Brain Topogr (2015) 28:494–505
DOI 10.1007/s10548-013-0338-2

ORIGINAL PAPER

Brain Prediction of Auditory Emphasis by Facial Expressions During Audiovisual Continuous Speech

Kuzma Strelnikov · Jessica Foxton · Mathieu Marx · Pascal Barone

Received: 19 July 2013 / Accepted: 21 November 2013 / Published online: 24 December 2013

Abstract The visual cues involved in auditory speech processing are not restricted to information from lip movements but also include head or chin gestures and facial expressions such as eyebrow movements. The fact that visual gestures precede the auditory signal implicates that visual information may influence the auditory activity. As visual stimuli are very close in time to the auditory information for audiovisual syllables, the cortical response to them usually overlaps with that for the auditory stimulation; the neural dynamics underlying the visual facilitation for continuous speech therefore remain unclear. In this study, we used a three-word phrase to study continuous speech processing. We presented video clips with even (*without emphasis*) phrases as the frequent stimuli and with one word visually emphasized by the speaker as the non-frequent stimuli. Negativity in the resulting ERPs was detected after the start of the emphasizing articulatory movements but before the auditory stimulus, a finding that was confirmed by the statistical comparisons of the audiovisual and visual stimulation. No such negativity was present in the control visual-only condition. The propagation of this negativity was observed between the visual and fronto-temporal electrodes. Thus, in continuous speech, the visual modality evokes predictive coding for the auditory speech, which is analysed by the cerebral cortex in the context of the phrase even before the arrival of the corresponding auditory signal.

Keywords Audio-visual speech · Prosody · Mismatch · Predictive coding

Introduction

Ecological speech has a multisensory nature and is mostly continuous, involving whole words and phrases; phonemes and syllables are encountered not separately (as presented in many speech perception studies) but embedded in the flow of continuous speech. Speech perception combines for typical listeners both auditory and visual features in face-to-face communication beginning from early childhood. Thus, audiovisual speech is the ecological modality of speech from early childhood. The information provided by the auditory and visual channels is relatively redundant and speech processing can often be supported solely by the auditory system. However, in degraded auditory conditions, congruent visual information from lip movements added to the auditory speech signal significantly increases the accuracy of speech comprehension (Sumby and Pollack 1954), a phenomenon which is equivalent to an increase in the signal to noise ratio (Ross et al. 2007). An important

Kuzma Strelnikov and Jessica Foxton are joint first authors
This is one of several papers published together in Brain Topography on the ''Special Issue: Auditory Cortex 2012''

Electronic supplementary material The online version of this article (doi:10.1007/s10548-013-0338-2) contains supplementary material, which is available to authorized users.

K. Strelnikov (✉) · J. Foxton · M. Marx · P. Barone
Université Paul Sabatier, CerCo, Université Toulouse, Toulouse, France
e-mail: kuzma@cerco.ups-tlse.fr

K. Strelnikov · J. Foxton · M. Marx · P. Barone
Faculté de Médecine de Purpan, CNRS, CERCO, UMR 5549, Pavillon Baudot CHU Purpan, BP 25202, 31052 Toulouse, France

M. Marx
Service d'Oto-Rhino-Laryngologie, Hopital Purpan, Toulouse, France

characteristic of audio-visual natural speech is that the visual cues originating from the mouth opening precede the auditory information by 100–200 ms and are strongly correlated with the auditory speech envelope (Chandrasekaran et al. 2009). This temporal dynamic implies that lip movements provide strong predictive cues for the auditory information. Such facilitatory audio-visual interactions can be observed at the behavioral level when analyzing speech intelligibility, but they are also present at the neuronal level, as expressed as a visual modulation of auditory speech processing in the human auditory cortex (Besle et al. 2004, 2008; Hertrich et al. 2007).

However, the visual cues involved in auditory speech processing are not restricted to information from the lip movements, but they also include head and chin gestures and facial expressions such as eyebrow movements (Munhall et al. 2004). These visual cues are especially important at the supra-segmental level of verbal or emotional communication, corresponding to the speech prosody. Locution prosody consists of modulations in the pitch, amplitude, and duration patterns of the words in a phrase, and enables listeners to distinguish between questions and statements, finished and unfinished phrases, or to recognize the emotional state of the speaker. Visual, non-verbal gestures contribute to speech prosody, and head and eyebrow movements have been shown to correlate with modulations in the pitch and amplitude of the talker's voice (Munhall et al. 2004; Hadar et al. 1984; Vatikiotis-Bateson and Yehia 2002). Again, based on the facilitatory rules that govern multisensory interactions, visual prosody-related information improves auditory speech comprehension in situations of degraded auditory information (Munhall et al. 2004; Barkhuysen et al. 2008). Recently, we have been able to show that prosodic visual cues can affect not only the linguistic aspects of speech but also the more fundamental levels of auditory processing. Indeed, we have demonstrated that the visual features of speech prosody can induce a crossmodal facilitation in detecting the auditory features of prosody, as expressed as a decrease in the threshold for detecting amplitude changes (Foxton et al. 2010).

