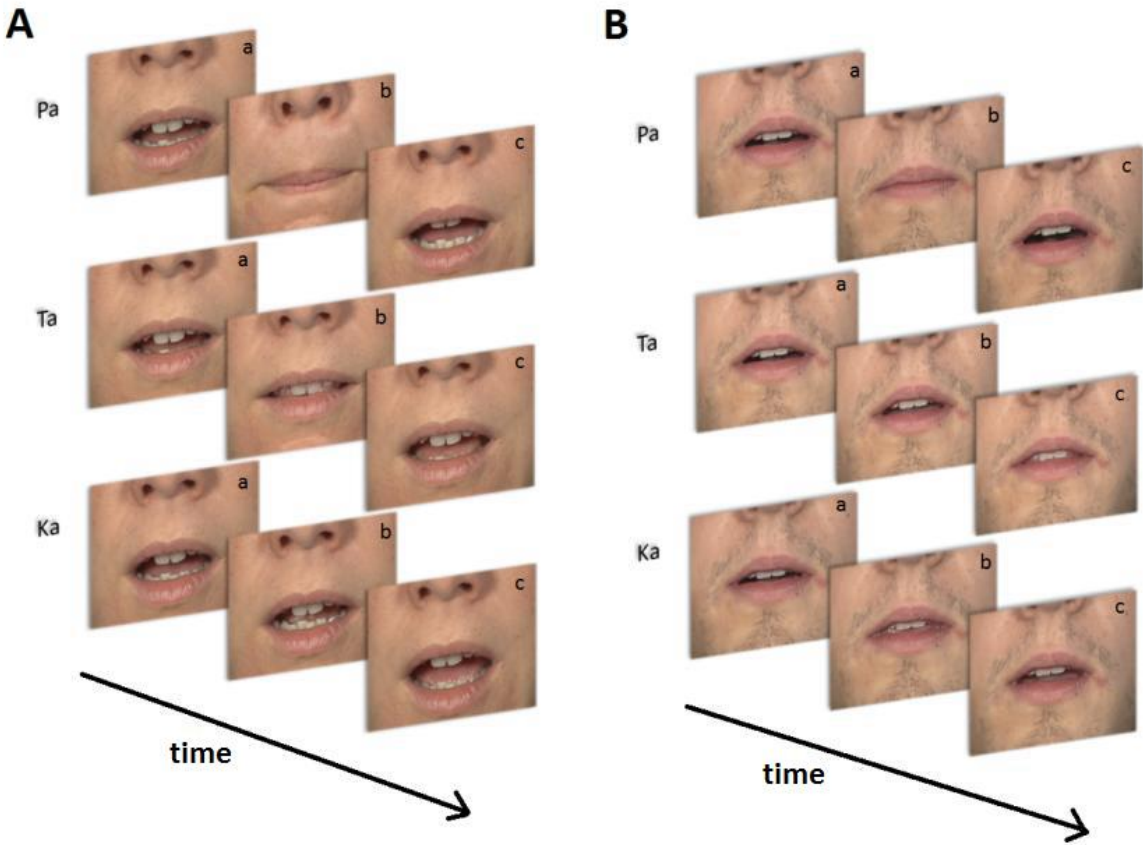
-Audiovisual modality (AV): The movie consisted of both the auditory and visual signals.

-Audiovisual modality with incongruent speakers (AV incongruent speakers): The movie consisted of the acoustic signal of the speaker dubbed with the visual signal of the same syllable but produced by another participant (see below for the matching method).

*Participant pair matching* – Because of possible idiosyncrasy or production differences between participants that might cause facilitation or perturbation of visual or auditory stimuli recognition, each

1 participant was associated to an unknown participant (same gender and equivalent age). Each pair of  
2 participants was therefore presented with the same set of stimuli from both participants. With this  
3 procedure, a possible self effect cannot therefore be attributed to possible idiosyncrasy differences.  
4

5  
6 Our experiment therefore consisted of 9 pairs of participants. To each participant we presented  
7 both her/his own productions and those of her/his unknown partner (see Figure 1). For each  
8 participant, eight conditions were tested, consisting on four distinct modalities applied either on the  
9 participant her/himself (self) or the unknown speaker (other): an auditory modality ( $A_{self}$ ,  $A_{other}$ ), a visual  
10 modality ( $V_{self}$ ,  $V_{other}$ ), an audiovisual modality ( $A_{self}V_{self}$ ,  $A_{other}V_{other}$ ) and an audiovisual modality with  
11 incongruent speakers in which the acoustic and visual signals were produced by the participant and  
12 the other speaker ( $A_{self}V_{other}$ ,  $A_{other}V_{self}$ ). The audiovisual modality with incongruent speakers was  
13 designed to determine whether a possible self-effect comes from auditory or visual information. With  
14 this procedure, a total of 864 stimuli were created (18 speakers x 4 modalities x 3 syllables x 4  
15 utterances).  
16  
17  
18  
19  
20  
21  
22  
23

