ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

The touch of your lips: haptic information speeds up auditory speech processing

Avril Treille, Camille Cordeboeuf, Coriandre Vilain, Marc Sato

GIPSA-lab, Department of Speech & Cognition, CNRS & Grenoble University, Grenoble, France
avril.treille@gipsa-lab.inpg.fr, marc.sato@gipsa-lab.inpg.fr

Abstract

The human ability to follow speech gestures through the visual modality is a core component of speech perception. Remarkably, 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 study, early cross-modal interactions were investigated by comparing early auditory evoked potentials during auditory, audio-visual and audio-haptic speech perception in natural dyadic interactions between a listener and a speaker. Although participants were not experienced with audio-haptic speech perception, shortened latencies of auditory evoked potentials were observed in both audio-visual and audio-tactile modalities compared to the auditory modality. These results demonstrate early cross-modal interactions during face-to-face and hand-to-face speech perception and highlight a predictive role of visual and haptic information on auditory speech processing in dyadic interactions.

Index Terms: audio-visual speech perception, audio-haptic speech perception, EEG.

1. Introduction

Although humans are proficient to extract phonetic features from the acoustic signal alone, interactions between auditory and visual modalities are beneficial in daily conversation. Notably, visual information is known to effectively improve speech perception in noise, the understanding of a semantically complex statement or a foreign language [1-3]. 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, recent electroencephalographic and magneto-encephalographic studies demonstrate that, as early as 100ms, auditory evoked potentials are attenuated and speeded up when an auditory syllable is accompanied by visual information from the speaker's face [4-7]. Given the temporal advantage of vision on the auditory signal during individual syllable production, the speeding-up and amplitude suppression of auditory evoked potentials is thought to reflect early multisensory integrative mechanisms reflecting visual prediction of the auditory syllable.

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 sources. From that question, researches on the Tadoma method demonstrate that deaf-blind individuals can understand spoken language remarkably well through the haptic modality [8-9]. 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 provide evidence for audio-tactile speech interaction in normally sensed adults, with inexperienced participants presented with syllables heard and felt from manual tactile contact with a speaker's face [10-12]. In keeping with these findings, evidence for cross-modal interactions during both face-to-face and hand-to-face speech perception would strength the hypothesis that sensory information from speech gestures conveys predictive information to the incoming auditory speech input.

The present electroencephalographic studies aimed at further investigating early cross-modal interactions through natural dyadic interactions between a listener and a speaker. We compared auditory evoked components in normally sensed adults, 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 (Experiment 1) or between /pa/, /ta/ or /ka/ syllables (Experiment 2). 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. In an auditory condition, participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the experimenter. In an audio-visual condition, they were asked to look at the experimenter's face. In an audio-haptic condition, they were asked to keep their eyes closed with their right hand placed on the experimenter's lips and jaw.

2. Method

2.1. Participants

Fourteen and eleven healthy adults, native French speakers, participated in Experiments 1 and 2. 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. A few participants entered both studies which were performed at least six months apart.

2.2. Procedure

In both experiments, the experimental procedure was adapted from the Tadoma method and similar to that previously used by Fowler and Dekle [10], Gick et al. [11] and Sato et al. [12]. Participants were individually tested in a sound-proof room and were seated at arm's length from a female experimenter (see Figure 1A). They were told that they would be presented with /pa/ or /ta/ syllables in Experiment 1, or with /pa/, /ta/ or /ka/ syllables in Experiment 2, either auditorily, visually and/or haptically over the hand-face contact.

In Experiment 1, five experimental conditions were tested. In an auditory condition (A), participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the experimenter. In an audio-visual condition (AV), they were asked to also look at the experimenter's face. In an audio-haptic condition (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) conditions were similar as the AV and AH conditions except that the experimenter silently produced each syllable. Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only conditions.

In Experiment 2, the same five experimental conditions were first performed in a behavioral session, without EEG acquisition. Then after, an EEG session was performed, including an auditory condition (A), an audio-visual condition (AV) and an audio-haptic condition (AH). Except these differences and the use of /pa/, /ta/ and /ka/ syllables, the experimental protocol was similar to that used in Experiment 1.

In both experiments, 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 1Hz, with the last display serving as the visual go-signal to produce the syllable. The intertrial interval was 3s. 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.