This study suggests that visual cues improve sensitivity to loudness changes and consequently can improve speech-sound processing. A fundamental question then emerges as to the underlying neuronal mechanisms of such an audio-visual facilitation. The fact that these visual gestures precede the auditory signal by several milliseconds (Hadar et al. 1984) means that the visual information may play a predictive role, which could influence the neuronal auditory responses. While it is known that visual cues from lip-reading can modulate auditory cortical responses (see Campbell 2008 for review), no evidence for such an interaction exists for prosody-related visual information.

Thus, previous studies have addressed the audiovisual integration for separate phonemes and syllables but no study addressed this issue in continuous speech at the level of a phrase. In contrast to studies of separate phonemes and syllables, this study focuses on the intonation emphasis of a word in a phrase. The purpose of the present study was to investigate the temporal-spatial neural dynamics underlying this visually driven facilitation in continuous speech.

In order to examine the effect of the visual prosodic information on the brain auditory responses during continuous speech, we used an oddball paradigm in which the standard and deviant stimuli contained the same auditory information, but where the deviant stimuli had visual emphasis on one word. Our hypothesis was that the visual cues would give the impression that there is an increase in the amplitude/pitch of this word, and that this would be detected as such by the auditory system and expressed as a brain response similar to the auditory mismatch negativity (MMN) response (Naatanen et al. 2007). This approach is analogous to the one developed for the McGurk effect to study linguistic audio-visual interactions, where the visual information alone can change the perceived syllable (McGurk and MacDonald 1976). In the studies of the McGurk effect, a significant MMN has been found, demonstrating a visual influence on speech processing in the auditory system (Colin et al. 2002; Saint-Amour et al. 2007; Besle et al. 2005). In our prosodic visual-auditory protocol, the frequent stimuli were presented by the neutral three-word phrase where both visual and auditory components were congruent. In the non-frequent stimuli, the auditory and visual information were not congruent with respect to the second word in the phrase. Here, the auditory information had a neutral prosodic pattern, while the visual information emphasized the second word in the phrase. The visual-only emphasis created an illusion of the auditory emphasis reported by all the subjects. This paradigm enabled us to explore the visual influence on prosodic processing independently of any auditory modulation.

Materials and Methods

Participants

Ten normally-hearing native French speakers (mean age 24, range 19–29) participated in the study. Seven of the participants were female and all were right-handed. All of the subjects reported no auditory, neurological, or psychiatric disease, and all had normal or corrected to normal vision. All of the participants gave their full-informed consent prior to their participation in this study, in accordance with the Declaration of Helsinki (1968). The study was approved by the local research ethics committee

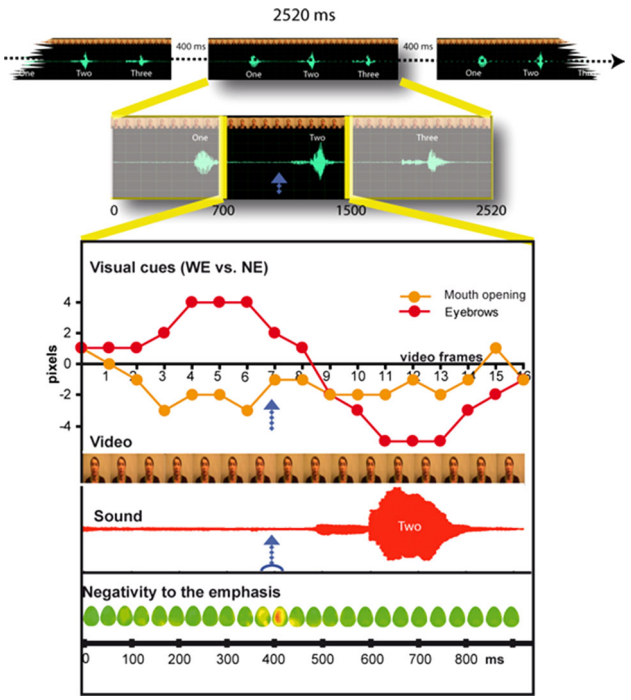

Fig. 1 Experimental protocol and analysis. The *upper part* of the figure shows the successive presentation of the video clip (duration 2,520 ms) in which an actress says the words "one-two-three" in French. **a** 400 ms silent interval separates each presentation of the video. Our EEG analysis is restricted to a period between 700 and 1,500 ms (*lower enlarged video clip*) that isolates the word "two" containing the visual emphasis. For this portion of the video clip, we quantified the differences (in pixels) between the facial features for the even and the emphasized "two" (*lower panel*). For the mouth opening, a distance was calculated for the mouth contour at the *middle vertical line*. For the eyebrows, a distance was calculated from the middle of the eyebrows to the *upper frame border*. The *graph* shows that the main feature of emphasis relates to the position of the eyebrows which are higher for the emphasized "two". The zero time point for the word "two" corresponds to 700 ms from the beginning of the video

(Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale Toulouse II Avis N°2-03-34/Avis N°2).