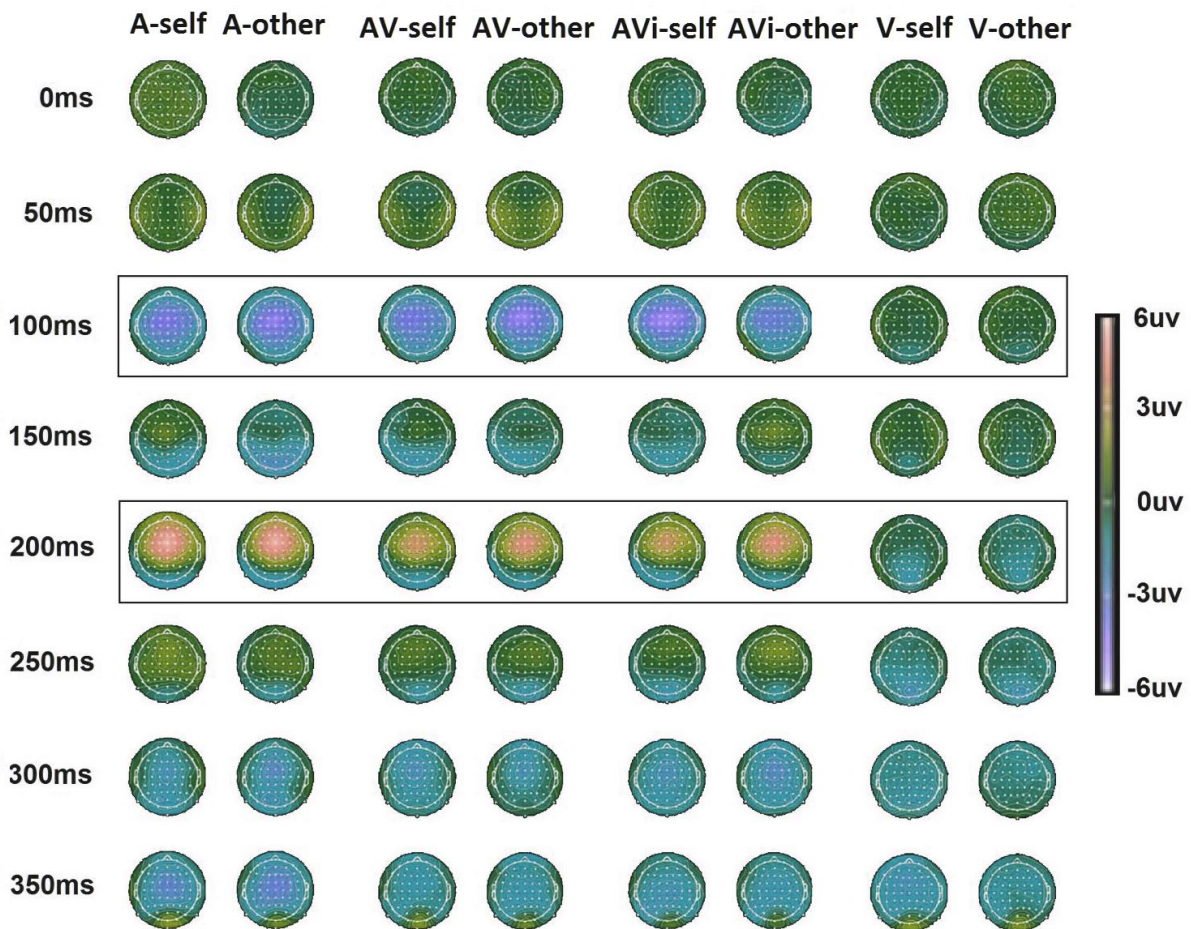


24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
**Figure 1: Examples of the visual stimuli for two participants (A,B). Each utterance begins with the mouth open (a); is followed by the stop consonant (b); and ends with the second /a/ vowel (c).**

## EXPERIMENTAL PROCEDURE

1 The participants sat in front of a computer monitor at a distance of approximately 50 cm. The  
2 acoustic stimuli were presented at a comfortable sound level through loudspeakers, with the same  
3 sound level set for all participants (frame-rate of the video recordings: 25 images/sec, refresh-rate of  
4 the monitor: 60 Hz). The software *Presentation* (Neurobehavioral Systems, Albany, CA) controlled  
5 stimulus presentation and recorded the participants' responses. The participants were instructed to  
6 identify the syllable presented by the movies by pressing a key on the keyboard with their left hand. It  
7 was a three-alternative /pa/, /ta/ and /ka/ forced-choice identification task. In order to dissociate  
8 sensory/perceptual responses from motor responses on EEG data, a brief single audio beep was  
9 delivered 600 ms after the end of each stimulus. The participants had to respond after this audio beep.  
10 The experiment consisted of 576 trials presented in a pseudo-randomized sequence, with 24 trials in  
11 each condition (4 modalities (A, V, AV, AV with incongruent speakers) x 2 speakers (self and other) x 3  
12 syllables (/pa/, /ta/ and /ka/) x 24 trials). The inter-trial interval was set at 3 s and the response key  
13 designation was fully counterbalanced across participants.  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## EEG ACQUISITION AND PROCESSING



**Figure 2: Topographic analysis conducted on all the participants and electrodes demonstrating a maximal response of N1/P2 auditory evoked potentials on fronto-central electrodes.**

