Neuropsychologia 57 (2014) 71–77

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions

Avril Treille *, Camille Cordeboeuf, Coriandre Vilain, Marc Sato

GIPSA-LAB, Département Parole and Cognition, CNRS and Grenoble Université, Grenoble, France

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form
2 February 2014
Accepted 4 February 2014
Available online 11 February 2014

Keywords:
Audio–visual speech perception
Audio–haptic speech perception
Multisensory interactions
EEG

ABSTRACT

Speech can be perceived not only by the ear and by the eye but also by the hand, with speech gestures felt from manual tactile contact with the speaker's face. In the present electro-encephalographic study, early cross-modal interactions were investigated by comparing auditory evoked potentials during auditory, audio–visual and audio–haptic speech perception in dyadic interactions between a listener and a speaker. In line with previous studies, early auditory evoked responses were attenuated and speeded up during audio–visual compared to auditory speech perception. Crucially, shortened latencies of early auditory evoked potentials were also observed during audio–haptic speech perception. Altogether, these results suggest early bimodal interactions during live face-to-face and hand-to-face speech perception in dyadic interactions.

1. Introduction

Interactions between auditory and visual modalities are beneficial in daily conversation. Visual speech information is known to effectively improve speech intelligibility in noise (Benoît, Mohamadi, & Kandel, 1994; Sumby & Pollack, 1954), the understanding of a semantically complex acoustic statement (Reisberg, McLean, & Goldfield, 1987) or a foreign language (Navarra & Soto-Faraco, 2005). Furthermore, seeing incongruent articulatory gestures may also modify auditory speech perception (McGurk & MacDonald, 1976). The fact that visual input may facilitate or even change the perceiver's auditory experience thus provides clear evidence for audio–visual integration in speech processing.

Despite no current agreement between theoretical models of audio–visual speech perception regarding the processing level at which the acoustic and visual speech signals fuse to a unified speech percept (for a review, see Schwartz, Robert-Ribes, and Escudier (1998)), recent electro-encephalographic (EEG) and magneto-encephalographic (MEG) studies demonstrate that early auditory evoked potentials N1 and P2 are attenuated (Arnal, Morillon, Kell, & Giraud, 2009; Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003; Pilling, 2010; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005; Vroomen & Stekelenburg, 2010) and speeded up (van Wassenhove et al., 2005) when an auditory syllable is accompanied by visual information from the speaker's face. The speeding-up and amplitude suppression of auditory evoked potentials is thought to reflect early multisensory integrative mechanisms. Given the temporal precedence of visible speech movements on the auditory signal for isolated syllables, the observed effects on early auditory evoked potentials might be due to the increased temporal predictability of the onset of the auditory stimulus (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010) and/or might reflect specific visual phonetic prediction of the incoming auditory syllable (Arnal et al., 2009; Arnal, Wyart, & Giraud, 2011; Arnal & Giraud, 2012; van Wassenhove et al., 2005).

From these studies, one fundamental issue is whether early cross-modal speech interactions only depend on well-known auditory and visual modalities or, rather, might also be triggered by other sensory modalities, namely the auditory and haptic modalities. Audio–haptic interactions are indeed frequently experienced in daily life, with auditory and tactile stimuli often perceived simultaneously (for instance, when we scratch ourselves, rub our hands together, knock at a door, or play a musical instrument). As in the McGurk audiovisual illusion (McGurk & MacDonald, 1976), incongruities between audio and tactile inputs may even result in unexpected percepts (Jousmäki & Hari, 1998). Regarding speech, past researches on the Tadoma method demonstrate that deaf-blind individuals can understand spoken language remarkably well through the haptic modality (Alcorn, 1932; Norton et al., 1977). In this method, speech is received by placing a hand on the face of the talker in order to monitor orofacial speech movements. Interestingly, a few behavioral studies also

* Correspondence to: Avril Treille, GIPSA-LAB, UMR CNRS 5216, Grenoble Université, 1180, avenue centrale, BP 25, 38040 Grenoble Cedex 9, France. Tel.: +33 476 827 784; fax: +33 476 824 335.
E-mail address: avril.treille@gipsa-lab.inpg.fr (A. Treille).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.02.004
0028-3932 © 2014 Published by Elsevier Ltd.

provide evidence for audio–tactile speech interaction in individuals without sensory impairement, with inexperienced participants presented with syllables heard and felt from manual tactile contact with a speaker's face (Fowler & Dekle, 1991; Gick, Jóhannsdóttir, Gibraiel, & Mühlbauer, 2008; Sato, Cavé, Ménard, & Brasseur, 2010). Fowler and Dekle (1991) demonstrated the influence of tactile information on speech perception in a completely untrained population, with felt syllables affecting judgments of the syllable heard and, conversely, acoustic syllables affecting judgments of the syllable felt. Interestingly, they also found evidence for audio–haptic McGurk-type illusion but only in few participants (but see Sato et al. (2010). Gick et al. (2008) further showed that manual tactile information improves both auditory and visual speech intelligibility in noise. Similarly, Sato et al. (2010) demonstrated that manual tactile information relevant to recovering speech gestures enhances auditory speech perception in case of degraded acoustic information and that audio–tactile interactions occur similarly in blind and sighted untrained listeners.