Two-alternative or three-alternative forced-choice identification tasks were used in Experiments 1 and 2, respectively, with participants instructed to categorize each perceived syllable by pressing on one key corresponding to /pa/ or /ta/ in Experiment 1, or to /pa/, /ta/ or /ka/ in Experiment 2, 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 600ms 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.

Experiment 1 consisted on five individual experimental sessions related to each modality of presentation (A, V, H, AV, AH). In each session, every syllable (/pa/ or /ta/) was presented 40 times in a randomized sequence for a total of 80 trials. Experiment 2 first consisted on five individual behavioral sessions related to each modality of presentation (A, V, H, AV, AH). In each session, every syllable (/pa/, /ta/ or /ka/) was presented 15 times in a randomized sequence for a total of 45 trials. In a subsequent EEG session, three individual experimental sessions related to each modality of presentation (A, AV, AH) were performed. In each session, every syllable (/pa/, /ta/ or /ka/) was presented 80 times in a randomized sequence for a total of 240 trials. In each experiment, the order of the modality of presentation and the response key designation were fully counterbalanced across participants.

Before the experiments, 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.

2.3. 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.

2.4. Data analyses

In Experiment 1, because the experimenter silently produced the syllables in the V and H conditions, acoustical analyses were only performed for A, AV and AH modalities. Because of no reliable acoustical triggers, EEG data were not analyzed in the visual-only and haptic-only conditions. Similarly, in Experiment 2, acoustical analyses were performed for A, AV and AH modalities from the EEG session.

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.

2.4.1. Acoustical analyses

All acoustical analyses were performed using Praat software [13]. A semi-automatic procedure was first devised for segmenting the experimenter's recorded syllables in the A, AV and AH conditions (3360 utterances in Experiment 1 and 7920 utterances in Experiment 2). This procedure involved the automatic segmentation of each vowel 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 vowel's boundaries on that basis. For each syllable, these boundaries were further hand-corrected, based on waveform and spectrogram information, with the individual syllable onsets serving as acoustical triggers for EEG analyses. Omissions and wrong productions were identified and removed from the analyses (less than 1% in Experiments 1 and 2).

In both experiments, in order to determine possible production differences between modality of presentation (A, AV, AH), the mean intensity and F_0 values averaged over syllables were calculated for each participant and each modality. These data were entered into repeated-measure ANOVAs with the modality (A, AV, AH) as within-subjects variable.

2.4.2. Behavioral analyses

In Experiment 1, the proportion of correct responses was individually determined for each participant, each syllable and each modality. A repeated-measure ANOVAs was performed on these data with the modality (A, V, H, AV, AH) and the syllable

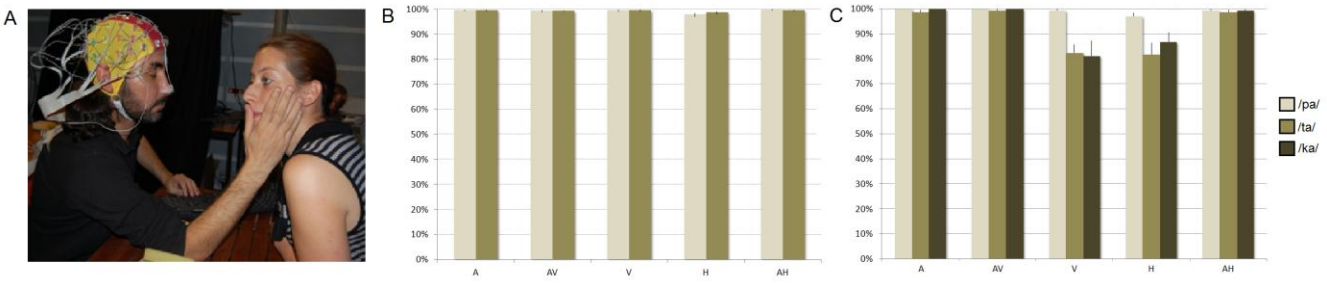

Figure 1: *(A) Experimental design used in the audio-haptic condition. (B-C) Mean percentage of correct identification in each modality of presentation in Experiments 1 (B: for /pa/ and /ta/ syllables) and 2 (C: for /pa/, /ta/ and /ka/ syllables). Error bars represent standard errors of the mean.*

(/pa, /ta/) as within-subjects variables. In Experiment 2, the proportion of correct responses was individually determined for each participant in the behavioral session, each syllable and each modality. A repeated-measure ANOVAs was performed on these data with the modality (A, V, H, AV, AH) and the syllable (/pa, /ta/, /ka/) as within-subjects variables.