EEG data were recorded continuously from 64 scalp electrodes (Electro-Cap International, INC, according to the international 10-20 system) using the Biosemi Active Two AD-box EEG system operating at a 256 Hz sampling rate. Two additional electrodes served as reference (Common Mode Sens [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). One other external reference electrode was set at the top of the nose. Horizontal (HEOG) and vertical (VEOG) eye movements were recorded using an electro-oculogram with electrodes positioned at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

EEG data were processed using the EEGLAB toolbox (Delorme and Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over central sites on the scalp (Scherg and Von Cramon, 1986; Näätänen and Picton, 1987), EEG data preprocessing and analyses were conducted on 6 representative fronto-central electrodes (F3, Fz, F4, C3, Cz, C4). This is in line with previous EEG studies on audiovisual speech perception and auditory

1 evoked potentials (e.g. Pilling, 2010; Stekelenburg and Vroomen, 2007; van Wassenhove et al., 2005;  
2 Vroomen and Stekelenburg, 2010). A topographic analysis conducted on all the participants and 64  
3 electrodes demonstrated a maximal response of N1/P2 auditory evoked potentials on fronto-central  
4 electrodes (see Figure 2). This confirmed the reliability of our selection of fronto-central electrodes.  
5 EEG data were first off-line re-referenced to the nose recording and band-pass filtered using a two-  
6 way least-square FIR filtering (1-20 Hz). Data were then segmented into 1000 ms epochs including a  
7 100 ms pre-stimulus baseline (from -500 ms to -400 ms relative to the acoustic syllable onset). Epochs  
8 with an amplitude change exceeding  $\pm 60$   $\mu$ V at any channel (including HEOG and VEOG channels) were  
9 rejected (on average, less than 6%). For each participant and condition ( $A_{self}$ ,  $A_{other}$ ,  $V_{self}$ ,  $V_{other}$ ,  $A_{self}V_{self}$ ,  
10  $A_{other}V_{other}$ ,  $A_{self}V_{other}$ ,  $A_{other}V_{self}$ ), the data were averaged on the 6 electrodes. Then the maximal  
11 amplitude and peak latency of auditory N1 and P2 evoked responses were determined on the EEG  
12 waveform using a fixed window (N1: 70-150 ms; P2: 150-250 ms).  
13  
14  
15  
16  
17  
18  
19  
20  
21

## 22 DATA ANALYSES

### 23 **Behavioral analyses**

24 The percentage of correct responses was determined for each participant, syllable and modality.  
25 We conducted a three-way repeated-measures ANOVAs with speaker type (self vs. other), modality  
26 (A, V, AV, AV with incongruent speakers) and the syllable (/pa/, /ta/ and /ka/) as within-participants  
27 variables.  
28  
29  
30  
31  
32

### 33 **EEG analyses**

34 *Audiovisual integration* – To test audiovisual speech integration, we used an additive model, with  
35 EEG responses in the bimodal conditions (AV) compared to the sum of auditory and visual EEG  
36 responses (A+V). We conducted three-way repeated-measures ANOVAs on N1/P2 amplitudes and  
37 latencies with signal type (bimodal vs. sum), auditory speaker (self vs. other) and visual speaker (self,  
38 other or none) as within-participants factors.  
39  
40  
41  
42  
43  
44

45 *Correlation between accuracy and EEG signals*–To test the relation between the perceptual visual  
46 saliency and degree of integration observed on the EEG signals, we conducted Pearson correlation  
47 analyses. The analyses concerned the relation between visual accuracy and the modulations of either  
48 N1/P2 amplitude or latency. They were related to the difference between the bimodal conditions and  
49 the sum of unimodal conditions (e.g., EEG responses on [ $A_{self}V_{self} - (A_{self} + V_{self})$ ] and [ $A_{other}V_{self} - (A_{other} +$   
50  $V_{self})$ ] correlated with  $V_{self}$  scores, or EEG responses on [ $A_{self}V_{other} - (A_{self} + V_{other})$ ] and [ $A_{other}V_{other} - (A_{other}$   
51  $+ V_{other})$ ] correlated with  $V_{other}$  scores).  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## RESULTS

### ACCURACY

Overall, the mean proportion of correct responses was 94% (see Figure 3). The analyses revealed a main effect of presentation modality ( $F(3,51)=67.6$ ;  $p<.0001$ ). The percentages of correct responses for the visual stimuli (83%) were lower than for auditory (A: 98%) and audiovisual stimuli (AV: 99%; AVi: 98%). In addition, consonant saliency also yielded a main effect ( $F(2,34)=23.3$ ;  $p<.0001$ ). The /pa/syllables were identified better (98%) than the /ta/ (92%) and in turn /ka/ (93%) ones. Finally, the interaction between the presentation modality and the syllable was reliable ( $F(6,102)=24.1$ ;  $p<.0001$ ). There was an effect of syllable saliency in the visual modality (V-/pa/: 99%; V-/ta/: 75%; V-/ka/: 74%).