The present electro-encephalographic study aimed at further investigating early cross-modal interactions through dyadic interactions between a listener and a speaker. We compared auditory evoked components in individuals without sensory impairement, not experienced in the Tadoma method, during auditory, audio–visual and audio–haptic speech perception during a forced-choice task between /pa/ and /ta/ syllables. To this aim, participants were seated at arm's length from an experimenter and they were instructed to manually categorize each syllable presented auditorily, visually and/or haptically.

Cross-modal speech interactions are usually thought to primarily depend on auditory and visual modalities, and have typically been attributed to the frequency with which event specific information from these two modalities are jointly encountered in daily conversation. To explore whether perceivers might integrate tactile information in auditory speech perception in a similar way as they do in visual information, we tested whether haptic and visual information from speech gestures both attenuate and speed-up early auditory evoked responses compared to auditory speech perception. Such evidence for early cross-modal interactions during both face-to-face and hand-to-face speech perception would further suggest that sensory information from speech gestures conveys predictive temporal and/or phonetic information to the incoming auditory speech input and would emphasize the multimodal nature of speech perception.

2. Methods

2.1. Participants

Two groups of fourteen and fifteen healthy adults, native French speakers, participated in the study (EEG experiment: 7 females, mean age of 34 years ± 11 years; behavioral experiment: 8 females, mean age of 28 years ± 9 years). All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. None of them was experienced in the Tadoma method.

2.2. Experimental procedure

2.2.1. EEG experiment

Early cross-modal speech interactions and auditory evoked components were first evaluated in an EEG experiment. The experimental procedure was adapted from the Tadoma method and similar to that previously used by Fowler and Dekle (1991), Gick et al. (2008) and Sato et al. (2010). Participants were individually tested in a sound-proof room and were seated at arm's length from a female experimenter (see Fig. 1A). They were told that they would be presented with /pa/ or /ta/ syllables either auditorily, visually, audio-visually, haptically, or audio-haptically over the hand–face contact.

Five modalities of presentation were tested. In the auditory modality (A), participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the experimenter. In the audio–visual modality (AV), they were asked to also look at the experimenter's face. In the audio–haptic modality (AH), they were asked to keep their eyes closed with their right hand placed on the experimenter's face (the thumb placed lightly and vertically against the experimenter's lips and the other fingers placed horizontally along the jaw line in order to help distinguishing both lip and jaw movements). The visual-only (V) and haptic-only (H) modalities were similar to the AV and AH modalities except that the experimenter silently produced each syllable. Because of no reliable acoustical triggers (see below), EEG data were not analyzed in the visual-only and haptic-only modalities.

The experimenter faced the participant and a computer screen placed behind the participant. On each trial, the computer screen specified the syllable to be produced. To this aim, the syllable was printed three times on the computer screen at 1 Hz, with the last display serving as the visual go-signal to produce the syllable. The inter-trial interval was 3 s. The experimenter previously practiced and learned to articulate each syllable in synchrony with the visual go-signal, with an initial neutral closed-mouth position and maintaining an even intonation, tempo and vocal intensity.

A two-alternative forced-choice identification task was used, with participants instructed to categorize each perceived syllable by pressing on one of two keys corresponding to /pa/ or /ta/ on a computer keyboard with their left hand. In order to dissociate sensory/perceptual responses from motor responses on EEG data, a brief single audio beep was delivered 600 ms after the visual go-signal (expecting to occur in synchrony with the experimenter production). Participants were told to produce their responses only after this audio go-signal.

The experiment included five individual experimental sessions related to each modality of presentation (A, V, H, AV, AH). Before each session, participants were informed about the modality of presentation. In each session, every syllable (/pa/ or /ta/) was presented 40 times in a randomized sequence for a total of 80 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities. They received no instructions concerning how to interpret visual and haptic information but they were asked to pay attention to both modalities during bimodal presentation. Because the experimental procedure was quite taxing for the experimenter and the participants, short breaks were offered between each experimental session.

Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the visual stimuli for the experimenter, the audio stimuli (beep) for the participant and to record key responses. In addition, all experimenter productions were recorded for off-line analyses.

2.2.2. Behavioral experiment

In order to test the temporal precedence of visible/tactile speech movements on the auditory signal for isolated syllables, reaction times (RTs) in a control behavioral experiment were evaluated in another group of fifteen participants during auditory, audio–visual and audio–haptic speech perception. Visual-only and haptic-only modalities were not included in the experiment because of no reliable acoustical triggers to estimate RTs. Importantly, the experimental procedure was perfectly identical to that used in the EEG experiment (with notably the same experimenter/speaker) except that the audio-go signal was removed and participants were instructed to categorize each perceived syllable as quickly as possible with their left hand. As in the EEG experiment, participants performed few practice trials in all modalities and were asked to pay attention to both modalities during bimodal presentation.

The experiment included three individual experimental sessions related to each modality of presentation (A, AV, AH). Before each session, participants were informed about the modality of presentation. In each session, every syllable (/pa/ or /ta/) was presented 20 times in a randomized sequence for a total of 40 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities.

2.3. EEG acquisition

EEG data were continuously recorded from 64 scalp electrodes (Electro-Cap International, INC., according to the international 10–20 system) using the Biosemi ActiveTwo AD-box EEG system operating at a sampling rate of 256 Hz. Two additional electrodes served as reference (Common Mode Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). One other external reference electrode was at the top of the nose. The electrooculogram measuring horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes 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.