Stimuli

An actress was asked to say the words "un-deux-trois" (translated into English as "one-two-three") twice: once with *no emphasis* (NE), and again *with emphasis* (WE) on the word "two". She was asked to try to produce the words with the same timing. Several video recordings were taken using a professional Sony digital camera (recording with 25 frames per second, with a resolution of 720 × 576 pixels, and a sound sample rate of 48,000 Hz) in a sound-attenuating chamber. Two video clips were chosen, one with no emphasis and the other with emphasis. Both videos were 2.52 s in duration. Using Adobe Premier Pro 7.0, the sound in the deviant video clip was replaced with the sound from the 'no-emphasis' video clip. In this way, the two video clips had exactly the same sound (always from the 'no-emphasis' clip), and differed only in terms of the visual signal. Thus the WE video can be considered as an incongruent audio-visual stimulus with respect to the prosodic information (see Foxton et al. 2010). The visual features of emphasis included raised eyebrows during the word "two" that were not present in the 'no-emphasis' video clip (Fig. 1).

These visual gestures were clearly dissociated from the auditory stimulus as they preceded the sound "two" by a period of about 500 ms, making them a good predictor of the emphasised word in the intact with-emphasis video (before sound replacement). In the dubbed video clip, we maintained this same delay between the visual and auditory cues.

To verify the neutral prosody of the phrase, we tested a supplementary set of subjects ($n = 17$) asking what word was emphasized. The majority (80 %) of them did not hear any emphasis on the word "two" but indicated randomly other words when tested with the auditory-only recording ($p < 0.001$, χ^2-test). In addition, in the deviant audio-visual clip, the illusion of the auditory word "two" as being emphasized was reported by every one of the ten subjects who participated in the EEG study (they were asked after the study), as well as in every one of eight additional controls.

Procedure

For this study, we adopted an oddball design, whereby participants were presented with repetitions of the NE video clip interspersed with occasional presentations of the deviant WE video clip, with a constant inter-stimulus onset asynchrony of 2,920 ms. As the deviant stimulation was located in the middle of the phrase, it could not cause an expectation effect for the whole phrase.

For half of the runs, the video clips were presented with the sound (audio-visual condition-AV); for the other half, the video clips were presented without the sound (visual-alone condition-V). During the runs, participants were asked to detect rare white crosses that occasionally appeared over the eyes during the video clips. They had to respond by pressing a button. The crosses appeared after one of four possible delays with respect to the start of the video (1.0, 1.2, 1.4, or 1.6 s). This task ensured that the participants focused their attention on the videos, and especially around the time of the word "two". By limiting the white crosses to the eyes, we focused the participants' attention on a facial area where visual features of emphasis are prominent, given the presence of raised eyebrows and widened eyes during the emphasised word.

There were six AV runs and six V runs, each lasting approximately 7 min. The order of the AV and V runs was randomised for each participant. In each run, there were 100 standard videos, 25 deviant videos, and a further 15 standard videos that contained a target white cross. 17.9 % of the videos were deviants and 10.7 % of the videos were targets, and there were a total of 150 deviants for each condition. Within each run, the order of the videos was pseudo-randomised, with the constraint that the first ten videos were always standards and that there were at least three standard videos between the deviants.

EEG Recordings and Analyses

EEG recordings were obtained using an elastic cap (Oxford Instruments, UK) fitted with tin electrodes and with a reference electrode placed on the tip of the nose. The cap electrode locations were in accordance with the 10–20 system with additional electrodes from the 10–10 system, and were at locations: FP1-FP2-F3-F4-C3-C4-P3-P4-F7-F8-T3-T4-T5-T6-CB2-CZ-Fz-Pz-T5′-T6′-O1-O2-O1′-O2′-P3′-P4′-Pz'-Oz-CB1-M1-M2-VEOG (32 electrodes). Additional electrodes were placed on the right and left mastoids, and the vertical electro-oculogram (VEOG) was recorded between electrodes placed above and below the eyes. The ground electrode was placed along the midline in front of Fz, and electrode impedances were kept below 10 kΩ. EEG and EOG data were recorded using a SynAmps amplifier (NeuroScan, El Paso, TX), with a sampling rate of 1,000 Hz, and low-pass filtered at 200 Hz.

