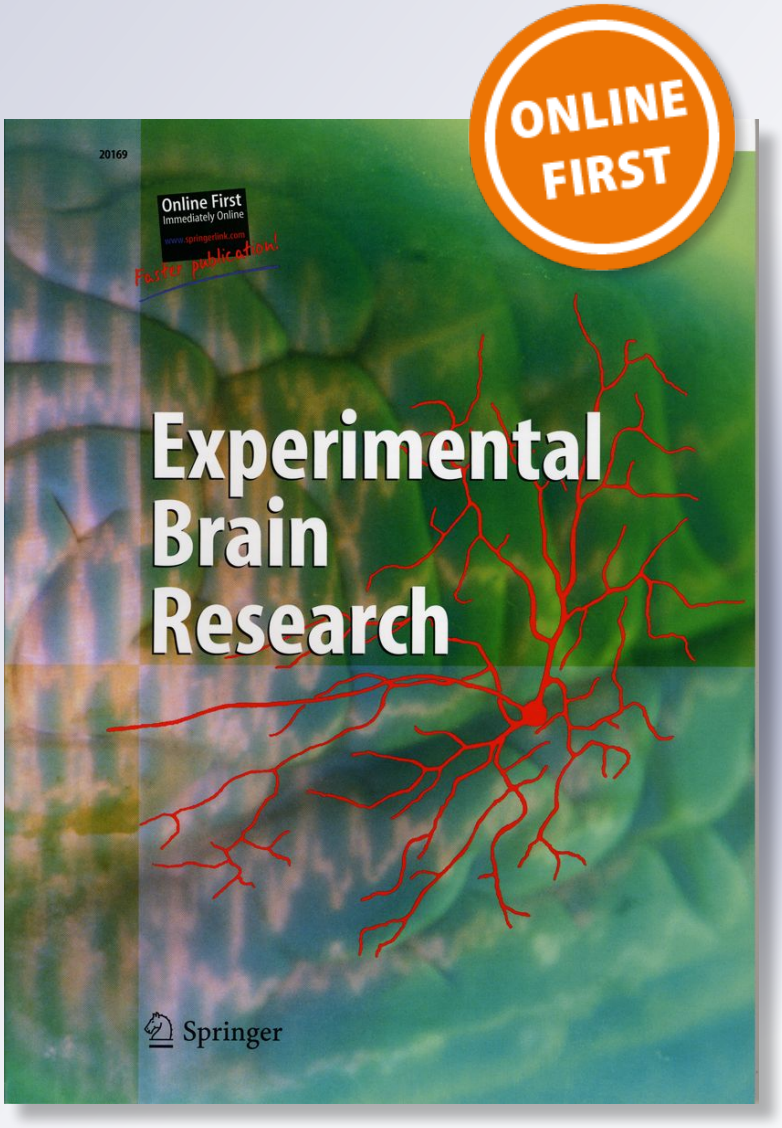
# *The impact of when, what and how predictions on auditory speech perception*

**Serge Pinto, Pascale Tremblay, Anahita Basirat & Marc Sato**

**Experimental Brain Research**

ISSN 0014-4819

Exp Brain Res
DOI 10.1007/s00221-019-05661-5




Springer

Experimental Brain Research
https://doi.org/10.1007/s00221-019-05661-5

RESEARCH ARTICLE

# The impact of when, what and how predictions on auditory speech perception

**Serge Pinto[1] · Pascale Tremblay[2,3] · Anahita Basirat[4] · Marc Sato[1]**

Received: 24 June 2019 / Accepted: 24 September 2019


## Abstract

An impressive number of theoretical proposals and neurobiological studies argue that perceptual processing is not strictly feedforward but rather operates through an interplay between bottom-up sensory and top-down predictive mechanisms. The present EEG study aimed to further determine how prior knowledge on auditory syllables may impact speech perception. Prior knowledge was manipulated by presenting the participants with visual information indicative of the syllable onset (*when*), its phonetic content (*what*) and/or its articulatory features (*how*). While w*hen* and *what* predictions consisted of unnatural visual cues (i.e., a visual timeline and a visuo-orthographic cue), *how* prediction consisted of the visual movements of a speaker. During auditory speech perception, *when* and *what* predictions both attenuated the amplitude of N1/P2 auditory evoked potentials. Regarding *how* prediction, not only an amplitude decrease but also a latency facilitation of N1/P2 auditory evoked potentials were observed during audiovisual compared to unimodal speech perception. However, w*hen* and *what* predictability effects were then reduced or abolished, with only *what* prediction reducing P2 amplitude but increasing latency. Altogether, these results demonstrate the influence of *when*, *what* and *how* visually induced predictions at an early stage on cortical auditory speech processing. Crucially, they indicate a preponderant predictive role of the speaker's articulatory gestures during audiovisual speech perception, likely driven by attentional load and focus.

**Keywords** Auditory speech perception · Audiovisual speech perception · Predictive coding · Predictive timing · EEG

## Introduction

It is widely acknowledged that subjective perceptual experience does not solely derive from sensory processing but is also constrained by prior knowledge or expectations. Contrary to the proposal that sensory information and higher level knowledge are integrated at a late, post-sensory, decision stage (Fodor 1983; Norris et al. 2000), long-standing perceptual theories postulate that the brain continuously predicts forthcoming sensory events, infers their most likely causes, to reduce sensory uncertainty (von Helmholtz 1909; Neisser 1967; Gregory 1980). From this view, perceptual processing is not strictly feedforward but partly operates through an interplay between bottom-up sensory and top-down predictive mechanisms. In its contemporary version based upon hierarchical Bayesian inference and probabilistic computations (Rao and Ballard 1999; Knill and Pouget 2004; Friston 2005, 2010; Clark 2013), the predictive coding theory proposes that, at each hierarchical cortical level, bottom-up sensory information is compared with top-down predictions from higher levels to estimate prediction errors. Over time, through perceptual learning, leverage prediction errors are thought to help reduce sensory uncertainty and to provide increasingly accurate recognition. Although some of the core tenets of predictive coding theory still remain debated (for a recent review, see Heilbron and Chait 2018), empirical evidence for the fundamental influence of expectations on neural responses and their anticipatory, predictive, nature has been accumulated over the past years, mostly using priming, adaptation and omission paradigms in which the predictability of the content and/or timing of a stimulus

✉ Marc Sato
marc.sato@lpl-aix.fr

[1] Laboratoire Parole et Langage, UMR 7309, CNRS, LPL, Aix Marseille Université, 5 avenue Pasteur, 13100 Aix-en-Provence, France

[2] Département de Réadaptation, Faculté de Médecine, Université Laval, Quebec City, Canada

[3] Cervo Brain Research Centre, Quebec City, Canada

[4] Univ. Lille, CNRS, CHU Lille, UMR 9193, SCALab, Sciences Cognitives et Sciences Affectives, Lille, France

Published online: 01 October 2019

is experimentally manipulated (e.g., oddball paradigms using electro- or magnetoencephalography, EEG/MEG, repetition suppression paradigms using fMRI).