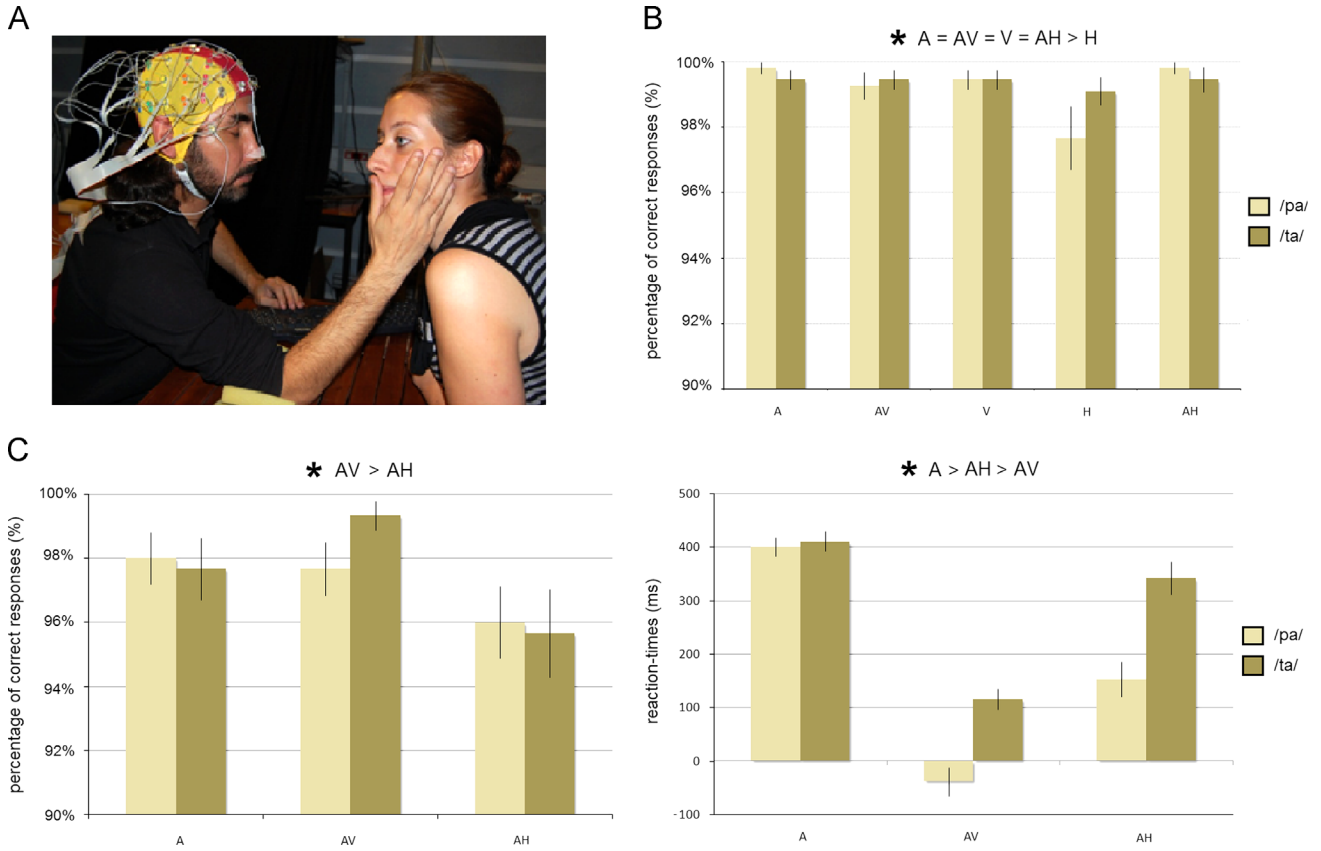

Fig. 1. (A) Experimental design used in the audio–haptic modality. Participants were asked to keep their eyes closed with their right hand placed on the experimenter's face and to categorize with their left hand each perceived syllable. (B) Mean percentage of correct identification for /pa/ and /ta/ syllables in each modality of presentation in the EEG experiment. (C) Mean percentage of correct identification and RTs (in ms) for /pa/ and /ta/ syllables in each modality of presentation in the behavioral experiment. Error bars represent standard errors of the mean.

2.4. Data analyses

2.4.1. Behavioral analyses

In both experiments, the proportion of correct responses was individually determined for each participant, each syllable and each modality. In addition, in the behavioral experiment, RTs were calculated from the consonantal onset of each produced syllable (see acoustical analyses). Two-way repeated-measure ANOVAs were performed on these data with the modality (A, V, H, AV, AH in the EEG experiment; A, AV, AH in the behavioral experiment) and the syllable (/pa/, /ta/) as within-subjects variables.

2.4.2. Acoustical analyses

In both experiments, acoustical analyses were performed on the experimenter's recorded syllables in order to determine the individual syllable onsets serving as acoustical triggers for EEG and RT analyses in the EEG and behavioral experiments, respectively. In the EEG experiment, because the experimenter silently produced the syllables in the V and H modalities, acoustical analyses were only performed for A, AV and AH modalities.

All acoustical analyses were performed using Praat software (Boersma & Weenink, 2013). A semi-automatic procedure was first devised for segmenting the experimenter's recorded syllables in the A, AV and AH modalities (5160 utterances). This procedure involved the automatic segmentation of each syllable based on an intensity and duration algorithm detection. Based on minimal duration and low intensity energy parameters, the algorithm automatically identified pauses between each syllable and set the syllable's boundaries on that basis. For each syllable, these boundaries were further hand-corrected, based on waveform and spectrogram information. Omissions and wrong productions were identified and removed from the analyses (less than 1%).