2.4.3. EEG analyses

In both experiments, EEG data in the A, AV and AH conditions were processed using the EEGLAB toolbox [14] running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over fronto-central sites on the scalp [4-6], 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 (1-20Hz). Data were then segmented into epochs of 1000ms including a 100ms prestimulus baseline (from -500ms to -400ms to the acoustic syllable onset, individually determined from the acoustical analyses). Epochs with an amplitude change exceeding ±100 µV at any channel (including HEOG and VEOG channels) were rejected (less than 5% in Experiments 1 and 2).

In each experiment, maximal amplitude and peak latency of auditory N1 and P2 evoked responses were individually determined for each participant, each modality and each electrode. Because of an insufficient number of trials per syllable for reliable EEG analyses, responses from /pa/ and /ta/ syllables were averaged together in Experiment 1. 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. In Experiment 2, repeated-measure ANOVAs were performed on N1 and P2 amplitude and latency with the modality (A, AV, AH), the syllable (/pa/, /ta/, /ka/), the rostro-caudal position (frontal, central) and the medio-lateral position (left, middle, right) of the electrodes as within-subjects variables.

3. Results

3.1. Experiment 1

3.1.1. Acoustical analyses

No differences were observed between modalities on mean syllable intensity ($F(2,26)=3.47$; A: 73dB, AV: 73dB, AH: 71dB). However, mean F_0 was significantly lower in the AH condition compared to A and AV conditions ($F(2,26)=5.93$, $p < .01$; A: 241Hz, AV: 240Hz, AH: 237Hz). This difference remains however quite low and cannot explain latency and amplitude differences observed on EEG data between A, AV and AH modalities.

3.1.2. Behavioral analyses (Figure 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 conditions than in the H condition (on average, A: 100%, V: 99%, AV: 99%, AH, 100%, H: 98%). No significant effect of the syllable or interaction was observed.

3.1.3. EEG analyses – N1 amplitude (Figure 2A)

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 (on average, left: -6.17µV, middle: -6.35µV, right: -5.66µV).

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 (on average, A: -6.80µV, AH: -6.84µV, AV: -4.55µV). 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 (on average, A-left: -6.87V, A-middle: -7.17µV, A-right: -6.36µV, AH-left: -7.16µV, AH-middle: -7.13µV, AH-right: -6.18µV). However, for the AV modality, no differences were observed (on average, AV-left: -4.49V, AV-middle: -4.73µV, AV-right: -4.44µV). No other effect or interactions were found to be significant.

These results thus appear 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.

3.1.4. EEG analyses – N1 latency (Figure 2A)

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 (on average, A:

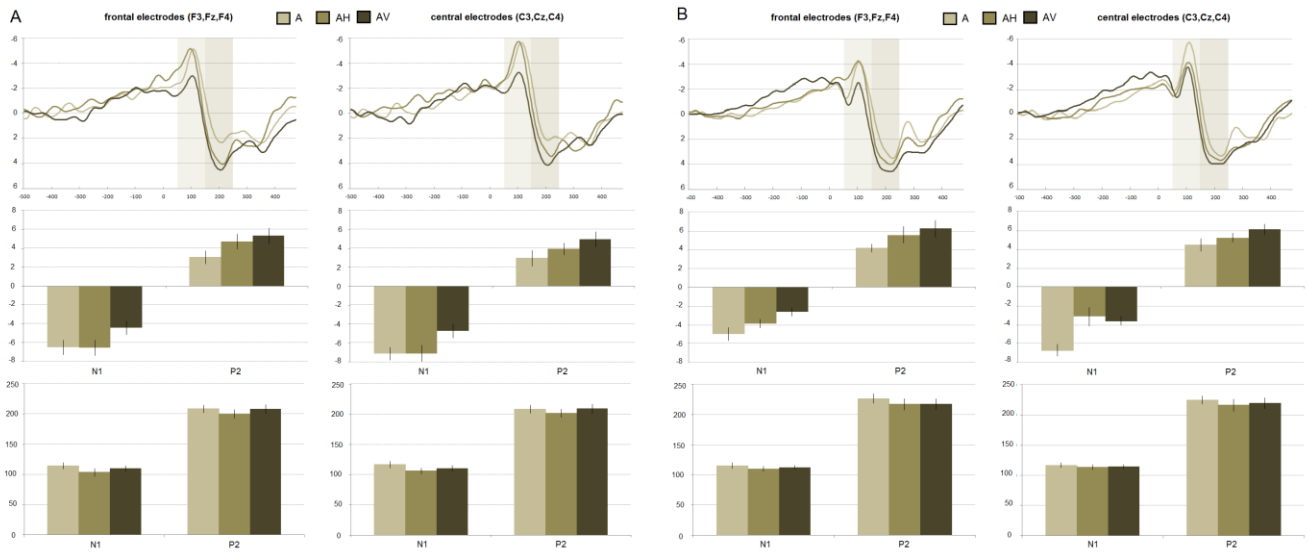

Figure 2: *Grand-average auditory evoked potentials (top), mean amplitude (in µV, middle) and mean latency (in ms, bottom) of N1 and P2 auditory components averaged over frontal (F3, Fz, F4) and central (C3, Cz, C4) electrodes in A, AV and AH conditions in Experiments 1 (A) and 2 (B). Error bars represent standard errors of the mean.*

116ms, AH: 105ms, AV: 111ms). 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 (on average, A-left: 118ms, AV-left: 111ms, AH-left: 104ms, A-middle: 113ms, AV-middle: 110ms, AH-middle: 104ms). In the right electrodes, a shorter negative N1 peak latency was observed for both the AH and AV modalities compared to the A modality (on average, A-right: 116ms, AV-right: 110ms, AH-right: 108ms). No other effects or interactions were significant.

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, even, compared to for the AH modality compared to the AV modality in the left and middle electrodes.

3.1.5. EEG analyses – P2 amplitude and latency (Figure 2A)

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 (on average, left: 3.83µV, middle: 4.81µV, right: 3.88µV). No other effects or interactions were found to be significant. Finally, regarding P2 peak latency, no effects or interactions were significant.

3.2. Experiment 2

3.2.1. Acoustical analyses

The mean syllable intensity was slightly higher in A compared to AV and AH conditions ($F(2,20)=27.32$, $p<.001$; A: 78dB, AV: 75dB, AH: 73dB). As in experiment 1, mean F_0 was significantly lower in AH compared to A and AV conditions ($F(2,20)=24.86$, $p<.0001$; A: 263Hz, AV: 256Hz, AH: 253Hz). It is to note that these differences remains however quite low and cannot explain latency and amplitude differences observed on EEG data between A, AV and AH modalities.

3.2.2. Behavioral analyses (Figure 1C)

Overall, the mean proportion of correct responses was of 95%. The main effect of modality of presentation was significant ($F(4,40) = 20.29$, $p < .001$), with more correct responses in the A, AV and AH conditions than in the V and H conditions (on average, A: 100%, AV: 100%, AH, 99%, V: 88%, H: 88%). The main effect of syllable was also significant, with more correct responses for /pa/ syllable than for /ta/ and /ka/ syllables ($F(2,20)=11.00$, $p<.001$; on average, pa: 99%, ta: 92%, ka: 93%). Finally, a significant interaction between the syllable and the modality of presentation was due to less correct responses for /ta/ and /ka/ syllables in V and H conditions compared to all other conditions ($F(8,80)=4.80$, $p<.001$; on average, V-ta: 82%, V-ka: 81%, H-ta: 82%, H-ka: 87%).

3.2.3. EEG analyses – N1 amplitude (Figure 2B)

The main effect of medio-lateral position was significant ($F(2,20) = 7.73$, $p < .004$), with a reduced negative N1 amplitude observed in right electrodes as compared to left and middle electrodes (on average, left: -4.43µV, middle: -4.55µV, right: -3.87µV). The main effect of rostro-caudal position was also significant ($F(1,10)=23.81$, $p<.001$) with a reduced negative amplitude observed in frontal as compared to central electrodes

(on average, frontal : -3.76 µV, central : -4.80 µV). In addition, the interaction between the modality and the medio-lateral position of electrodes was also reliable ($F(4,40) = 4.32$, $p < .006$). For both A and AH modalities, a reduced negative N1 amplitude was observed in right electrodes as compared to both left and middle electrodes while, for the AV modality, no significant differences were observed (on average, A-left: -5.35V, A-middle: -5.47µV, A-right: -4.85µV, AH-left: -4.78µV, AH-middle: -4.83µV, AH-right: -3.83µV, AV-left: -4.49µV, AV-middle: -4.73µV, AV-right: -4.44µV).