Another line of evidence in favor of predictive mechanisms involves audiovisual perception, in which perceptual experience is facilitated by prior cross-modal associations and online integrative mechanisms. From a Bayesian perspective, perceptual experience here derives from the processing and integration of multisensory inputs based on their predictability and joint probability (Massaro 1998; van Wassenhove 2013; Rosenblum et al. 2016). For instance, previous EEG studies in which visual cues made the content/nature or the onset of an ongoing sound predictable showed that visual-induced predictions can modify activity in the auditory cortex as early as 100 ms post-stimulus (Widmann et al. 2004; Laine et al. 2007; Vroomen and Stekelenburg 2010; Paris et al. 2016, 2017). For speech perception, following a seminal investigation by Klucharev et al. (2003), neurophysiological studies have consistently showed that adding visual articulatory movements to auditory speech modulates activity early in the supratemporal auditory cortex, with an attenuated amplitude and earlier latency of auditory evoked potentials (AEPs, N1/P2 or M100) during audiovisual compared to unimodal speech perception (Klucharev et al. 2003; Besle et al. 2004; van Wassenhove et al. 2005; Stekelenburg and Vroomen 2007; Arnal et al. 2009; Pilling 2009; Winneke and Phillips 2011; Frtusova et al. 2013; Baart et al. 2014; Ganesh et al. 2014; Treille et al. 2014a, b, 2017, 2018; for a recent review and discussion, see Baart 2016). N1/P2 auditory evoked responses are thought to reflect synchronous neural activation in the thalamic-cortical segment of the central nervous system, with their sources mainly originating from the supratemporal plane of the auditory cortex, in response to spectral and temporal cues of an auditory stimulation (e.g., Näätänen and Picton 1987and Woods 1995). Given the temporal precedence of visual articulatory movements on the auditory speech signal in these studies (by tens to hundreds of milliseconds, Chandrasekaran et al. 2009; but see also Schwartz and Savariaux 2014), the earlier latency (timing of auditory neural processing) and amplitude suppression (size of neural population and activation synchrony) of AEPS are thought to reflect an increased temporal and/or phonetic predictability of the auditory speech stimulus through visual predictions (van Wassenhove et al. 2005; Stekelenburg and Vroomen 2007; Arnal et al. 2009, 2011; Vroomen and Stekelenburg 2010; Baart et al. 2014).

Although these results are well explained within the framework of visual-to-auditory predictive mechanisms (for reviews, see Arnal and Giraud 2012, van Wassenhove 2013, and Talsma 2015), one debated issue is whether the amplitude and latency modulation of AEPs reflect non-speech-specific temporal and/or phonetic visual predictions on the auditory speech signal. Stekelenburg and Vroomen (2007), Vroomen and Stekelenburg (2010), and Baart et al. (2014) argue that temporal, non-speech specific, visual predictions are reflected in N1 latency facilitation and amplitude reduction. In agreement with this hypothesis, they observed an amplitude and a latency reduction of auditory-evoked N1 responses during audiovisual perception for natural speech and non-speech actions (Stekelenburg and Vroomen 2007), as well as for artificial audiovisual stimuli (Vroomen and Stekelenburg 2010). In addition, using sine-wave speech that was perceived as speech by half of participants, they also provided evidence for a P2 amplitude reduction specifically dependent on the phonetic predictability of the visual speech input (Baart et al. 2014). In contrast, van Wassenhove et al. (2005) observed a visually induced suppression of both N1 and P2 auditory components independently of the visual phonetic saliency of the speech stimuli, but a latency reduction of N1 and P2 peaks depending on the degree of their visual phonetic predictability (i.e., the higher visual recognition of the syllable, the larger is the latency facilitation; see also Arnal et al. 2009; but Treille et al. 2014b for inconclusive findings). Based on their results, the authors argued for two distinct integration stages: a global bimodal perceptual stage and a featural phonetic stage. According to the authors, the global bimodal perceptual stage would be reflected in the amplitude reduction, independent of the featural content of the visual stimulus and possibly reflecting phase-coupling of auditory and visual cortices. The featural phonetic stage, in contrast, would be reflected in the latency facilitation, in which articulator-specific and predictive visual information are taken into account during auditory phonetic processing (for further discussion, see van Wassenhove 2013).

The goal of the present EEG study was to further determine how prior knowledge on auditory syllables is processed in the brain during auditory speech perception. To this aim, we investigated the extent to which experimentally induced visual predictability of an incoming auditory speech stimulus might modulate N1/P2 AEPs. Prior knowledge on auditory syllables was manipulated by presenting participants with visual information (see Fig. 1) indicative of the timing (*when*), the phonetic content (*what*) and/or the articulatory features (*how*). While *when* and *what* predictions consisted on unnatural visual cues (i.e., a visual timeline indicative of the syllable onset or a visuo-orthographic cue indicative of the syllable content), *how* prediction was operationalized by manipulating the presentation modality [auditory (A), visual (V) and audiovisual (AV)] to determine whether adding natural speech movements of a speaker modulates auditory speech perception.

First, the influence and possible interaction of *when* and *what* predictions on cortical auditory speech processing were determined by comparing the latency and amplitude of N1/P2 AEPs between these conditions in the auditory modality.