A standardised EEG analysis procedure was followed using Neuroscan software. The data was band-pass filtered at 2–30 Hz (slope 48 dB/octave, FIR) and epoched from 700 to 1,500 ms with respect to the beginning of the video clip. In the text and figures, this corresponds to the "zero" of the epoch used for the analysis. This timing was chosen to target the analysis towards the visual and auditory parts of the word of interest: "two" (Fig 1). We chose this onset on the basis of the timing of the visual gestures, and the "zero" corresponds closely to the minimum difference between the NE and WE videos for both the eyebrow movements and mouth opening. The epochs were baseline-corrected for a time period of 100 ms before the epoch. This baseline corresponded to the interval between the auditory "one" and "two". Epochs containing excessive residual artefacts were excluded (artefact rejection outside −40:40 μV), as were epochs for videos immediately following the deviants or targets to exclude the possible contamination by the mismatch negativity to the standard stimulus following the deviant one (Sams et al. 1984). After the rejection, for each condition we had the average of 102 deviant and 408 standard trials per subject. The recording sites plus an electrode placed on the right mastoid were referenced online to the left mastoid electrode and digitally re-referenced offline to the algebraic average of the left and right mastoids (Naatanen et al. 2007). Averages were calculated separately for the standard and the deviant videos.

Statistical analyses were carried out in order to determine whether there were significant differences between the waveforms for the standard and deviant videos. These focused on the time period during the word "two" where the visual differences related to emphasis were present. Average waveforms were calculated for each participant, and then waveforms for standards and deviants were compared with a paired *t* test for the electrode of interest, Fz. To sharpen surface topographies, we conducted the current source density (CSD) analysis separately for standard and deviant stimuli and calculated the differential map using the CSD toolbox (http://psychophysiology.cpmc.columbia.edu/software/CSDtoolbox/index.html), which computes scalp surface Laplacian or current source density (CSD) estimates for surface potentials.

The basic methods of statistical estimation were a 2×2 repeated measures ANOVA including the modality (audio-visual, visual) and type of stimulation (standard, deviant) factors, and a bootstrap test with bias-corrected and accelerated confidence intervals (Carpenter and Bithell 2000). The family-wise error rate was controlled when necessary using the cluster correction in the temporal and spatial domains (Maris and Oostenveld 2007).

Results

Behavioural Results

The participants obtained high scores on the detection task (a white cross), which ranged from 88 to 100 % correct detection (median 99 %) for the audio-visual condition, and from 91 to 100 % detection (median 99 %) for the visual-alone condition. The number of false alarms ranged from 0 to 2 for both conditions. These results demonstrate that the participants' attention was focused on the videos during the recordings. Because there was a variable delay in the presentation of the visual target, we were confident that participants were attentive to the visually emphasised word 'two'.

EEG Results

To search for a MMN response, we performed a contrast analysis by first comparing the EEG signal for the standard (NE) and deviant (WE) audio-visual conditions at of interest, as is normally done for MMN analyses because the Fz electrode provides a summary response from the