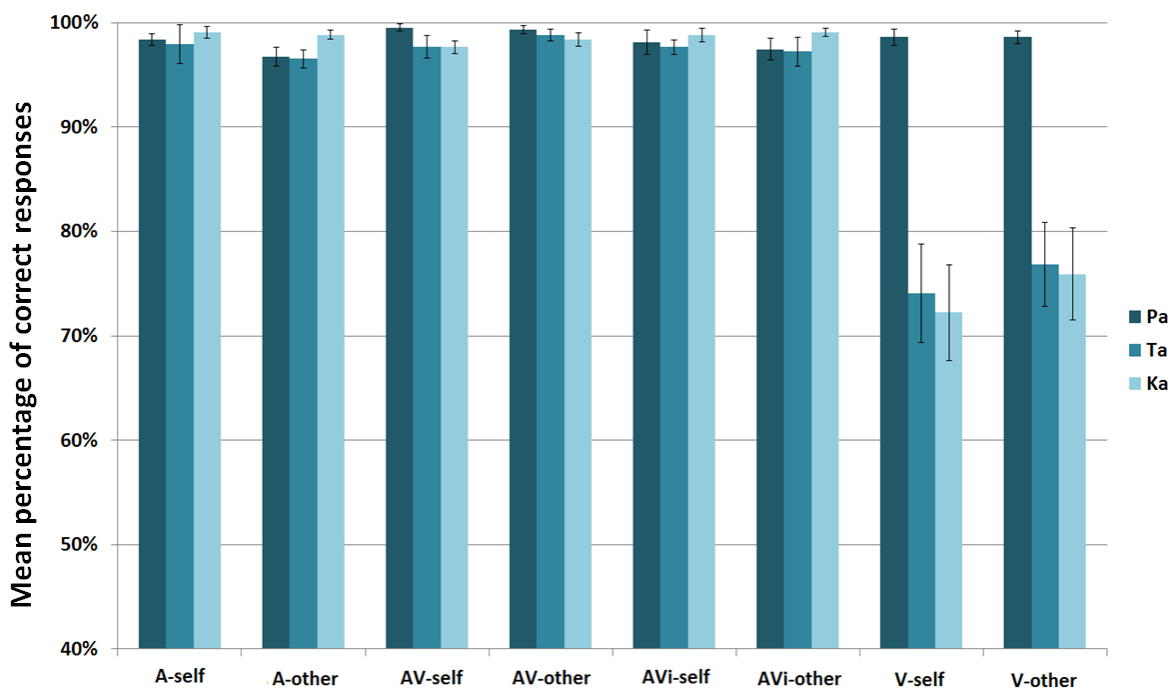


Figure 3: Mean percentage of correct responses observed for each speaker type, presentation modality and each syllable.

### EEG RESULTS

*Amplitude* – None of the effects reached significance for N1 amplitude. There was a main effect of signal type for P2 amplitude ( $F(1,16)=6.9$ ;  $p<.02$ ; see Figure 4). The amplitude was smaller for the bimodal conditions (3.8  $\mu$ V) than the sum of the auditory and visual signals (4.7  $\mu$ V).

*Latency* – Regarding the analyses on N1 latency, there was a significant effect of the visual speaker ( $F(1,16)=8.2$ ;  $p<.02$ ; see Figure 4). There was a temporal facilitation during the perception of visual-self speech movements (107 ms) compared to visual-other speech movements (113 ms). No significant effects were found for P2 latency.