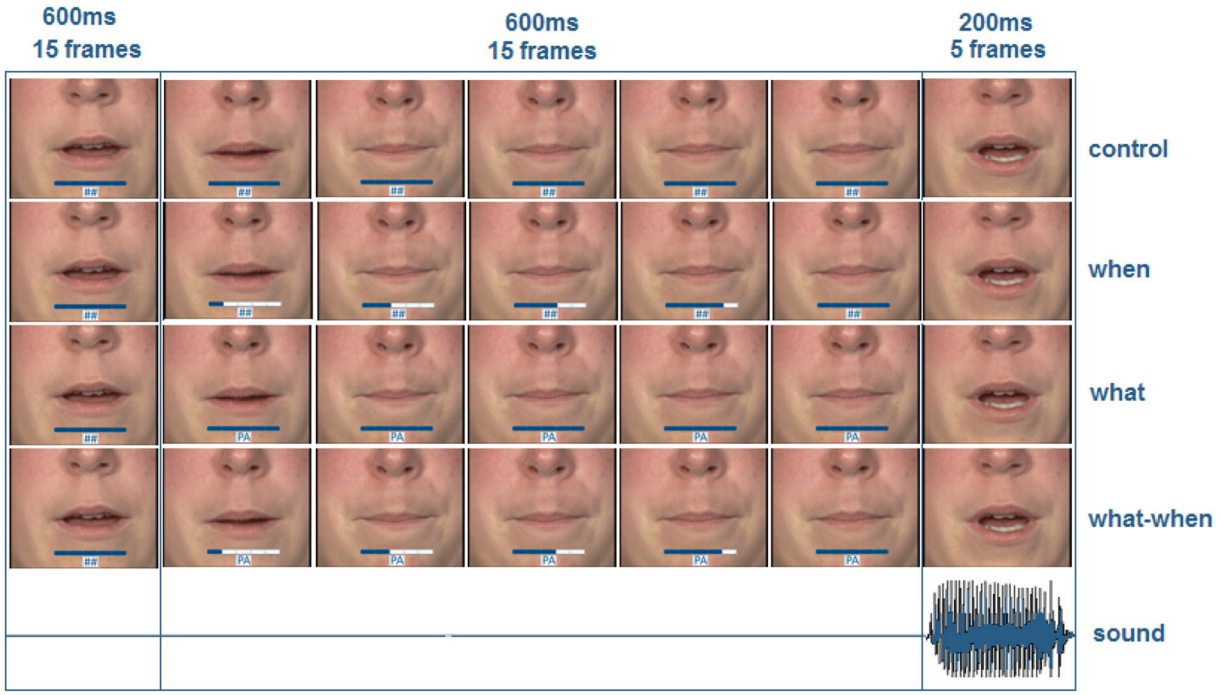



**Fig. 1** Examples of the four prediction conditions (control, *when*, *what*, *what–when*) for the /pa/ syllable in the audiovisual modality. The visual modality included the same prediction conditions and visual speech movements but did not include any sound. The auditory modality included the same prediction conditions with the acoustic syllable dubbed on a static image of a neutral mid-open mouth position of the speaker

Previous EEG studies provided evidence for predictive timing (Schafer et al. 1981; Clementz et al. 2002; Lange 2009; see also Vroomen and Stekelenburg 2010, Paris et al. 2016) or coding (Widmann et al. 2004 and Laine et al. 2007; see also Sohoglu et al. 2012 and Paris et al. 2017) during auditory speech perception but they did not examine whether these two predictions might interact. Second, the impact of *how* prediction on cortical auditory speech processing was determined by estimating whether adding speech movements to the acoustic signal might modulate the latency and amplitude of N1/P2 AEPS. To do so, we used an additive model to compare N1/P2 AEPs in the bimodal condition with the sum of those observed in the unimodal conditions (i.e., $AV \neq A + V$; for a recent review, see Baart 2016).[1] Based on the above-mentioned studies, we hypothesized a reduced amplitude and a shorter latency of N1/P2 auditory evoked responses in the bimodal compared to the sum of unimodal EEG signals. Finally, additional influence of *when* and *what* predictions during audiovisual speech perception was assessed.

## Methods

### Participants

Eighteen healthy adults (14 females and 4 males), with a mean age of 28 years (± 7 SD), ranging from 18 to 42 years, participated in the study after giving informed consent. All participants were native French speakers, with a mean of 15 years (± 3 SD) of education, ranging from 11 to 20 years. They were all right handed according to the standard handedness inventory (Oldfield 1971) with a mean score of 84% (± 14 SD), had normal or corrected-to-normal vision, and reported no history of hearing, speaking, language, neurological and/or neuropsychological disorders. The cognitive functioning of all participants was evaluated using the Montreal Cognitive Assessment scale (MoCA; Nasreddine et al. 2005). The mean score was 28.9/30 (± 1.4 SD). Hearing thresholds were assessed using a screening audiometer (Resonance R17A, MRS, Italy), at three frequencies (0.5, 1

[1] The use of an additive model in EEG studies is usually required to test audiovisual speech integration, defined by differences between the summed unimodal auditory and visual activity and activity generated by the audiovisual stimuli. Although the above-mentioned studies support the view that audiovisual speech integration partly operates through visually based predictions from the speaker's articulatory movements, it is important to note that audiovisual speech integration does not solely rely on visually based articulatory predictions. To avoid a conceptual confusion between audiovisual integration and visually based articulatory predictions, we will only refer to audiovisual speech perception throughout the manuscript.

and 2 kHz). A pure tone average was computed, with mean hearing thresholds of 5.6 (±4.2 SD) and 4.8 (±5.3 SD) dB HL for the left and right ear, respectively. The protocol was carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki and participants were compensated for the time spent in the study. The dataset from one female participant was removed from the study because of technical issues during EEG acquisition.

## Stimuli

Multiple utterances of /pa/, /ta/, and /ka/ syllables were individually recorded by three French speakers in a soundproof room. These syllables were selected to ensure a gradient of visuo-labial saliency (with the bilabial /p/ consonant known to be more visually salient than alveolar /t/ and velar /k/ consonants) in the visual and audiovisual modalities. Video digitizing was done at 25 frames per second with a resolution of 720×576 pixels. Audio digitizing was done at 44.1 kHz with 16-bit quantization recording. On the basis of visual and acoustical signals (using VirtualDub, VirtualDub.org, and Praat software, Boersma and Weenink 2013), one set of clearly articulated /pa/, /ta/, and /ka/ tokens were selected per speaker, providing a total of nine syllables.

One hundred and eight movies were then created consisting of the nine distinct /pa/, /ta/, and /ka/ syllables, each presented in three modalities associated to the above-mentioned *how* prediction [auditory (A), visual (V), audiovisual (AV)] and under four conditions manipulating the *what* and *when* predictions (control, *when*, *what*, *what–when*; see below). Each movie was 35-frame long (1400 ms). For all stimuli, the auditory signal intensity was normalized using a common maximal amplitude criterion. The audiovisual stimuli started with an initial neutral mid-open mouth position followed by visual speech movements (30 frames, 1200 ms) before the acoustic consonantal burst and the syllable (5 frames, 200 ms). The auditory stimuli consisted on the acoustic syllable dubbed on a static image of a neutral mid-open mouth position of the speaker. The visual stimuli consisted of the visual speech movements displayed without any sound.

The experiment included 12 experimental conditions related to the 3 modalities of presentation (A, V, AV) and the 4 prediction conditions: control (A, V, AV), when ($A_{when}$, $V_{when}$, $AV_{when}$), what ($A_{what}$, $V_{what}$, $AV_{what}$) and what–when ($A_{what\text{-}when}$, $V_{what\text{-}when}$, $AV_{what\text{-}when}$). In all conditions, the "##"orthographic symbols and a static timeline were visually presented during the first 15 frames (0–600 ms; see Fig. 1). In the *when* conditions, a moving visual timeline indicative of the temporal consonantal onset of the acoustic syllable replaced the static timeline during the subsequent 15 (±2) frames (600 (±80) ms to 1200 ms). In the *what* conditions, a visuo-orthographic cue indicative of the syllable (/pa/, /ta/ or /ka/) replaced the "##" symbols during the subsequent 15 (±2) frames (600 (±80) ms to 1200 ms). In the *what–when* conditions, both the visual timeline and visuo-orthographic cues were presented. Finally, in the control conditions, the speech signals were presented without any predictions regarding the acoustic syllable but with the "##"orthographic symbols and the static timeline during the next 15 (±2) frames (600 (±80) ms to 1200 ms). Note that in all conditions, the last five frames related to the syllable were presented without any visuo-orthographic symbols or cues, nor with any static or moving visual timeline. Importantly, the audiovisual stimuli were first created. The visual and auditory-only stimuli were based on these stimuli, by removing the acoustic signal or by replacing the visual speech movements by a static face. With this method, the acoustic signals were the same across the auditory-only and audiovisual stimuli, as well as the variable duration of the *when* and *what* predictive cues across all three modalities.