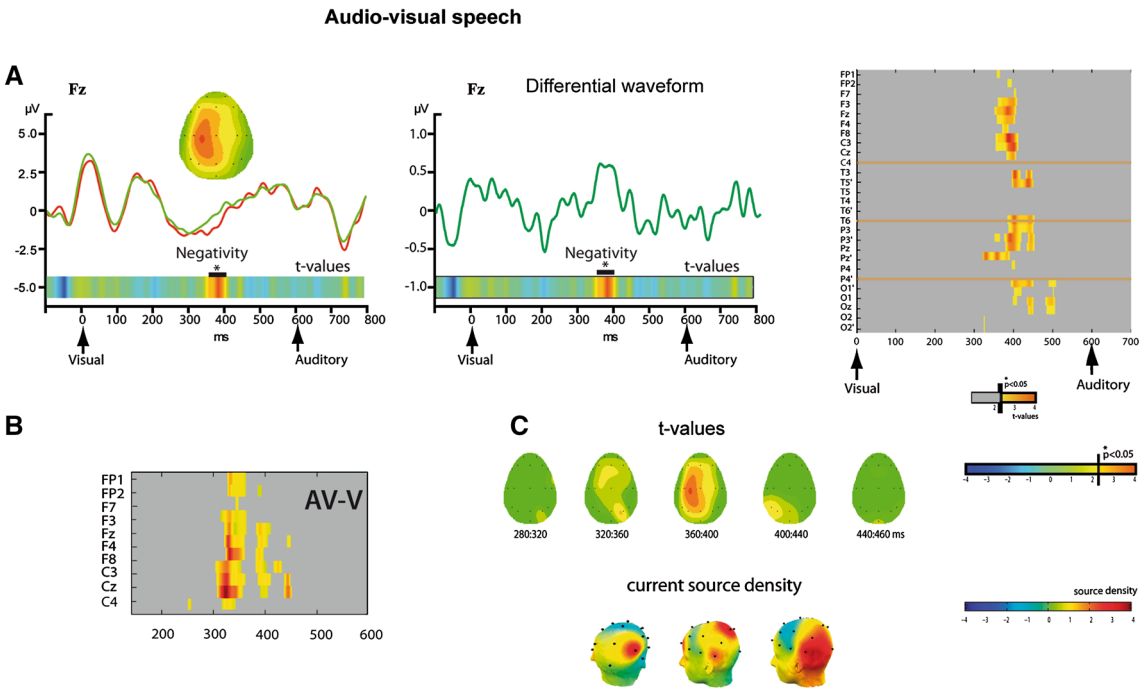

Fig. 2 MMN-like component for the visually-evoked auditory predictive coding. **a** Event-related potentials for the word "two" presented either emphasized in the WE condition (*red waveform*) or non-emphasized in the NE condition (*green waveform*). The critical t-value corresponding to $p = 0.05$ is 2.3 (df = 9). The spatial distribution of the t-values at peak is presented. The zero time point for the word "two" corresponds to 700 ms from the beginning of the video. The *black bar* indicates the t-values, which survived the cluster correction in the temporal domain. The *right panel* displays a statistical cluster plot of the negativities observed in the ERPs. *Color values* indicate the time intervals where the difference "deviant versus standard" is significant at $p < 0.05$ in the point-wise paired t-test. Electrode positions are arranged from frontal to posterior regions. Only statistically significant differences are depicted. In panel **b**, the significance of the results from the "AV-V" contrast is presented per subject in the fronto-central electrodes compared between standards and deviants. As in the previous analysis (see panel **a**), we found a robust negativity that precedes the apparition of the auditory stimulus. **c** Spatial-temporal pattern of the MMN-like component. The negativity during audio-visual stimulation originates from the occipito-parietal electrodes, propagates to the fronto-temporal electrodes, and then returns back to the occipital electrodes. This loop occurred after the visual deviance but before the auditory stimulation. In the CSD analysis (time range 320–440 ms), the propagation involves the left frontal and the left posterior temporal sites. The *colour scale* of t-values is the same for all parts of the figure

auditory cortices (Naatanen et al. 2007). The SNR at Fz was 1.7, which permits to distinguish ERPs even in the continuous EEG without averaging (Quian Quiroga and Garcia 2003). The SNR reflects the relation of the variances in the poststimulus and prestimulus intervals; for the whole set of electrodes it was 2.0 ± 0.6 (SD).

Given the absence of EEG studies using oddball paradigm for continuous audiovisual speech, we had no a priory idea in what time window after visual deviance the negativity is located. As we had no a priori hypothesis, we performed a temporal cluster correction of the t-values obtained from paired t-tests of the whole waveform; the correction controlled for the family-wise error rate in the temporal domain. Using this procedure, significant negativity was detected at 360–400 ms relative to the beginning of the epoch, which corresponds to 1,060–1,100 ms with respect to the onset of the video ($p_{corr} < 0.05$ (10,000 permutations), see Fig. 1). In this time window, we detected that the cluster with negativity covers a set of electrodes (Fz, Cz, Pz, F3, F4, C3, C4, P3, P3′) (Fig. 2), the significance of negativity in these electrodes was confirmed using the cluster correction in the spatial domain ($p_{corr} < 0.05$ (10,000 permutations).

We computed the size of the effect using the Cohen's test (1992). This method, applied at the temporal window of 360–400 ms at the electrode Fz, resulted in d values of 1.46, which classified the size of the observed effect as large.

To check for the possible late cognitive components of the response, we analyzed the data during 600 ms after the onset of the auditory word "two". Neither at the Fz electrode nor at other electrodes we did not observe a significant negativity during 600 ms after the onset of the auditory word "two," even at the uncorrected level of significance.

Concerning the between-subject stability of the effect at the Fz electrode, the negativity at 360–400 ms was found for all but one of the subjects. However, it was present for