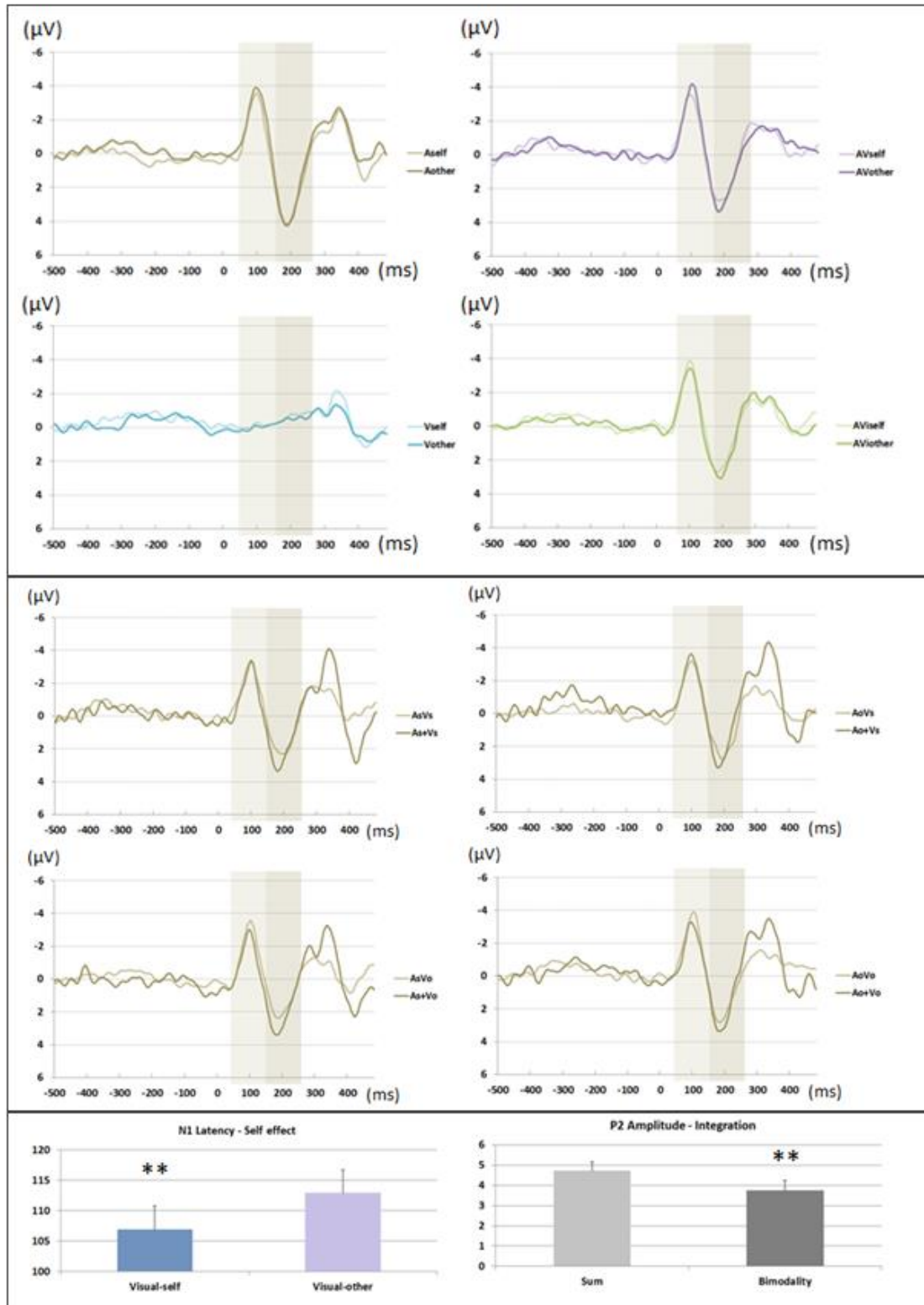


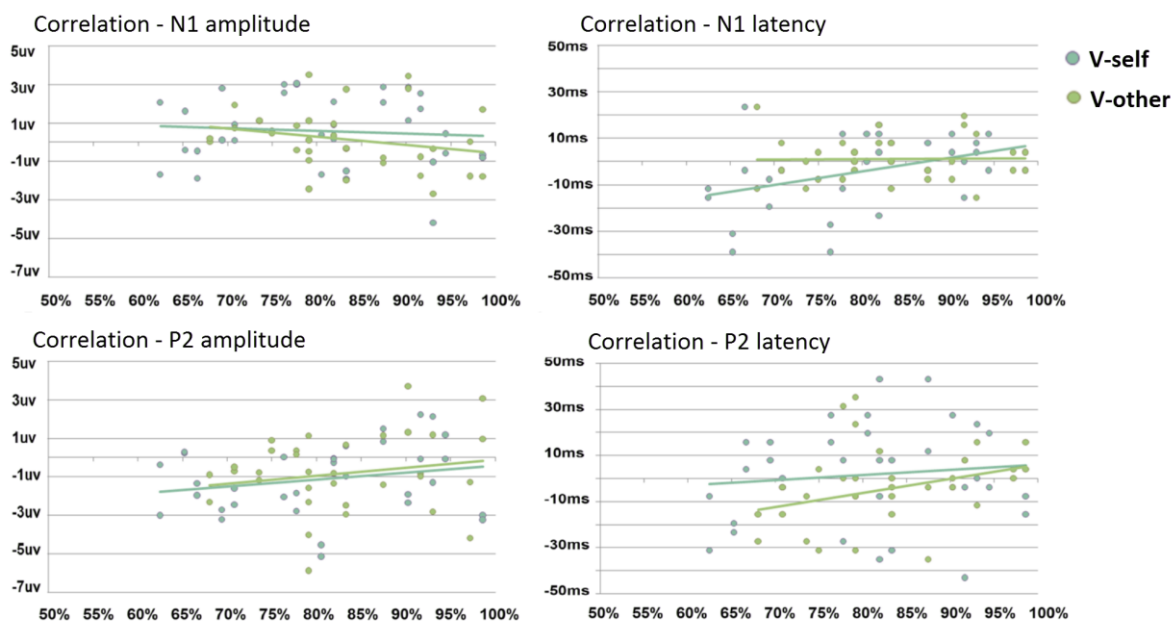
Figure 4: Top: Average event-related potentials on fronto-central electrodes related to each modality (A, AV, V, AVi) for “self” (light colour) and “other” (dark colour) conditions. Middle: Averaged event-related potentials on fronto-central electrodes related to the audiovisual conditions (AV) and the sum of unimodal conditions (A+V) according to the auditory and the visual speakers (s: self; o: other). Bottom: Significant effects on N1 and P2. Left: Latency N1 decrease observed in audiovisual conditions for self compared to other visual movements. Right: Amplitude P2 decrease observed for AV compared to A+V.