The individual syllable onsets served as acoustical triggers for EEG and RT analyses. In addition, to determine possible production differences between modalities of presentation in the EEG experiment, the mean duration, relative intensity and f0 values (calculated from a period defined as ± 25 ms of the maximum peak intensity of each syllable) averaged over /pa/ and /ta/ syllables, as well as the mean delay between the visual go-signal and the produced syllable were then calculated for each participant and each modality. These data were entered into one-way repeated-measure ANOVAs with the modality (A, AV, AH) as the within-subjects variable.

2.4.3. EEG analyses

Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only modalities. EEG data in the A, AV and AH modalities were processed using the EEGLAB toolbox (Delorme & Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over fronto-central sites on the scalp (Näätänen & Picton, 1987; Scherg & Von Cramon, 1986), and in line with previous EEG studies on audio–visual speech perception and auditory evoked potentials (e.g. Pilling (2010), Stekelenburg and Vroomen (2007), van Wassenhove et al. (2005), Vroomen and Stekelenburg (2010)), EEG data preprocessing and analyses were conducted on 6 representative frontal and central electrodes (F3, Fz, F4, C3, Cz, C4). EEG data were first re-referenced off-line to the nose recording and band-pass filtered using a two-way least-squares FIR filtering (1–20 Hz). Data were then segmented into epochs of 1000 ms including a 100 ms prestimulus baseline (from -500 ms to -400 ms to the acoustic syllable onset, individually determined from the acoustical analyses). Epochs with an amplitude change exceeding $\pm 100\,\mu V$ at any channel (including HEOG and VEOG channels) were rejected (on average, $2\% \pm 3\%$).

Because of an insufficient number of trials per syllable for reliable EEG analyses, responses from /pa/ and /ta/ syllables were averaged together. For each participant and each modality, the EEG waveforms of the six electrodes were first carefully inspected by two experimenters. Two temporal windows were then defined in order to include N1 and P2 peaks for all electrodes (on average, 90–130 ms for N1 and 180–220 ms for P2). From these temporal windows, maximal amplitude and peak latency of auditory N1 and P2 evoked responses were then determined for the 6 electrodes. The mean latencies for N1 and P2 peaks were of 116 ms (± 6 ms)/209 ms (± 7 ms), 111 ms (± 5 ms)/209 ms (± 7 ms) and 105 ms (± 6 ms)/201 ms (± 7 ms) in the A, AV and AH modalities, respectively. Three-way repeated-measure ANOVAs were performed on N1 and P2 amplitude and latency with the modality (A, AV, AH), the rostro-caudal position (frontal, central) and the medio-lateral position (left, middle, right) of the electrodes as within-subjects variables.

3. Results

For all the following 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.

3.1. Behavioral analyses (see Fig. 1)

3.1.1. EEG experiment (see Fig. 1B)

Overall, the mean proportion of correct responses was of 99%. The main effect of modality of presentation was significant ($F(4,52)=3.63$, $p<.01$), with more correct responses in the A, V, AV and AH modalities than in the H modality (as shown by post-hoc analyses, all comparisons significant; on average, A: 100%, V: 99%, AV: 99%, AH, 100%, H: 98%). No significant effect of the syllable or interaction was observed. These results thus confirm a near perfect identification of the perceived syllables in all modalities, although a slightly lower accuracy in the H modality was observed (on average, 2%).

3.1.2. Behavioral experiment (see Fig. 1C)

The mean proportion of correct responses was of 97%. The main effect of modality of presentation was significant ($F(2,28)=3.34$, $p=.05$), with more correct responses in the AV than in the AH modality (as shown by post-hoc analyses; on average, A: 98%, AV: 99%, AH: 96%. No significant effect of the syllable or interaction was observed.

Regarding RTs, the main effect of modality was significant ($F(2,28)=94.81$, $p<.001$), with faster RTs observed in the AH than in the A modality, and in the AV modality than in the AH modality (as shown by post-hoc analyses, all comparisons significant; on average, A: 406 ms, AV: 39 ms, AH: 248 ms). Faster RTs were also observed for /pa/ compared to /ta/ syllables ($F(1,14)=74.10$, $p<.001$). Finally, the interaction between the modality and the syllable was reliable ($F(2,28)=36.38$, $p<.001$), with no differences between /pa/ and /ta/ syllables in the A modality but faster RTs for /pa/ than for /ta/ in both the AV and AH modalities (as shown by post-hoc analyses; on average, A-/pa/: 401 s, A-/ta/: 411 ms, AV-/pa/: −38 ms, AV-/ta/: 116 ms, AH-/pa/: 153 ms, AH-/ta/: 343 ms).

In sum, despite near perfect syllable recognition (although lower in the AH modality), faster RTs in both the AV and AH modalities provide evidence for the temporal precedence of the dynamic configurations of the articulators on the auditory signal. RTs observed in the AV modality were around 400 ms faster than in in the auditory modality. Compared to previous studies showing that the visual signal typically precedes the onset acoustic speech signal by tens to a few hundred milliseconds (see van Wassenhove et al. (2005)), this indicates that movements of the experimenter/speaker were here highly anticipated (likely due to the experimental procedure and, more specifically, to the 1 Hz visual go-signals for the experimenter). It is also worthwhile noting that negative RTs for /pa/ syllables demonstrate that participants under time pressure (and with subsequent auditory feedback) even recognized the syllable on a visual basis. Interestingly, the haptic advantage was not as strong as the visual one. Given the causal relationship between visual and haptic onsets, this might indicate less natural and more complex processing to extract relevant speech information. Finally, faster RTs for for /pa/ than for /ta/ syllables in both the audio–visual and audio–haptic modalities are likely due to the strong visual and haptic perceptual saliencies of bilabial movements.