## Procedure

The EEG experiment was carried out in a sound-attenuated room. Participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented through loudspeakers at a comfortable sound level, with the same sound level set for all participants. The Presentation software (Neurobehavioral Systems, Albany, USA) controlled stimulus presentation and recorded participants' responses.

During the EEG recording, participants were instructed to pay attention to all visual cues (i.e., the articulatory movements of the speaker, the written orthographic syllable to be produced and/or the timeline of the temporal consonantal onset). Importantly, they were told that visual cues were always coherent with the syllable (no incongruency). They were asked to identify (forced-choice identification task) the syllable presented by pressing a key on a keyboard with their left hand (three response-key alternatives for /pa/, /ta/ and /ka/). To dissociate sensory/perceptual responses from motor responses on EEG data, a single brief auditory tone as well as the '?' symbol were delivered 600 ms after the end of each stimulus. The participants were asked to respond only after the presentation of these cues.

The experiment consisted of 864 trials presented in a pseudo-randomized order (avoiding the same condition in 2 consecutive trials), with 72 trials in each condition (3 speech modalities [A, V, AV]×4 predictions [control, *when*, *what*, *what–when*]×3 speakers×3 syllables (/pa/, /ta/, /ka/)×8 trials). The inter-trial interval was set at 3 s and the response key designation was fully counterbalanced. The experiment lasted approximately 45 min and was divided in four experimental sessions of equal duration. Short breaks were offered between sessions.

### EEG setup

EEG data were recorded continuously from nine 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 512-Hz sampling rate. Since N1/P2 AEPs have maximal response over central sites on the scalp (Scherg and Von Cramon 1986; Näätänen and Picton 1987), EEG were only collected from fronto-central electrodes (F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2). This is in line with previous EEG studies on audiovisual speech perception and auditory evoked responses (e.g., Pilling 2009, Stekelenburg and Vroomen 2007, Treille et al. 2014a, b, 2017, 2018, van Wassenhove et al. 2005, and Vroomen and Stekelenburg 2010). Two additional electrodes served as ground electrodes (Common Mode Sense (CMS) active and Driven Right Leg (DRL) passive electrodes). These two electrodes form a feedback loop driving the average potential of the subject in the Biosemi system (see https://www.biosemi.com). In addition, one external reference electrode was set at the top of the nose. Horizontal (HEOG) and vertical (VEOG) eye movements were recorded using 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, USA). EEG data were first off-line re-referenced to the nose recording and band-pass filtered using a two-way least-square FIR filtering (3–30 Hz) to reduce slow drift and high-frequency noise. Though the high-pass filtering is less typical than a standard 0.1 Hz or 1 Hz, it was most appropriate for our dataset. First, we used a more conventional 1-Hz high-pass filter. However, for a few subjects, we observed a constant slow drift of the EEG signal. Although the cause of the drift was uncertain (cardiac artifact, perspiration artifact, etc.), this slow modulation of the EEG signal was present in all experimental conditions. Since a 3-Hz high-pass filter fully removed the drift in the EEG signal of these subjects in all experimental conditions, we, therefore, decided to use it instead of the more widely used high-pass cutoff of 0.1 Hz/1 Hz. Data were then segmented into 500-ms epochs including a 100-ms prestimulus baseline (from −100 to −0 ms relative to the acoustic syllable onset). Epochs with an amplitude change exceeding ±100 µV at any channel (including HEOG and VEOG channels) were rejected (mean (±SD): 2% (±2%) trials).

Responses from /pa/, /ta/ and /ka/ syllables were first averaged together to provide 72 trials per condition. For each participant and each condition (i.e., A, V, AV, $A_{when}$, $V_{when}$, $AV_{when}$, $A_{what}$, $V_{what}$, $AV_{what}$, $A_{what\text{-}when}$, $V_{what\text{-}when}$, $AV_{what\text{-}when}$), data were then averaged over the nine electrodes. Finally, the maximal amplitude and peak latency of N1 and P2 AEPs were determined on the EEG waveform using a fixed temporal window (N1: 70–150 ms; P2: 150–250 ms).

### Analyses

#### Accuracy

The percentage of correct responses was determined for each participant and condition. We conducted a three-way repeated measures ANOVA with modality (A, V, AV), *when* (yes, no) and *what* (yes, no) predictions as within-participant factors.

#### EEG: auditory speech perception

To test the influence of *when* and *what* predictions on auditory EEG responses, we first conducted two-way repeated measures ANOVAs, separately on N1 and P2 amplitudes and latencies in the auditory modality with *when* (yes, no) and *what* (yes, no) predictions as within-participant factors.

#### EEG: audiovisual speech perception

To assess the impact of *how* prediction on auditory EEG responses as well as additional influence of *when* and *what* predictions during audiovisual speech perception. To do so, we used an additive model in which the bimodal audiovisual EEG signal was compared to the sum of auditory and visual unimodal EEG signals ($AV \neq A + V$, $AV_{what} \neq A + V_{what}$, $AV_{when} \neq A + V_{when}$, $AV_{what\text{-}when} \neq A + V_{what\text{-}when}$). We conducted three-way repeated measure ANOVAs on both N1 and P2 amplitudes and latencies with signal type (bimodal vs. sum), *when* (yes, no) and *what* (yes, no) predictions as within-participant factors.

In all analyses, the alpha level was set at $p = 0.05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, post hoc analyses were conducted with Bonferroni corrections.

## Results

### Accuracy (see Fig. 2)

The mean proportion of correct responses was 86%. As was expected, a strong effect of modality was observed ($F(2,32) = 110.8$, $p < 0.00001$, $\eta^2 = 0.87$) with the percentage of correct responses in V modality (87%) lower than in A (98%) and AV (98%). The main effect of *what* was also significant ($F(1,16) = 91.5$, $p < 0.00001$, $\eta^2 = 0.85$), with the percentage of correct responses higher when

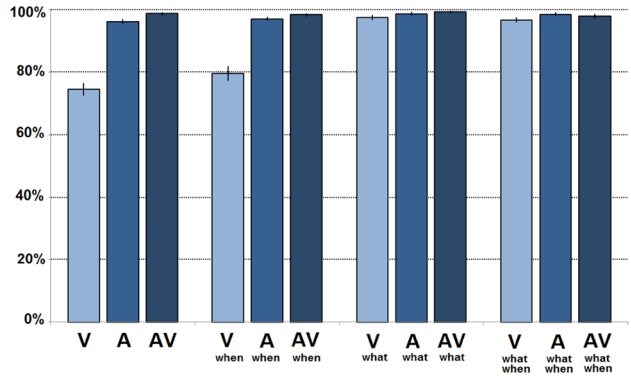



**Fig. 2** Mean percentage of correct responses observed in the auditory, visual and audiovisual modalities in relation to *when* and *what* predictions. Error bars represent the standard error of the mean

the prediction was present (98%) than when it was not (91%). Finally, all interactions were significant (modality × *when*: $F(2,32) = 7.5$, $p < 0.006$, $\eta^2 = 0.32$; modality × *what*: $F(2,32) = 92.2$, $p < 0.00001$, $\eta^2 = 0.85$; *when* × *what*: $F(1,16) = 29.8$, $p < 0.00006$, $\eta^2 = 0.65$; modality × *when* × *what*: $F(2,32) = 6.1$, $p < 0.006$, $\eta^2 = 0.28$). The modality × *when* × *what* interaction revealed a significantly lower percentage of correct responses for V (75%) compared to $V_{when}$ (80%), and for $V_{when}$ compared to all other stimuli (all above 96%).