## CORRELATION BETWEEN BEHAVIORAL SCORES AND EEG SIGNALS

Because a significant reduction of N1 latency was observed for self stimuli, we conducted additional correlation analyses between visual recognition scores and both amplitude and latency of N1 and P2 PERs in order to test a possible relationship between the perceptual visual saliency and degree of integration observed on the EEG signals.

*Amplitude* - No significant correlation was found between EEG signals related to AV integration and the visual saliency of syllables for both N1 and P2 amplitude (N1: self:  $r=.09$ ;  $p<.63$ ; other:  $r=.24$ ;  $p<.16$ ; P2: self:  $r=.22$ ;  $p<.22$ ; other:  $r=.18$ ;  $p<.30$ ; see Figure 5).

*Latency* - N1 latency difference between AV and A+V EEG responses related to the visual-self syllables was negatively correlated with the visual recognition scores (V-self:  $r=.41$ ;  $p<.02$ ). No significant correlation was observed for the visual syllables from an unknown speaker (V-other:  $r=.01$ ;  $p<.94$ ). Finally, no significant correlation was observed between P2 latency data related to the degree of integration of self and other visual information and visual accuracy (V-self:  $r=.11$ ;  $F(1,32)=0.32$ ;  $p<.54$ ; V-other:  $r=.29$ ;  $F(1,32)=2.95$ ;  $p<.10$ ; see Figure 5).



**Figure 5: Correlation between the visual recognition scores for self and other visual movements (x-axis) and the difference in amplitude and latency of N1 and P2 auditory evoked potentials between AV and A+V (y-axis).**



## DISCUSSION

1  
2 The present EEG study investigated a possible self-processing advantage during speech  
3 perception, and its related impact on audiovisual integration mechanisms. Two main results were  
4 observed. First and in line with previous EEG studies on audiovisual speech integration, we observed  
5 an amplitude decrease on P2 auditory evoked potentials during the bimodal presentation compared  
6 to the sum of auditory and visual unimodal responses. Crucially, during audiovisual speech integration,  
7 a temporal facilitation related to self lip movements was observed on N1 auditory evoked potentials,  
8 a facilitation that appears negatively correlated with the saliency of visual stimuli.  
9

10  
11 Previous studies on audiovisual speech integration demonstrated that bimodal presentations  
12 produce a decrease in N1 and/or P2 latency and amplitudes (Besle et al., 2004; van Wassenhove et al.,  
13 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014; Treille et al., 2014a, 2014b)  
14 and latency (van Wassenhove et al., 2005; Stekelenburg et Vroomen, 2007; Baart et al., 2014; Treille  
15 et al., 2014a; see also Arnal et al., 2009 for similar results with MEG) when compared to auditory  
16 responses or to the sum of auditory and visual responses. These modulations of the N1/P2 responses  
17 are thought to reflect specific stages of audiovisual speech integration. N1 latency and amplitude  
18 modulations would reflect a non speech-specific stage while P2 latency shifts or amplitude decreases  
19 would rather be speech-specific and related to a featural phonetic stage (Baart et al., 2014). Using an  
20 additive model, our results revealed a P2 amplitude decrease during the bimodal presentation  
21 compared to the sum of the unimodal auditory and visual conditions. In line with previous studies (van  
22 Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014; Treille et  
23 al., 2014b), this result suggests that visual speech information affects ongoing auditory activity and  
24 further demonstrates the integration of auditory and visual speech signals. However, there were no  
25 differences on P2 latency, nor on N1 amplitude and latency. This contrasts with previous studies  
26 reporting latency shifts of auditory evoked responses and/or N1 amplitude decreases in the bimodal  
27 condition. Some aspects of the present experimental procedure might explain these differences. A first  
28 important point is related to the stimulus variability. In our experiment we presented four tokens of  
29 three syllables produced by two speakers. The above mentioned studies only presented one token of  
30 each presented syllable (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al.,  
31 2009; Baart et al., 2014) and/or a more limited number of syllables (i.e., one or two; Stekelenburg and  
32 Vroomen, 2007; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Treille et al.,  
33 2014a). In the present EEG experiment, the higher stimulus variability might have decreased eventual  
34 habituation/learning effects. This might have limited latency shifts on auditory evoked potentials.  
35 From that view, a recent meta-analysis suggests that variability across EEG/MEG studies on audiovisual  
36 speech integration may potentially be driven by many experimental, procedural, and methodological  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 differences, such as the number and quality of stimuli, the sound intensity, the inter-trial interval, the  
2 task, the degree of selective attention, the preprocessing and the analysis of the data (Baart, 2016).