Of more interest and consistent with Experiment 1, the main effect of modality was significant ($F(2,20) = 7.71$, $p < .004$), with a reduced negative N1 amplitude observed for the AV modality as compared to both A and AH modalities (on average, A: -5.23µV, AH: -4.48µV, AV: -3.14µV). Furthermore, the interaction between the modality and the rostro-caudal position of electrodes was found to be also reliable ($F(2,20)=19.42$, $p<.0001$). In frontal electrodes, a reduced negative N1 amplitude was observed for AV as compared to A and AH modalities while, in central electrodes, a reduced negative N1 amplitude was observed for AV as compared to AH, and to AH as compared to A (on average, A-frontal: -4.42µV, AH-frontal: -4.43µV, AV-frontal: -2.44µV, A-central: -6.02µV, AH-central: -4.53µV, AV-central: -3.84µV). Finally, the interaction between the modality and the syllable was also reliable ($F(4,40)=4.15$, $p<.006$). For /pa/ syllable, a reduced negative N1 amplitude was observed for AV and AH as compared to A while, for /ta/ and /ka/ syllables, a reduced negative N1 amplitude was observed for AV as compared to A and AH (on average, A-pa: -5.73µV, A-ta:-4.84µV, A-ka: -5.10µV, AH-pa: -3.04µV, AH-ta: -5.32µV, AH-ka: -4.90 µV, AV-pa: -3.21µV, AV-ta: -3.24µV, AV-ka: -3.14µV). No other effects or interactions were significant.

Altogether, these results thus confirm a visually-induced amplitude suppression of the auditory evoked N1 component and appear in line with results observed in previous EEG studies and in Experiment 1. Interestingly, a haptically-induced amplitude suppression was here observed depending on the rostro-caudal position of electrodes (i.e., in central electrodes) and on the perceived syllable (i.e., for /pa/ syllable).

3.2.4. EEG analyses – N1 latency (Figure 2B)

The main effect of medio-lateral position was significant ($F(2,20) = 4.71$, $p < .03$), with a shorter N1 peak latency observed in the middle compared to right electrodes (on average, left: 117ms, middle: 115ms, right: 118ms). The interaction between the medio-lateral and rostro-caudal positions of electrodes was also reliable ($F(2,20) = 9.79$, $p < .002$). For frontal electrodes, no significant differences were found while, for central electrodes, a shorter N1 peak latency was observed for middle compared to left, and for left compared to right electrodes (on average, frontal-left: 115ms, frontal-middle: 115ms, frontal-right: 117ms, central-left: 117ms, central-middle: 113ms, central-right: 119ms).

Crucially, a significant interaction between the modality of presentation, the syllable and the rostro-caudal position of electrodes was observed ($F(4,40) = 3.45$, $p < .02$). For frontal electrodes, a shorter N1 peak latency was observed for /ka/ syllable in audio-visual and audio-haptic modalities compared to the auditory modality (on average: frontal-A-ka: 125ms, frontal-AV-ka: 110ms, frontal-AH-ka: 115ms) while, for central electrodes, a shorter N1 peak latency was observed for both /pa/ and /ka/ syllables in audio-visual and audio-haptic modalities compared to the auditory modality (on average: central-A-pa: 121ms, central-A-ka: 122ms, central-AV-pa: 115ms, central-AV-ka: 115ms, central-AH-pa: 114ms, central-AH-ka: 117ms). No other effects or interactions were significant.

In sum, a shorter N1 latency was observed in audio-visual and audio-haptic modalities depending on the rostro-caudal position of electrodes and the perceived syllable (i.e., in frontal electrodes for /ka/ and in central electrodes for /pa/ and /ka/).

3.2.5. EEG analyses – P2 amplitude and latency (Figure 2B)

The analysis on P2 amplitude showed a significant effect of the medio-lateral position ($F(2,20) = 16.64$, $p < .0001$), with an higher positive P2 amplitude observed in middle electrodes as compared to both left and right electrodes (on average, left: 4.38µV, middle: 5.57µV, right: 4.50µV). A significant interaction between the rostro-caudal position and the medio-lateral position of electrodes further demonstrate a lower positive P2 amplitude for left frontal compared to right frontal electrodes and for right frontal compared to middle frontal electrodes, as well as for left and right central electrodes compared to central middle electrodes ($F(2,20) = 3.90$, $p < .037$; on average, left-frontal: 4.28 µV, middle-frontal: 5.50 µV, right-frontal: 4.65 µV, left-central: 4.49 µV, middle-central: 5.54 µV, right-central: 4.35 µV). No other effects or interactions were significant. Finally, regarding P2 peak latency, no effects or interactions were significant.