## EEG: auditory speech perception (see Fig. 3a, b)

### Amplitude

The mean N1 and P2 amplitudes were −5.4 μV and 5.3 μV, respectively. For N1, there was a main effect of *when* ($F(1,16) = 8.1$, $p = 0.01$, $\eta^2 = 0.34$), with a lower negative amplitude when the prediction was present compared to when it was absent (on average, −5.1 μV vs. −5.6 μV). Similarly, the main effect of *what* was also significant ($F(1,16) = 6.2$, $p = 0.02$, $\eta^2 = 0.28$), with a lower negative amplitude when the prediction was present (on average, −5.0 μV vs. −5.8 μV). Finally, a significant *when* × *what* interaction was observed ($F(1,16) = 9.7$, $p < 0.007$, $\eta^2 = 0.38$). While a significantly reduced negative amplitude was observed for *what* regardless of *when*, the latter was found to lower the amplitude only in the absence of *what* (on average, no prediction: −6.2 μV, *when*: −5.4 μV, *what*: −5.1 μV, *what–when*: −4.9 μV). For P2, there was a main effect of *when* ($F(1,16) = 17.3$, $p < 0.0008$, $\eta^2 = 0.52$), with a lower positive amplitude when the prediction was present compared to when it was absent (on average, 4.9 μV vs. 5.7 μV). The main effect of *what* was also significant ($F(1,16) = 8.3$, $p = 0.01$, $\eta^2 = 0.34$) with a lower positive amplitude when the prediction was present (on average, 5.0 μV vs. 5.6 μV). The *what* × *when* interaction was not reliable ($F(1,16) = 3.0$).

### Latency

The mean N1 and P2 latencies were 118 ms and 197 ms, respectively. No main effect or interaction reached significance for both N1 and P2 latencies.

## EEG: audiovisual speech perception (see Fig. 4a, b)

### Amplitude

For AV and A + V EEG signals, the mean N1 and P2 amplitudes were −5.5 μV and 5.5 μV, respectively. For N1, there was a main effect of signal type ($F(1,16) = 13.3$, $p < 0.003$, $\eta^2 = 0.45$) with a higher negative amplitude for A + V compared to AV signals (on average, −6.0 μV vs. −5.0 μV). No other main effect or interaction was found. For P2, a main effect of signal type was observed $F(1,16) = 35.1$, $p < 0.001$, $\eta^2 = 0.69$) with a higher positive amplitude for A + V compared to AV signals (on average, 6.5 μV vs. 4.5 μV). The main effect of *what* was also significant ($F(1,16) = 6.6$, $p = 0.02$, $\eta^2 = 0.29$) with a lower positive amplitude when the prediction was present compared to when it was absent (on average, 5.4 μV vs. 5.7 μV). Moreover, a significant signal type × *what* interaction was observed ($F(1,16) = 12.9$, $p < 0.003$, $\eta^2 = 0.45$) with a stronger amplitude reduction between AV and A + V signals with the prediction than without (on average, −2.3 μV vs. −1.7 μV). No other main effect or interaction was found to be significant.

### Latency

For AV and A + V EEG signals, the mean N1 and P2 latencies were 118 ms and 198 ms, respectively. For N1, there was a main effect of signal type ($F(1,16) = 9.0$, $p < 0.009$, $\eta^2 = 0.36$) with a longer latency for A + V compared to AV signals (on average, 121 ms vs. 114 ms). No other main effect or interaction was found. For P2, there was a main effect of signal type ($F(1,16) = 5.0$, $p = 0.04$, $\eta^2 = 0.24$) with a longer latency for A + V compared to AV signals (on average, 201 ms vs. 194 ms). In addition, the main effect of *what* was also reliable ($F(1,16) = 4.7$, $p = 0.05$, $\eta^2 = 0.23$) with a shorter latency for A + V compared to AV signals (on average, 196 ms vs. 199 ms). No other main effect or interaction was found.

# Discussion

The aim of the present EEG study was to determine the influence of visual predictions on auditory speech processing. Prior knowledge on auditory syllables was