3.2. Acoustical analyses

The mean delay between the visual go-signal and the produced syllable was constant across A, AV and AH modalities ($F(2,26)=0.86$; on average, A: +3 ms, AV: +14 ms, AH: −12 ms). Similarly,

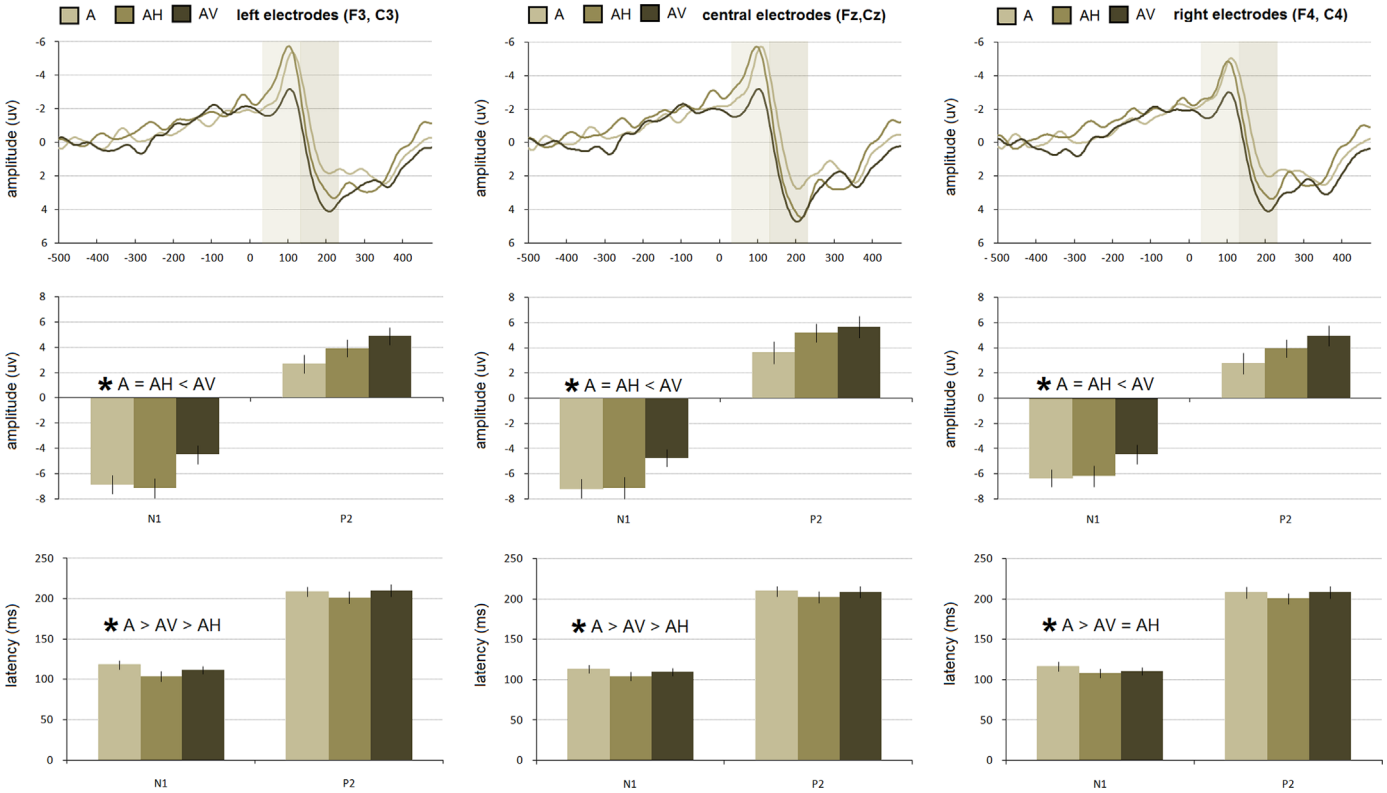

Fig. 2. Grand-average auditory evoked potentials (top), mean amplitude (in μV, middle) and mean latency (in ms, bottom) of N1 and P2 components averaged over left (F3, C3), central (Fz, Cz) and right (F4, C4) electrodes. EEG data were segmented into epochs of 1000 ms, from −500 ms to +500 ms to the acoustic syllable onset. Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only modalities. Error bars represent standard errors of the mean.

no differences were observed between modalities on the mean syllable duration ($F(2,26)=1.12$; on average, A: 194 ms, AV: 192 ms, AH: 197 ms) and intensity ($F(2,26)=3.47$; on average, A: 73 dB, AV: 73 dB, AH: 71 dB). However, the main effect of f0 was significant ($F(2,26)=5,93$, $p<.01$), with a lower f0 in the AH modality compared to A and AV modalities (as shown by post-hoc analyses, all comparisons significant; A: 241 Hz, AV: 240 Hz, AH: 237 Hz). These results show that, on average, the experimenter articulated the syllables in synchrony with the visual go-signal and maintained a quite constant intonation, tempo and vocal intensity across modalities and participants. Lower f0 observed in the AH modality is likely due to the participant's contact with the experimenter's face. However, this difference remains quite low (on average, 3 Hz) and, in our view, cannot explain latency and amplitude differences observed on EEG data between A, AV and AH modalities. Although the N1/P2 complex is known to depend on acoustic features of the speech signal, in our best knowledge, no EEG study demonstrated significant N1 and/or P2 latency and amplitude changes related to such a small f0 difference (i.e., 3 Hz) between isolated speech sounds. In addition, it can be noted that no f0 difference was observed between A and AV modalities despite significant difference on N1 amplitude and latency on EEG data between these modalities (see below).