3 It is noteworthy that our behavioral results did not reveal any visual, auditory or audiovisual self-  
4 processing advantage. This contrasts with a behavioral study conducted by Tye-Murray and colleagues  
5 (2013). They showed that we lip-read more accurately sentences produced by ourselves than by other  
6 speakers. For the authors, these results provide support to the common coding theory (Prinz, 1997;  
7 Hommel et al., 2001), which posits that producing and perceiving share the same representations of  
8 motor plans. Because of this perceptuo-motor coupling, observing one's own action activates these  
9 motor plans to a greater extent than observing someone else's action. A reason for this divergence  
10 could reside on stimulus length. In the present study we used syllables whereas Tye-Murray et al used  
11 sentences. The use of short CV syllables therefore limited the quantity of visual information and  
12 facilitated correct responses (mean 94%). Our results appear consistent however with the study by  
13 Aruffo and colleagues (2012) who did not find any visual self-processing advantage with participants  
14 presented with incongruent audiovisual syllables (McGurk stimuli), although self-voice appeared to  
15 weaken the illusion effect.  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25

26 The major contribution of our EEG study is that it provides evidence for a visual self-processing  
27 advantage on N1 latency during audiovisual speech perception. More specifically, a temporal  
28 facilitation of audiovisual speech processing was observed when participants watched their own  
29 productions compared to those of another speaker. This facilitation was negatively correlated with the  
30 recognition score of visual self-stimuli. This suggests that the visual self-processing effect is linked to  
31 specific visual speech features of the presented syllables, like the place of articulation of the  
32 consonants (with their acoustic bursts here used as onsets for EEG analyses). Interestingly, this effect  
33 seems to be largely driven by visually "ambiguous" syllables, i.e. syllables that were the most difficult  
34 to identify (see Figure 5). Although this correlational result precludes any causal inferences, a plausible  
35 explanation could be that the difficulty to decode our own speech gestures would increase the degree  
36 of audiovisual integration and temporally facilitate auditory process.  
37  
38  
39  
40  
41  
42  
43  
44  
45

46 In conclusion, the present EEG study provides the first electrophysiological evidence for a self-  
47 processing advantage during audiovisual speech integration. The observed temporal facilitation of N1  
48 responses during the visual perception of self speech movements compared to those of another  
49 speaker suggest that perceiving our own articulatory gestures speed up auditory speech perception,  
50 possibly through a better processing and integration of sensory inputs with our own sensory-motor  
51 knowledge.  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## ACKNOWLEDGEMENTS

This study was supported by research funds from the European Research Council (FP7/2007-2013 Grant Agreement no. 339152).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

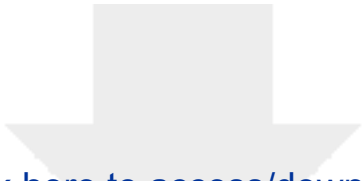
## REFERENCES

- 1  
2 Alcorn, S. (1932). The Tadoma method. *Volta Review*, 34:195-198
- 3  
4 Arnal, L.H., Morillon, B., Kell, C.A. & Giraud, A.L. (2009). Dual neural routing of visual facilitation in  
5 speech processing. *The Journal of Neuroscience*, 29(43):13445-13453
- 6  
7 Aruffo, C., & Shore, D.I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual  
8 speech. *Psychonomic Bulletin & Review*, 19:66–72
- 9  
10 Baart, M., Stekelenburg, J. J. & Vroomen, J. (2014). Electrophysiological evidence for speech-specific  
11 audiovisual integration. *Neuropsychologia*, 65:115–211
- 12  
13 Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between  
14 lip-read and lexical context during speech sound processing. *Journal of Memory and Language*,  
15 85:42–59
- 16  
17 Baart, M. (2016). Quantifying lip-read induced suppression and facilitation of the auditory N1 and P2  
18 reveals peak enhancements and delays. *Psychophysiology*, 53(9):1295-306
- 19  
20 Benoit, C., Mohamadi, T. & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility  
21 of French speech in noise. *Journal Speech Hearing Research*, 37:1195–1203
- 22  
23 Besle, J., Fort, A., Delpuech, C. & Giard, M.H. (2004). Bimodal speech: early suppressive visual effects  
24 in human auditory cortex. *European journal of Neuroscience*, 20:2225-2234
- 25  
26 Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer. Computer program, Version  
27 5.3.42, retrieved 2 March 2013 from <http://www.praat.org/>
- 28  
29 Burfin, S., Pascalis, O., Ruiz Tada, E., Costa, A., Savariaux, C. & Kandel S. (2014). Bilingualism affects the  
30 audio-visual processing of non-native phonemes. *Frontiers in Psychology (Research Topic “New  
31 advances on the perception and production of non-native speech sounds” – Section Language  
32 Sciences)*, 5: 1179
- 33  
34 Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural  
35 processes underlying perceptual enhancement by visual speech gestures. *Neuro Report*, 14:2213-  
36 2217
- 37  
38 Callan, D.E., Jones, J.A., Munhall, K.G., Callan, A.M., Kroos, C. & Vatikiotis-Bateson, E. (2004).  
39 Multisensory integration sites identified by perception of spatial wavelet filtered visual speech  
40 gesture information. *Journal of Cognitive Neuroscience*, 16:805-816
- 41  
42 Calvert, G.A. & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates  
43 of visible speech. *Journal of Cognitive Neuroscience*, 15:57–70
- 44  
45 Calvert, G.A., Campbell, R. & Brammer, M.J. (2000). Evidence from functional magnetic resonance  
46 imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10(11):649-657
- 47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65


- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M.J. & David, A.S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, 12:233-243
- Fowler, C. & Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *Journal of Experimental Psychology- Human Perception and Performance*, 17:816–828
- Frtusova, J. B., Winneke, A. H., & Phillips, N. A. (2013). ERP evidence that auditory–visual speech facilitates working memory in younger and older adults. *Psychology and Aging*, 28(2), 481–494
- Gick, B., Jóhannsdóttir, K.M., Gibrael, D. & Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of Acoustical Society of America*, 123:72-76
- Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Review Neurosciences*, 8:393-402
- Hommel, B., Musseler, J., Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24:849–878
- Jones, J.A. & Callan, D.E (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuro report*, 14:1129–1133
- Kaganovich, N., & Schumaker, J. (2014). Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. *Brain and Language*, 139:36–48
- Klucharev, V., Möttönen, R. & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res. Cogn. Brain Res.*, 18:65-75
- Liberman, A.M. & Mattingly, I.G (1985). The motor theory of speech perception revised. *Cognition*, 21:1-36
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746-748
- Näätänen, R. & Picton, T.W. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24:375–425
- Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. & Sams, M. (2005). Processing of audiovisual speech in Broca’s area. *NeuroImage*, 25:333–338
- Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, L.P., Kujala, T. & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3T. *NeuroImage*, 29(3):797–807
- Pickering, M.J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.*, 36:329–347

- 1 Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal*  
2 *of Speech, Language, and Hearing Research*, 52(4):1073-1081
- 3 Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9:129–  
4 154
- 5  
6  
7 Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Braida, L.D., Conway-Fithian, S. & Schultz, M.C. (1985).  
8 Research on the Tadoma method of speech communication. *J Acoust Soc. Am.*, 77(1):247-257
- 9  
10 Reed, C.-M., Rabinowitz, W.-M., Durlach, N.-I., Delhorne, L.-A., Braida, L.-D., Pemberton, J.-C.,  
11 Mulcahey, B.-D. & Washington, D.-L. (1992). Analytic study of the Tadoma method: Improving  
12 performance through the use of supplementary tactual displays. *Journal of Speech and hearing*  
13 *Research*, 35:450-465
- 14  
15  
16  
17 Reisberg, D., McLean, J. & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading  
18 advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The*  
19 *Psychology of LipReading*, 97-114
- 20  
21  
22  
23 Sato, M., Buccino, G., Gentilucci, M. & Cattaneo, L. (2010). On the tip of the tongue: modulation of the  
24 primary motor cortex during audiovisual speech perception. *Speech Communication*, 52(6): 533-541
- 25  
26 Scherg, M., and VonCramon, D. (1986). Evoked dipole source potentials of the human auditory cortex.  
27 *Electroencephalogr. Clin. Neurol.*, 65:344–360
- 28  
29  
30 Schwartz, J.L. & Savariaux, C. (2001). Is it Easier to Lipread One's Own Speech Gestures Than Those of  
31 Somebody Else? It Seems Not! *Auditory-Visual Speech Processing*, Aalborg, Denmark, 18-23
- 32  
33  
34 Schwartz, J.L., Ménard, L., Basirat, A. & Sato, M. (2012). The Perception for Action Control Theory  
35 (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336-354
- 36  
37  
38 Skipper, J., Van Wassenhove, V, Nussman, H. & Small, S. (2007). Hearing lips and seeing voices: How  
39 cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral*  
40 *Cortex*, 17:2387-2399
- 41  
42  
43 Skipper, J.I., Nusbaum, H.C. & Small, S.L. (2005). Listening to talking faces: motor cortical activation  
44 during speech perception. *NeuroImage*, 25:76–89
- 45  
46  
47 Stekelenburg, J.J. & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically  
48 valid audiovisual events. *Journal of Cognitive Neuroscience*, 19:1964–1973
- 49  
50  
51 Sumbly, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of*  
52 *the Acoustical Society of America*, 26:212-215
- 53  
54  
55 Treille, A., Cordeboeuf, C., Vilain, C. & Sato, M. (2014a). Haptic and visual information speed up the  
56 neural processing of auditory speech in live dyadic interactions. *Neuropsychologia*, 57:71-77
- 57  
58  
59 Treille, A., Vilain, C. & Sato, M. (2014b). The sound of your lips: electrophysiological cross-modal  
60 interactions during hand-to-face and face-to-face speech perception. *Frontiers in Psychology*,  
61 5(420):1-9  
62  
63  
64  
65


- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- Tye-Murray, N., Spehar, B., Myerson, J., Hale, S. & Sommers, M.S. (2013). Reading your own lips: Common coding theory and visual speech perception. *Psychonomic Bulletin & Review*, 20:115-119
- Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. (2014). Lipreading in school-age children: The roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research*, 57:556–565
- van Wassenhove, V., Grant, K.W. & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences U.S.A.*, 102:1181-1186
- Vroomen, J. & Stekelenburg, J.J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, 22:1583-1596
- Watkins, K.E., Strafella, A.P. & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41:989-994
- Watkins, K.E. & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *Journal of Cognitive Neuroscience*, 16(6):978-987




Click here to access/download  
**Supplementary Material**  
s05-pa-2-AV\_norm.avi








Click here to access/download  
**Supplementary Material**  
s07-ta-2-AV\_norm.avi





Click here to access/download  
**Supplementary Material**  
s12-ka-3-AV\_norm.avi