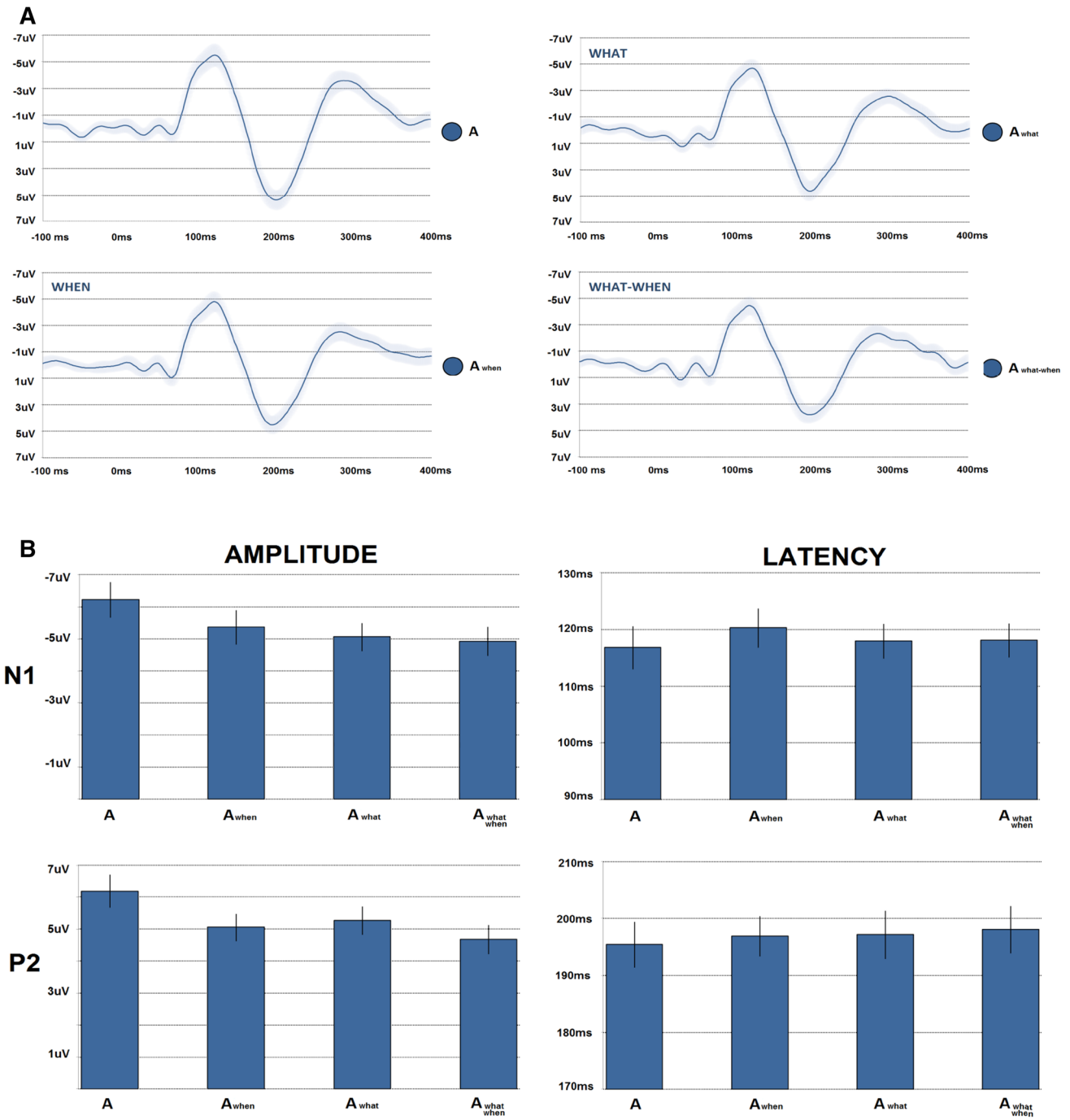



**Fig. 3** **a** Averaged event-related potentials on fronto-central electrodes in the auditory modality in relation to *when* and *what* predictions. **b** Mean N1 and P2 amplitudes and latencies in the auditory modality in relation to *when* and *what* predictions. Error bars represent the standard error of the mean

manipulated by presenting the participants with visual information indicative of the syllable onset (*when*), its phonetic content (*what*) and/or its articulatory features (*how*). There are three main findings. First, during auditory speech perception, *when* and *what* predictions attenuated the amplitude of N1/P2 AEPs. Second, regarding *how* prediction, an amplitude decrease and a latency facilitation of N1/P2 AEPs were observed when comparing bimodal audiovisual to unimodal auditory and visual conditions. Finally, only *what* but not *when* prediction was found to reduce P2 amplitude and to increase latency during audiovisual speech perception.

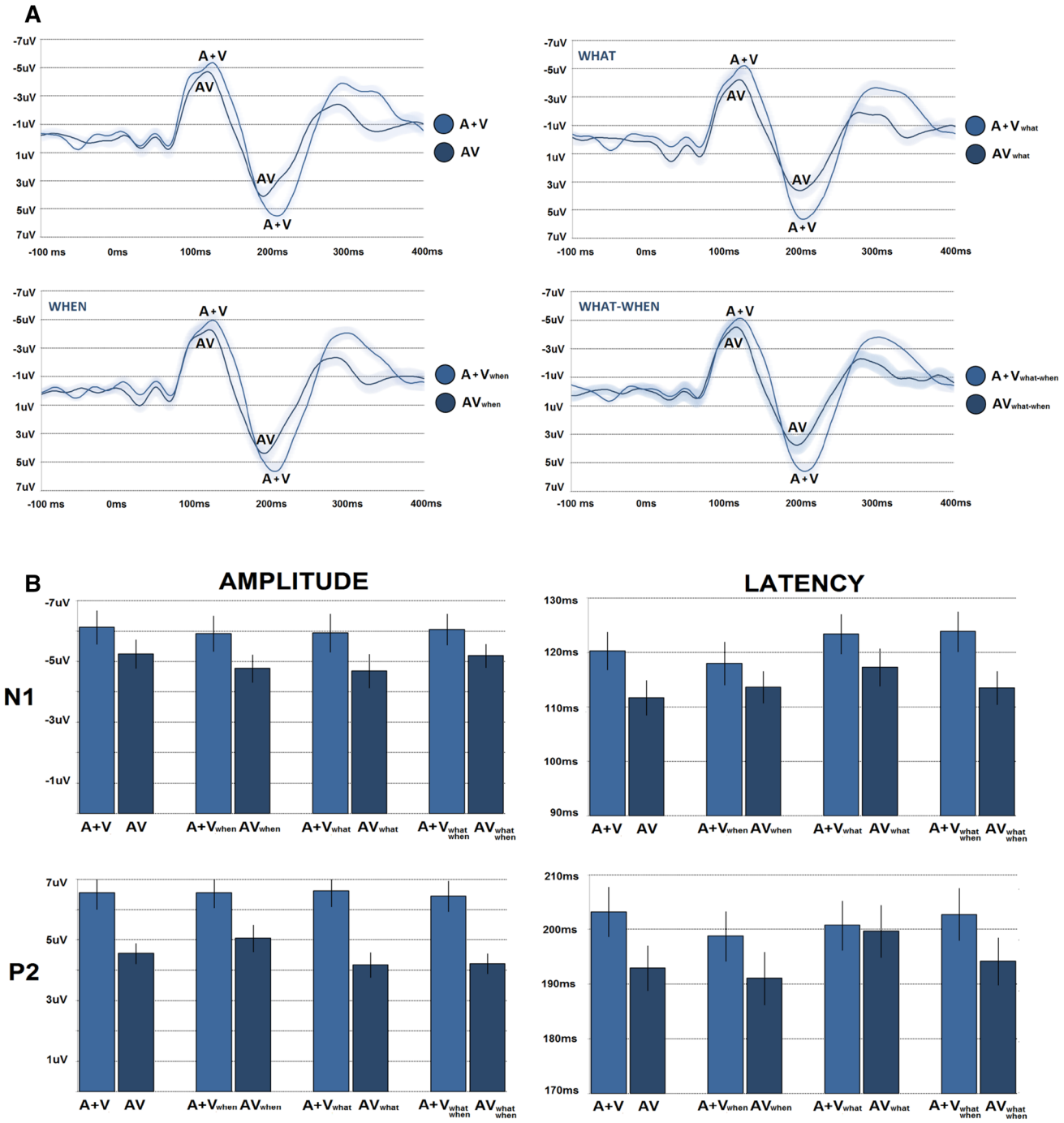



**Fig. 4** **a** Averaged event-related potentials on fronto-central electrodes related to AV and A + V EEG signals in relation to *when* and *what* predictions. **b** Mean N1 and P2 amplitudes and latencies related to AV and A + V EEG signals in relation to *when* and *what* predictions. Error bars represent the standard error of the mean

## Accuracy

A near-perfect classification of syllables was observed for auditory and audiovisual speech stimuli, independently of *what* and *when* predictions. As was expected, lip-reading in the visual modality was associated with the lowest syllable recognition rate, though performance was well above chance level (75%). Adding the *when* prediction modestly but significantly increased the percentage of correct responses from 75% to 80%, while adding the *what* prediction led to almost perfect accuracy. Overall, these results confirm that participants benefited from watching predictive visual cues.