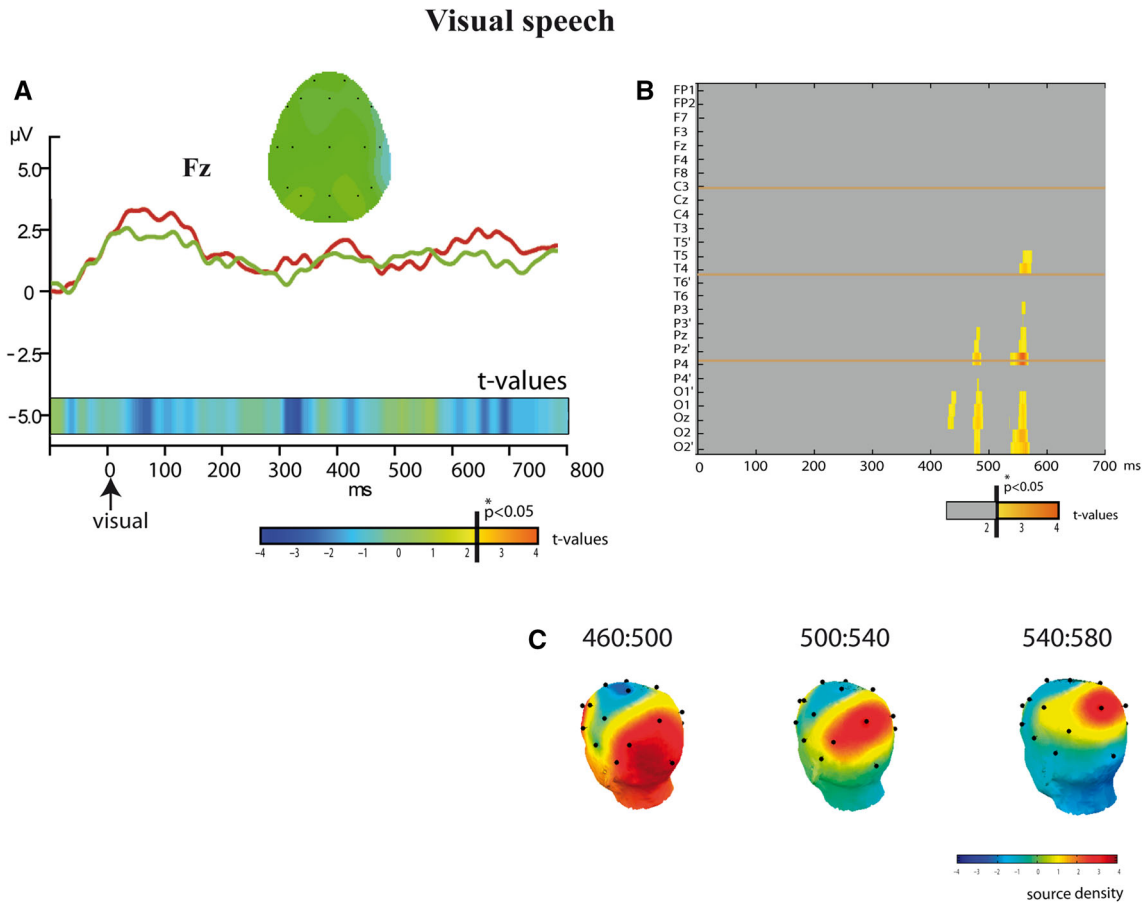

Fig. 3 MMN-like component during the visual-only conditions. At the Fz electrode, we did not detect a MMN-like component, as observed during the audio-visual condition (see Fig. 2). However, the statistical cluster plot reveals a late negativity, restricted to the occipital electrodes corresponding to the visual cortex. Conventions including the colour code as in Fig. 2. **a** Event-related potentials for the visual word "two" presented either emphasized in the WE condition (*red waveform*) or non-emphasized in the NE condition (*green waveform*). **b** Statistical cluster plot of the negativities observed in the ERPs. *Colour values* indicate the time intervals where the difference "deviant versus standard" is significant at $p < 0.05$ in the point-wise paired t-test. **c** CSD analysis indicates sources in the occipito-parietal areas

each subject at the Cz electrode, a location that is usually found to record an auditory MMN, including previous audio-visual speech MMN studies (Kislyuk et al. 2008).

As the detected negativity occurred at about 200 ms before the beginning of the auditory "two," it is probably related to the prosodic visual gestures. In an attempt to relate this MMN-like response to the visual cues, we performed a frame-by-frame analysis of the differences (in pixels, see Fig. 1) between the NE standard and WE deviant video clips. This analysis showed differences in the size of the mouth opening (mean 1.5 ± 0.8 pixels), a feature known to be related to the spectral structure of voice sounds (see Chandrasekaran et al. 2009). But the largest variations concerned the eyebrow movements, which were found to present a mean amplitude difference between the NE and WE conditions of 2.7 (±1.5 pixels) with bimodal temporal dynamics. It is important to note that these eyebrow movements began about 500 ms before the related sound "two." As can be seen from the frame-by-frame timing in Fig. 1, the MMN-like response occurred immediately after the initiation of a large eyebrow lift in the emphasized video clip. This timing strongly suggests a causal link between the MMN-like response and the facial prosodic gestures.

However, this MMN-like activity could be a visual MMN, as has been demonstrated in some McGurk studies (see Saint-Amour et al. 2007). To explore this possibility, we analyzed the ERPs during the visual-only runs in which the visual deviant and standard were the same as in the AV conditions. The 2×2 repeated measures ANOVA at 360–400 ms revealed a significant interaction between the modality of stimulation and the type of stimulation (standard and deviant): $F(1,9) = 7.2$, $p < 0.05$, Power of the effect 0.7. The same effect was observed using 2×2

MANOVA on the group of electrodes from the earlier described cluster of electrodes with significant negativity (Fz, Cz, Pz, F3, F4, C3, C4, P3, P3′): $F(1,9) = 27.5$, $p < 0.001$.

No significant negativity was detected for the emphasized relative to the even speech in the visual modality at any time period for the above described electrodes even at the uncorrected level of significance (Fig. 3a).

A bootstrap analysis showed no significant effect for the visual-only condition ($p > 0.05$) at central, frontal, and temporal groups of electrodes for the 360–400 ms time window, where negativity had been observed in the audio-visual condition. However, at the occipital pole (electrode Oz), a slight difference was observed at the uncorrected level ($p > 0.05$) between the deviant and standard ERP signals at a later stage (440 ms, Fig. 3b). The CSD analysis revealed the occipito-parietal distribution of the current source density for this late negativity (Fig. 3c).