3.3. EEG analyses—N1 amplitude (see Fig. 2—Middle)

The main effect of medio-lateral position was significant ($F(2,26)=6.49$, $p<.005$), with a reduced negative N1 amplitude observed in right electrodes as compared to left and middle electrodes (as shown by post-hoc analyses, all comparisons significant; on average, left: −6.17 μV, middle: −6.35 μV, right: −5.66 μV). Of more interest is the significant effect of modality ($F(2,26)=12.84$, $p<.001$), with a reduced negative N1 amplitude observed for the AV modality as compared to both A and AH modalities (as shown by post-hoc analyses, all comparisons significant; on average, A: −6.80 μV, AH: −6.84 μV, AV: −4.55 μV). The interaction between the modality and the medio-lateral position of electrodes was also reliable ($F(4,52)=4.33$, $p<.005$). For both A and AH modalities, a reduced negative N1 amplitude was observed in right electrodes as compared to both left and middle electrodes (as shown by post-hoc analyses, all comparisons significant; on average, A-left: −6.87 V, A-middle: −7.17 μV, A-right: −6.36 μV, AH-left: −7.16 μV, AH-middle: −7.13 μV, AH-right: −6.18 μV). However, for the AV modality, no differences were observed (on average, AV-left: −4.49 V, AV-middle: −4.73 μV, AV-right: −4.44 μV). No other effect or interactions were found to be significant.

The main effect of medio-lateral position confirms that auditory N1 had a central maximum (Näätänen & Picton, 1987; Scherg & Von Cramon, 1986) and possibly indicate left-lateralized auditory responses. Of more interest, the main effect of modality appears in line with previous EEG studies on audio–visual speech perception and confirm a visually-induced amplitude suppression of the auditory evoked N1 component. Interestingly, no haptically-induced amplitude suppression was observed, with similar amplitude for A and AH modalities and a higher negative N1 amplitude observed for the AH modality compared to the AV modality.

3.4. EEG analyses—N1 latency (see Fig. 2—Bottom)

The main effect of modality was significant ($F(2,26)=4.62$, $p<.02$), with a shorter negative N1 peak latency observed for the AH modality compared to the A modality (as shown by post-hoc analyses, all p's $<.05$; on average, A: 116 ms, AH: 105 ms, AV: 111 ms). The fact that the main effect did not provide evidence for shorter auditory evoked responses in the AV modality compared to the A modality is probably due response variability between the medio-lateral position of the electrodes. Indeed, a significant interaction between the modality and the medio-lateral position of electrodes ($F(4,52)=5.89$, $p<.001$) further demonstrate a shorter negative N1 peak latency for both AH and AV modalities compared to the A modality. Posthoc analyses showed that, in the left and middle electrodes, a shorter negative N1 peak latency was observed for the AH modality compared to the AV modality, and for the AV modality compared to the A modality (all p's $<.05$; on average, A-left: 118 ms, AV-left: 111 ms, AH-left: 104 ms, A-middle: 113 ms, AV-middle: 110 ms, AH-middle: 104 ms). In the right electrodes, a shorter negative N1 peak latency was observed for both the AH and AV modalities compared to the A modality (all p's $<.05$; on average, A-right: 116 ms, AV-right: 110 ms, AH-right: 108 ms). No other effects or interactions were significant. Altogether, these results appear in line with previous EEG studies with a visually-induced speeding-up of the auditory evoked N1 component. Similarly, a shorter negative N1 peak latency was also observed for the AH modality compared to the A modality, and compared to the AV modality in the left and middle electrodes.

3.5. EEG analyses—P2 amplitude and latency (see Fig. 2—Middle and bottom)

The analysis on P2 amplitude showed a significant effect of the medio-lateral position ($F(2,26)=14.56$, $p<.001$), with an higher positive P2 amplitude observed in middle electrodes as compared to both left and right electrodes (all p's $<.05$; on average, left: 3.83 μV, middle: 4.81 μV, right: 3.88 μV). In the same line to what was found with N1 amplitude, this effect confirms that auditory P2 had a central maximum. No other effects or interactions were found to be significant. Finally, regarding P2 peak latency, no effects or interactions were significant.

4. Discussion

In the present study, early cross-modal interactions were investigated by comparing early auditory evoked potentials and behavioral performance during auditory, audio–visual and audio–haptic speech perception in dyadic interactions between a listener and a speaker. Congruent with the possibility that face-to-face and hand-to-face dyadic interactions speed up the processing of auditory speech, reduced latency of early auditory evoked responses and faster reaction times were observed during both audio–visual and audio–haptic speech perception compared to auditory speech perception.