More interestingly, the significant *when* predictability effect on syllable recognition demonstrated that adding temporal predictive cue helped disambiguating phonetic content from the visual speech stimuli. Although the underlying mechanism remains unclear, one possibility could be that the visual timeline focused the participants' attention on the lip forms and kinematics, enhancing the visual categorization between syllables.

### Auditory speech perception

In the auditory modality, *when* and *what* predictions modulated the amplitude of N1 and P2 AEPs, despite similar recognition scores across conditions. These results are in line with previous EEG studies that showed significant effects of prior temporal knowledge on auditory stimuli (Schafer et al. 1981; Clementz et al. 2002; Lange 2009; Vroomen and Stekelenburg 2010; Paris et al. 2016). For example, Lange (2009) presented participants with a sequence of tones, with expectations induced by varying the temporal regularity of the sequence. Results showed an attenuation of N1 AEPs by temporal expectations. Relatedly, Vroomen and Stekelenburg (2010) showed a reduced auditory N1 amplitude during tone perception when auditory onsets were made temporally predictable by two critically timed moving disks. Importantly, in all these studies as in the present one, amplitude reduction was observed for temporal predictable compared to unpredictable auditory stimuli. By contrast, an enhancement of N1 amplitude was observed in EEG studies investigating temporal auditory attention, mostly by manipulating the time interval between two stimuli (Lange et al. 2003, 2006; Lange and Röder 2006; Röder et al. 2007; Lange 2013). For example, Lange et al. (2003) presented participants with a sequence of two sounds separated by a shorter or longer temporal interval. Participants were asked to respond to the second sound only after a specified interval, marking the attended time point. Stimuli presented at attended compared to unattended moments in time elicited an amplitude enhancement of the auditory N1. These results are usually regarded as evidence for a gating or filter mechanism of attention, by which the processing of attended stimuli is favored over that of unattended ones (for a review, Lange 2013). These opposite effects on AEPs measured in studies investigating temporal attention vs. expectation suggest that the visual temporal cues used in the present study, which increased stimulus onset predictability, may have counteracted a possible enhancing effect of attention on N1/P2 AEPs, leading to a decline in N1 amplitude. This interpretation is also relevant for the observed effect of the *what* prediction on AEPs. Indeed, a number of studies in which visual cues made the content/nature of an ongoing sound predictable also revealed an amplitude reduction of N1 AEPs (Widmann et al. 2004; Laine et al. 2007; Paris et al. 2017; see also Sohoglu et al. 2012 for an enhanced amplitude on left frontal electrodes but using spoken words).

These studies, therefore, argue for a predictive rather than an attentional influence of *when* and *what* visual predictions on the auditory processing of syllables in the present study. Interestingly, while a reduced amplitude of AEPs was observed for the *what* prediction, independent of the presence or absence of the *when* prediction, the *when* prediction reduced the amplitude of AEPs only when the *what* prediction was absent. This suggests that when presented with both predictions (*what–when* condition), phonetic compared to temporal knowledge of the incoming syllable had a preponderant predictive role on auditory speech processing, possibly due to higher visual attentional load and specific focus (see below). Another important element to discuss is that, as in previous studies, these effects appeared within 100 ms after stimulus onset. This supports the view that predictive timing and coding mechanisms occur at an early sensory stage of cortical auditory speech processing (Talsma 2015), which argues against the hypothesis that sensory information and prior knowledge are integrated at a late decision stage (Fodor 1983; Norris et al. 2000).

### Audiovisual speech perception

We observed a reduced amplitude and a shorter latency of both N1 and P2 AEPs when comparing the audiovisual EEG signal with the sum of the auditory and visual EEG signals (i.e., using an additive model; for a recent review, see Baart 2016). This result is in line with previous EEG studies on audiovisual speech perception (Klucharev et al. 2003; Besle et al. 2004; van Wassenhove et al. 2005; Stekelenburg and Vroomen 2007; Arnal et al. 2009; Pilling 2009; Vroomen and Stekelenburg 2010; Winneke and Phillips 2011; Frtusova et al. 2013; Baart et al. 2014; Ganesh et al. 2014; Treille et al. 2014a, b, 2017, 2018) and demonstrates that visual movements of the speaker affect ongoing cortical auditory activity.

Importantly, compared to the auditory modality, adding speech movements reduced or even abolished *when* and *what* predictability effects. Only the *what* prediction was still found to reduce P2 amplitude but to also increase latency. These results suggest a preponderant predictive role of *how* prediction during audiovisual speech perception as well as a competing effect between *how* and *what* predictability effects on the auditory P2. Adding the speaker's visual movements to the timeline and/or phonetic cues undoubtedly increased visual attentional load and, as a result, decreased the amount of available processing resources for each visually induced predictions. In addition, although participants were asked to pay attention to all visual cues, their display sizes were

very different, the visual movements being largely predominant (see Fig. 1). Since attention resources were here overloaded, participants might have strategically focused on visual speech movements, which would have reduced or abolished the influence of *what* and *when* predictions. This hypothesis does not mean that participants fully ignored these visual cues when the speaker movements were displayed. Recognition scores indeed confirmed that participants benefited from watching temporal and phonetic predictive cues in the visual modality (a condition that was indistinguishable from the audiovisual modality until the speech sound was presented). One additional explanation would be that of a ceiling predictive effect of visual speech movements that already includes, to some degree, temporal and phonetic information on auditory syllables (e.g., van Wassenhove et al. 2005, Stekelenburg and Vroomen 2007, and Baart et al. 2014). This would also partly explain the absence or reduced *what* and *when* predictability effects during audiovisual perception. However, a longer latency of the auditory P2 was also observed between AV and A + V signals when the *what* prediction was concomitantly presented with the visual movements. A final, more cautious, interpretation is, therefore, that of a combination of predictive but also attentional influences on auditory speech processing. Indeed, since attention and expectation are known to have an opposite effect on N1/P2 AEPs (for a review, Lange 2013), a greater attentional load due to multiple visual cues and a visual preference for the speaker movements may have partly blurred the contribution of *what* and *when* predictive cues. Consistent with this interpretation, previous behavioral and EEG studies have demonstrated that audiovisual speech perception is indeed modulated by attentional load (Alsius et al. 2005, 2014). In a single- vs. dual-task paradigm, these studies showed a lower McGurk effect (i.e., a visually driven alteration in the auditory speech percept; McGurk and MacDonald 1976), a poorer lip-reading and a slower latency of N1/P2 AEPs when participants were concurrently performing an unrelated visual task compared to when attention was fully focused on speech stimuli in a single task. The increased latency of the auditory P2 observed in the present study when the *what* prediction was concomitantly presented with the visual movements appears in line with these results. Interestingly, we did not observe such detrimental influence of the visual timeline on AEPs during audiovisual speech perception. Moreover, we observed a beneficial influence on performance of *when* and *what* visual cues when added to the speaker movements in the visual modality. Although a full explanation here appears elusive, this suggests that high attentional load can influence audiovisual speech perception in different ways, at multiple sensory and decision stages, depending on the predictive value of the visual cues, its naturalness and the attentional locus.