As a supplementary part of the analysis, to further delineate the audiovisual nature of the observed response, we subtracted the ERP signal of the visual condition from the audiovisual one, prior to contrasting the standard and deviant conditions for each subject (AV-V contrast, see Saint-Amour et al. 2007). A bootstrap test was used to compare the differences between the standard and deviant waveforms in a window that covered the 360–400 ms time window of interest (see Fig 2b). This analysis revealed the presence of negativity at the Fz electrode ($p < 0.05$), thus confirming that the MMN-like response was specific to the audio-visual condition.

The spatio-temporal pattern of the MMN-like response in the audio-visual condition is shown in the the right panel of Fig. 2a. This statistical cluster plot (shown at the uncorrected level of $p < 0.05$) reveals a clearly organized pattern of statistically significant negativity across the electrodes, which was confirmed by a cluster correction in the temporal and spatial domains. It can be seen that the MMN-like response was more pronounced at the central and frontal electrode sites and had no significant left–right asymmetries ($p > 0.05$, bootstrap test). No gender effect on the negativity was detected ($p > 0.5$).

The spatial dynamics of the negativity propagation are presented in Fig. 2c. The earliest part (320–360 ms) may suggest that the MMN-like response originates in the posterior parieto-occipital visual cortex ($p < 0.05$, cluster correction in the temporal and spatial domains with 10,000 permutations). Later, between 360–400 ms, corresponding to the peak of the MMN-like response, it propagates forwards to the auditory temporo-frontal electrodes, before ending in a backward propagation towards the occipital electrodes at 400–440 ms ($p < 0.05$, cluster correction in the temporal and spatial domains with 10,000 permutations). In the CSD analysis, the current source density in the occipital regions confirm the early occipital negativities, but for the main negativity effect at 360–400 ms, the large central negativity corresponds to the current source density in the left parietal region, in the left posterior temporal region, and in the inferior frontal region (Fig. 2c). Though the tendency to the left side lateralization was not significant in the classical analysis, it became more pronounced in the CSD analysis. This spatio-temporal analysis confirms the multisensory visuo-auditory nature of the negativity, which was only observed in the audiovisual condition.

To obtain a three-dimensional image of the underlying sources in the brain, we performed the distributed source reconstruction where current sources are estimated at a large number of fixed points (8,196 in our case) within a cortical mesh, rather than approximated by a small number of equivalent dipoles. In this approach, multiple sparse priors are assessed using the "greedy search" method (Friston et al. 2008) implemented in SPM8. For the observed negativity, this method permitted to detect sources in the left posterior temporal and frontal regions (supplementary Fig. 1) at 360–400 ms, which roughly correspond to the locations of the CSD analysis. In spite of the complex nature of the summary effects, which are observed on the scalp, if one considers Fig. 2 there is also a tendency for the left-side asymmetry in the scalp distribution of the negativity. However, no spatio-temporal dynamics of this negativity could be detected in the 3D space, possibly due to the small amplitudes of the potentials at the early and late time periods used in the CSD analysis. This could be explained by the lower temporal sensitivity of the 3D source reconstruction in the suboptimal for the 3D source reconstruction 32-channel EEG system. The CSD analysis is performed in the 2D space and reflects more directly the real data. However, the spatial similarity of the CSD analysis and distributed source analysis strengthens the conclusions on the role of the left posterior temporal and frontal brain activities in the observed negativity. Thus, we consider the 3D source reconstruction as a supplementary result confirming the spatial distribution of the negativity in the 2D space by the CSD analysis.

Discussion

Speech processing represents what is probably the most striking example of multisensory interactions, in which complementary information from lip movements and sound signals are merged into coherent percepts for phonetic perception. Our present results provide new evidence that facial movements play an important role in continuous speech processing and can be used to predict certain

features of the auditory utterance. Using an oddball audio-visual paradigm with the same sound across standards and deviants, we detected a significant negativity after the start of the emphasizing visual movements but, surprisingly, before the auditory stimulus. This negativity was confirmed by the statistical comparisons of the audio-visual and visual stimulations. No such negativity was present in the control visual-only condition, in spite of the same emphasizing movements. Thus, the observed brain response is specific to the auditory expectation in a audio-visual condition; in the multisensory context, the visual modality evokes facilitating predictive coding for the auditory speech. The behaviourally observed illusion is based on the visual prediction, which is not rejected by the brain during the auditory stimulation even though the auditory part is not emphasized. These results provide the first evidence at the neuronal level of predictive relationships between prosody and related facial movements, such as eyebrow movements and speech acoustic signals.

In continuous speech the preceding context, purely auditory or visuo-auditory, influences the perception of the forthcoming information by creating predictions based on the context and on the previous experience (Strelnikov 2008). This is a principal difference from the perception of the separate syllables and phonemes, which are not embedded in the linguistic context. However, because in continuous speech, the audio-visual asynchrony derived from lip movements could be reduced compared to that observed in isolated words (Schwartz and Savariaux 2013), one might consider that the visual information derived from prosody-related facial movements become more important. Thus, we believe that the observed context-dependent effect is an important phenomenon for continuous speech.