Before we discuss these results, it is important to consider a clear limitation of the present study. Since face-to-face and hand-to-face speech perception were here examined in live dyadic interactions, we used individual syllable onsets of the experimenter's productions as acoustical triggers for EEG and RT analyses. For the visual-only and haptic-only modalities, the use of electromyographic and/or visual recordings of the experimenter's lip movement would not allowed to determine such reliable triggers, due to the variability or temporal limitation of these signals (but see Leotta, Rabinowitz, Reed, and Drulach (1988), for a synthetic Tadoma system which allows realistic and precisely timed synthetic facial inputs). Because of no reliable triggers, these two unimodal conditions were not analyzed in the EEG experiment and were not included in the behavioral RT experiment. For that reason, we could not use an additive model (i.e., $AV \neq A+V$) to determine whether the results observed in the audio–visual and audio–haptic modalities simply come from a superposition of the sum of the unimodal signals or truly reflect crossmodal interactions. Notably, without EEG recordings in the visual-only and

haptic-only modalies, it is impossible to determine whether the observed N1 and P2 auditory evoked potentials in the audio–visual and audio–haptic modalities are not contamined by visual and haptic ERPs. From that question, although auditory ERPs are rarely observed in the visual-only modality in fronto-central electrodes (see Besle et al. (2004), Pilling (2010), van Wassenhove et al. (2005)), haptic ERPs (notably, the P30, P40, P100 and N140 components) are known to also occur in fronto-central electrodes (e.g., Desmedt and Tomberg (1989)). Importantly, it should be noted that RTs observed in the behavioral experiment strongly suggest that visual, and therefore haptic, speech movements of the experimenter/speaker started well before (around 400 ms) the acoustical onset of the syllables. Given the difference between sensory onsets, it is therefore unlikely that haptic (and visual) ERPs might arise at the same time-latency of auditory ERPs in the audio–haptic (and audio–visual) modality on fronto-central electrodes. Hence, although our results cannot fully demonstrate early bimodal integration mechanisms in the audio–visual and audio–haptic modalities, they strongly suggest that haptic and visual information speed up the neural processing of auditory speech.

In spite of this important limitation, our results appears fully in line with previous EEG studies on audio–visual speech perception (Besle et al., 2004; Klucharev et al., 2003; Pilling, 2010; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010), with N1 auditory evoked potentials attenuated and speeded up during audio–visual compared to auditory speech perception. Given the temporal precedence of visible speech movements on the auditory signal for isolated syllables (the visual signal, and therefore the haptic signal, preceding the onset acoustic speech signal by tens to a few hundred milliseconds during individual syllable production; see van Wassenhove et al. (2005)), the speeding-up and amplitude suppression of auditory evoked potentials likely reflect early multisensory integrative mechanisms. Despite near perfect recognition, this temporal precedence of the dynamic configurations of the articulators on the auditory signal is attested in the behavioral experiment. Indeed, faster RTs were observed in both the audio–visual and audio–haptic modalities compared to the unimodal auditory modality.

Crucially, although participants were not experienced with audio–haptic speech perception, haptic information was also found to speed up auditory speech processing, with a shorter latency of N1 auditory evoked potentials in audio–haptic compared to auditory speech perception. Compared to a strong visually-induced N1 amplitude suppression observed in the present experiment and in previous studies on audio–visual speech perception, no haptically-induced N1 amplitude suppression was however observed. Although speculative, one possibility is that this difference might partly be explained by higher attentional demands in the audio–haptic modality, which is well known to enhance amplitude of early auditory evoked potentials (Näätänen, 1992; Giard et al., 2000).

Importantly, two qualitatively different integrative mechanisms, although not mutually exclusive, can been proposed to explain these findings. A first mechanism implies that early cross-modal audio–visual and audio–haptic interactions are not speech specific and rather depend on the temporal relationship of sensory input. Congruent with this hypothesis, Stekelenburg and Vroomen (2007) demonstrated visually-induced amplitude suppression and latency reduction of auditory-evoked N1 responses during audio–visual perception of both speech and non speech actions, like clapping hands. They further showed that early cross-modal interactions are not restricted to actions but can be observed also with artificial stimuli if their timing is made predictable (like two moving disks predicting a pure tone; Vroomen and Stekelenburg (2010)). It can be also noted that audio–tactile interactions are not restricted to speech and have been previously observed with simultaneous presented tone and vibration (e.g., Lütkenhöner, Lammertmann, Simões, and Hari (2002)). A second mechanism implies the existence of specific phonetic prediction, extracted from the visual signal, of the incoming auditory speech target. From that view, early auditory-visual latency facilitation has been shown to correlate with visual identification of the speech stimuli, with stronger latency facilitation coupled with higher visual identification. In a previous study by van Wassenhove et al. (2005), auditory-visual facilitation effects were shown to systematically vary according to the identification scores observed in the visual modality. In their study, a higher visual accuracy was observed for /pa/ compared to /ta/ syllables, and for /ta/ compared to /ka/ syllables. Consistent with an articulator-specific facilitation, latency of auditory evoked potentials were found to be shorter for /pa/ then for /ta/ syllables, and for /ta/ than for /ka/ syllables (for similar results using MEG, see also Arnal et al. (2009)). It is worthwhile noting that a potential limit of these studies comes from the use of a limited number of tokens used to represent each syllable, repeatedly presented to the participants and possibly enhancing stimulus predictability. Conversely, one clear limit of the present study comes from the use of a forced-choice task between /pa/ and /ta/ syllables and a ceiling effect on perceptual scores. Although our results do not contradict the hypothesis that sensory inputs convey phonetic predictive information with respect to the incoming auditory speech input, future studies using a larger sample of syllables are therefore required to test whether haptic information is used to specifically predict the incoming auditory syllable.