## Conclusion

Altogether, our results demonstrate the impact of *when*, *what* and *how* visually induced predictions at an early sensory stage on cortical auditory speech processing. Importantly, they indicate a preponderant role of *how* prediction during audiovisual speech perception and suggest a competing effect between *how* and *what* predictability effects. Finally, the present findings strongly suggest an interaction of predictive and attentional influences on auditory speech processing due to multiple visual cues.

**Acknowledgements** The authors thank Avril Treille and Coriandre Vilain for their help with the stimuli. We also thank all the participants.

### Compliance with ethical standards

**Conflict of interest** The authors declare no competing financial interests.

## References

Alsius A, Navarra J, Campbell R, Soto-Faraco S (2005) Audiovisual integration of speech falters under high attention demands. Curr Biol 15:839–843

Alsius A, Möttönen R, Sams ME, Soto-Faraco S, Tiippana K (2014) Effect of attentional load on audiovisual speech perception: evidence from ERPs. Front Psychol 5:727

Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. Trends Cogn Sci 16(7):390–398

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29(43):13445–13453

Baart M (2016) Quantifying lip-read induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. Psychophysiology 53(9):1295–1306

Baart M, Stekelenburg JJ, Vroomen J (2014) Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia 65:115–211

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20:2225–2234

Boersma P, Weenink D (2013) Praat: doing phonetics by computer. Computer program, Version 5.3.42. http://www.praat.org/. Accessed Sept 2019

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar A (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5:e1000436

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–204

Clementz BA, Barber SK, Dzau JR (2002) Knowledge of stimulus repetition affects the magnitude and spatial distribution of low-frequency event-related brain potentials. Audiol Neurootol 7:303–314

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. J Neurosci Methods 134:9–21

Fodor J (1983) The modularity of mind. Massachusetts Institute of Technology, Cambridge

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138

Frtusova JB, Winneke AH, Phillips NA (2013) ERP evidence that auditory–visual speech facilitates working memory in younger and older adults. Psychol Aging 28(2):481–494

Ganesh AC, Berthommier F, Vilain C, Sato M, Schwartz JL (2014) A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. Front Psychol 5:1340

Gregory RL (1980) Perceptions as hypotheses. Philos Trans R Soc Lond B Biol Sci 290:181–197

Heilbron M, Chait M (2018) Great expectations: is there evidence for predictive coding in auditory cortex? Neuroscience 389:54–73

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

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719

Laine M, Kwon MS, Hämäläinen H (2007) Automatic auditory change detection in humans is influenced by visual-auditory associative learning. NeuroReport 18(16):1697–1701

Lange K (2009) Brain correlates of early auditory processing are attenuated by expectations for time and pitch. Brain Cogn 69:127–137

Lange K (2013) The ups and downs of temporal orienting: a review of auditory temporal orienting studies and a model associating the heterogeneous findings on the auditory N1 with opposite effects of attention and prediction. Front Integr Neurosci 7:263

Lange K, Röder B (2006) Orienting attention to points in time improves stimulus processing both within and across modalities. J Cogn Neurosci 18:715–729

Lange K, Rösler F, Röder B (2003) Early processing stages are modulated when auditory stimuli are presented at an attended moment in time: an event-related potential study. Psychophysiology 40:806–817

Lange K, Krämer UM, Röder B (2006) Attending points in time and space. Exp Brain Res 173:130–140

Massaro DW (1998) Perceiving talking faces. MIT Press, Cambridge

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 265:746–748

Näätänen R, Picton TW (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

Nasreddine ZS, Phillips NA, Bedirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment (MoCA): a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53:695–699

Neisser U (1967) Cognitive psychology. Appleton-Century-Crofts, New York

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23:299–370

Oldfield RC (1971) The Assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–113

Paris T, Kim J, Davis C (2016) The processing of attended and predicted sounds in time. J Cogn Neurosci 28(1):158–165

Paris T, Kim J, Davis C (2017) Visual form predictions facilitate auditory processing at the N1. Neuroscience 343:157–164

Pilling M (2009) Auditory event-related potentials (ERPs) in audiovisual speech perception. J Speech Lang Hear Res 52(4):1073–1081

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87

Röder B, Krämer UM, Lange K (2007) Congenitally blind humans use different stimulus selection strategies in hearing: an ERP study of spatial and temporal attention. Restor Neurol Neurosci 25:311–322

Rosenblum LD, Dorsi J, Dias JW (2016) The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. Ecol Psychol 28(4):262–294

Schafer EWP, Amochaev A, Russell MJ (1981) Knowledge of stimulus timing attenuates human evoked cortical potentials. Electroencephalogr Clin Neurophysiol 52:9–17

Scherg M, Von Cramon D (1986) Evoked dipole source potentials of the human auditory cortex. Electroencephalogr Clin Neurol 65:344–360

Schwartz JL, Savariaux C (2014) No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Comput Biol 10(7):e1003743

Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. J Neurosci 32:8443–8453

Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19:1964–1973

Talsma D (2015) Predictive coding and multisensory integration: an attentional account of the multisensory mind. Front Integr Neurosci 19:9

Treille A, Cordeboeuf C, Vilain C, Sato M (2014a) Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. Neuropsychologia 57:71–77

Treille A, Vilain C, Sato M (2014b) The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. Front Psychol 5(420):1–9

Treille A, Vilain C, Kandel S, Sato M (2017) Electrophysiological evidence for a self processing advantage during audiovisual speech integration. Exp Brain Res 235(9):2867–2876

Treille A, Vilain C, Schwartz JL, Hueber T, Sato M (2018) Electrophysiological evidence for audio-visuo-lingual speech integration. Neuropsychologia 109:126–133

van Wassenhove V (2013) Speech through ears and eyes: interfacing the senses with the supramodal brain. Front Psychol 4:1–17

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 102:1181–1186

von Helmholtz H (1909) In treatise on physiological optics, vol III, 3rd edn. Voss, Leipzig

Vroomen J, Stekelenburg JJ (2010) Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. J Cogn Neurosci 22:1583–1596

Widmann A, Kujala T, Tervaniemi M, Kujala A, Schröger E (2004) From symbols to sounds: visual symbolic information activates sound representations. Psychophysiology 41(5):709–715

Winneke AH, Phillips NA (2011) Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. Psychol Aging 26(2):427–438

Woods D (1995) The component structure of the N1 wave of the human auditory evoked potential. Electroencephalogr Clin Neurophysiol Suppl 44:102–109