4. Discussion and Conclusions

In two electroencephalographic studies, early cross-modal interactions were investigated by comparing early auditory evoked potentials during auditory, audio-visual and audio-haptic speech perception in natural dyadic interactions between a listener and a speaker.

In line with previous studies [4-6], results from both experiments demonstrate that N1 auditory evoked potentials are attenuated and speeded up during audio-visual compared to auditory speech perception. Given the temporal advantage of vision on the auditory signal during individual syllable production, the speeding-up and amplitude suppression of auditory evoked potentials likely reflect early multisensory integrative mechanisms reflecting visual prediction of the auditory syllable [4-7].

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. However, compared to a strong visually-induced amplitude suppression observed in both experiments and in previous studies, a haptically-induced amplitude suppression was only observed in Experiment 2, depending on the rostro-caudal position of electrodes and on the perceived syllable. In our view, these differences might partly be explained by higher attentional demands in the audio-haptic modality, which is known to enhance amplitude of early auditory evoked potentials [7]. Despite these differences, our results provide clear evidence for early cross-modal interactions during both face-to-face and hand-to-face speech perception and

highlight a predictive role of visual and haptic information on auditory speech processing.

Finally, it is worthwhile noting that previous studies on audio-visual speech perception used a limited set of stimuli, repeatedly presented to the participants [4-6]. This is particularly important since it has been argued that latency facilitation systematically depends on the degree to which the visual signal predicts possible auditory targets. For example, in the study by van Wassenhove and colleagues [4], 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 (see also [6] for similar results). On the contrary, we did not systematically observed a correlation between visual accuracy and latency of /pa/, /ta/ and /ka/ syllables in Experiment 2. Indeed, while a higher visual accuracy was observed for /pa/ compared to /ta/ and /ka/ syllables, a shorter N1 latency was observed for /pa/ but also for /ka/ syllables in the audio-visual and audio-haptic modalities in central electrodes. In our view, our results do not contradict and even reinforce the hypothesis that sensory inputs convey predictive information with respect to the incoming auditory speech inputs. However, they also suggest that systematic conclusions on sensory predictability have to be taken with caution when using a limited set of speech stimuli.

In conclusion, our results demonstrate early integrative mechanisms between auditory, visual and haptic modalities and highlight a predictive role of haptic and visual information from speech gestures on auditory speech processing. The observed cross-modal interactions during face-to-face and hand-to-face speech perception likely suggest that multisensory speech perception is partly driven by listener's knowledge of speech production [15].

5. References

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

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

[3] Navarra, J. and Soto-Faraco, S. "Hearing lips in a second language: visual articulatory information enables the perception of second language sounds". Psychological research. 2005.

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

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

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

[7] Arnal, L.H., Wyart, V. and Giraud, A.L. "Transitions in neural oscillations reflect prediction errors generated in audiovisual speech". Nature Neuroscience, 14(6):797-801, 2011.

[8] Alcorn, S. "The Tadoma method". Volta Review, 34:195-198, 1932.

[9] Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M., et al. "Analytic study of the Tadoma method: Background and preliminary results". Journal of Speech and Hearing Research, 20, 574–595.

[10] Fowler, C. and Dekle, D. "Listening with eye and hand: crossmodal contributions to speech perception". Journal of Experimental Psychology- Human Perception and Performance, 17:816–828, 1991.

[11] Gick, B., Jóhannsdóttir, K.M., Gibraiel, D. and Mühlbauer, M. "Tactile enhancement of auditory and visual speech perception in untrained perceivers". Journal of Acoustical Society of America, 123:72-76, 2008.

[12] Sato, M., Cavé, C., Ménard, L. and Brasseur, L. "Auditory-tactile speech perception in congenitally blind and sighted adults". Neuropsychologia, 48(12): 3683-3686, 2010.

[13] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer". Computer program, Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/, 2013.

[14] Delorme, A. and Makeig, S. "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics". Journal of Neuroscience Methods, 134:9-21, 2004.

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