Despite the above-mentioned limitations of this study, the present and previously observed audio–haptic advantages in untrained participants may have profound implications for further understanding the basis of cross-modal speech integration (Fowler & Dekle, 1991; Gick et al., 2008; Sato et al., 2010). Since cross-modal speech interactions have previously been attributed to the frequency with which event specific information from auditory and visual modalities are jointly encountered in daily conversation, evidence for audio–haptic speech interactions, although less natural, further emphasize the multisensory and predictive nature of speech perception. It is first important to recall that, apart from speech, auditory and tactile stimuli are usually perceived simultaneously in daily life (as when knowing at a door). Although largely unexplored compared to audio–visual interactions, previous studies have provided evidence for an efficient integration of auditory and somatosensory processing at both the behavioral and neural levels (e.g., Desmedt and Tomberg (1989), Jousmäki and Hari (1998)). Importantly, the integration of sensory inputs can be viewed as a key feature of action control, as when we first knock on a door and then adjust the force applied in the subsequent knock (Wolpert, Ghahramani, & Jordan, 1995). Regarding speech, audio–haptic interactions likely depend on the temporal precedence of the dynamic configurations of the articulators on the auditory signal (as attested in the behavioral experiment). As previously mentioned, despite less natural and apparently more complex processing to extract relevant speech information from the haptic modality, this temporal precedence might increase the temporal predictability of the onset of the auditory stimulus and/or provide specific phonetic prediction of the incoming auditory syllable. The possibility that the brain might extract predictive temporal and/or phonetic relevant information for auditory processing when being provided with such information, although we rarely if ever touch the speaker's face to understand speech, raised important question on the representational format of speech. From that point, haptic perception is likely to be partly driven by listener's knowledge of speech production (Sato et al., 2010; Schwartz, Ménard, Basirat, & Sato, 2012), as for example

when relating the tactile sensation of lip protusion to the production of /pa/ syllables. Although speculative, thanks to the temporal precedence of tactile inputs, the observed audio–haptic interactions might also partly arise from auditory, visual and/or motor imagery processes and, possibly, a crossmodal mapping between the related sensory and motor speech representations.

In conclusion, our results suggest early integrative mechanisms between auditory, visual and haptic modalities. Since audio–haptic and audio–visual speech interactions were never assessed at the brain level in dyadic interactions between a listener and a speaker, these results provide new insights on the multisensory nature of speech perception.

Acknowledgments

This study was supported by research grant from the Centre National de la Recherche Scientifique (CNRS). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency. We thank Jean-Luc Schwartz and Marco Congedo for helpful discussions on multisensory speech perception and on EEG analyses.

References

Alcorn, S. (1932). The Tadoma method. *Volta Review*, *34*, 195–198.
Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Science*, *16*(7), 390–398.
Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience*, *29*(43), 13445–13453.
Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801.
Benoît, C., Mohamadi, T., & Kandel, S. D. (1994). Effects on phonetic context on audio–visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*, 1195–1203.
Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European journal of Neuroscience*, *20*, 2225–2234.
Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer. Computer program*, Version 5.3.42, retrieved 2 March 2013 from ⟨http://www.praat.org/⟩.
Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, *134*, 9–21.
Desmedt, J. E., & Tomberg, C. (1989). Mapping early somatosensory evoked potentials in selective attention: Critical evaluation of control conditions used for titrating by difference the cognitive P30, P40, P100 and N140. *Electroencephalography and Clinical Neurophysiology*, *74*(5), 321–346.
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.
Giard, M. H., Fort, A., Mouchetant-Rostaing, Y., & Pernier, J. (2000). Neurophysiological mechanisms of auditory selective attention in humans. *Front. Bioscience*, *5*, 84–94.
Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of Acoustical Society of America*, *123*, 72–76.
Jousmäki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology*, *8*(6), 190.
Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research. Cognitive Brain Research*, *18*, 65–75.
Leotta, D. F., Rabinowitz, W. M., Reed, C. M., & Drulach, N. I. (1988). Preliminary results of speech-reception tests obtained with the synthetic Tadoma system. *Journal of Rehabilitation Research and Development*, *25*(4), 45–52.
Lütkenhöner, B., Lammertmann, C., Simões, C., & Hari, R. (2002). Magnetoencephalographic correlates of audiotactile interaction. *NeuroImage*, *15*(3), 509–522.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.
Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Lawrence Erlbaum Associates.
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.
Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12.
Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M., et al. (1977). Analytic study of the Tadoma method: Background and preliminary results. *Journal of Speech and Hearing Research*, *20*, 574–595.
Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1073–1081.
Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard tounderstand: A lipreading advantage with intact auditory stimuli. In: R. Campbell, & B. Dodd (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 97–113). London (UK): Lawrence Erlbaum Associates.
Sato, M., Cavé, C., Ménard, L., & Brasseur, L. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, *48*(12), 3683–3686.
Scherg, M., & Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalography and Clinical Neurology*, *65*, 344–360.
Schwartz, J. L., Ménard, L., Basirat, A., & Sato, M. (2012). The perception for action control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354.
Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after summerfield: A taxonomy of models for audio–visual fusion in speech perception In: R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). Hove, UK: Psychology Press.
Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*, 1964–1973.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America*, *26*, 212–215.
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 of the United States of America*, *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.
Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.