A MMN-like Response Induced by Face Movements

The MMN in oddball protocols corresponds to the automatic brain responses evoked by changes in repeated auditory stimuli. This negativity has been observed in response to changes across a large number of different acoustic features, such as intensity, frequency, and duration (review in Naatanen et al. 2007). Several MMN studies have used incongruent situations such as the McGurk effect to study audio-visual interactions at the level of the auditory cortex. MMN-like responses have been reported for natural ecological stimuli (such as a hammer hitting a nail Ullsperger et al. 2006), and for emotional face/voice interactions (de Gelder et al. 1999), but they are especially prominent for visuo-auditory speech sounds where the deviance is limited to the visual stimulus (Colin et al. 2002; Saint-Amour et al. 2007; Ponton et al. 2009; Sams et al. 1991). In this latter case, it has been shown that the McGurk effect elicits a significant MMN-like waveform at Fz, despite the auditory stimulus in the deviants being the same as in the standards. Source analysis has located the generators of this response to the auditory cortex (Saint-Amour et al. 2007; Mottonen et al. 2002), confirming the modulating influence of visual information at the early stages of auditory processing, a concept that is now widely accepted for multisensory interactions (Cappe et al. 2010; Schroeder and Foxe 2005). Our present results support there being a visual influence on auditory speech processing, and they clarify the major features of these audio-visual interactions. In addition, whereas in all previous reports the MMN response occurred at around 150–250 ms after the auditory onset, the MMN-like response precedes the auditory stimulation in our study.

Concerning the role of attention, when crosses appeared over the eyes in the video, subjects could still distract their attention from the face though keeping gaze in the face area. In this case, the observed negativity to the deviant face expression could be interpreted as a pre-attentive processing making it closer to the pre-attentive nature of MMN (Naatanen et al. 2007).

MMN-like Response and the Auditory Prosodic Illusion

It is probable that the significant negativity we observed in the present study is involved in evoking the illusory percept of emphasis on the auditory word "two," as was reported by the subjects. Importantly, we did not observe a MMN-like negativity at the Fz or Cz electrodes when contrasting the standard and deviant conditions in the visual-only condition. Only an ancillary difference was observed at later stages of visual processing. Furthermore, when the visual signals were subtracted from the audio-visual ones, the MMN-like response remained robust. These results rule out the possibility that the MMN-like response observed in the audio-visual condition was induced by the visual mismatch alone. Rather, it shows that the MMN resulted from the visual features of emphasis being used to predict a change in the auditory utterance, such as in the pitch or intensity of the stressed word. Prosodic visual gestures have been previously shown to be closely related to sound modulations during speech production (Guaitella et al. 2009; Hadar et al. 1984; Munhall et al. 2004), and several studies have revealed a correlation between the acoustic and visual features of speech (Scarborough et al. 2009; Vatikiotis-Bateson and Yehia 2002; Barker and Berthommier 1999; Jiang et al. 2002); for example, between the fundamental frequency of the voice and the speaker's eyebrow movements (Cavé et al. 1996). In addition, the visual features of speech prosody have been shown to exert a cross-modal facilitation on auditory thresholds for intensity changes (Foxton et al. 2010).

In previous MMN studies using the McGurk effect, audio-visual reactions were not found to precede the auditory stimulus (Colin et al. 2002; Saint-Amour et al. 2007; Ponton et al. 2009; Sams et al. 1991) as in the present study. In these studies, an auditory MMN was reported at short latencies (~175 ms, see Colin et al. 2002; Saint-Amour et al. 2007; Ponton et al. 2009; Sams et al. 1991) after the acoustic phoneme, but never before. This absence of a preceding MMN signal may be explained by the fact that the visual stimuli, such as the mouth opening, were very close in time to the auditory information (Chandrasekaran et al. 2009) so that the brain reaction to them would overlap with the auditory stimulation. In our study, we have been able to show the brain response to the visual deviance, because the facial movements preceded the auditory information by more than 400 ms. The absence of an illusionary "auditory" MMN in our study could be explained by a limit in the sensitivity of the MMN approach. Our previous behavioural finding (Foxton et al. 2010) showed that visual cues increase the auditory threshold to prosody only by a few dB. In classical oddball protocols based on auditory intensity changes (Naatanen et al. 2007), MMN can be observed with intensity differences of more than 3 dB, corresponding to at least 10 % between the deviant and standard stimuli. The occipital negativity at about 480 ms in the visual-only condition may be related to the N400 component, which is known to be elicited in response to visual incongruence (Proverbio and Riva 2009), which is in our case an incongruence with respect to visual expectations.

Though the observed negativity in the audio-visual condition was before auditory stimulation, one could speculate that it may also reflect an analogy with the electrophysiological N2b response as part of the orienting complex (Halgren et al. 1995), overlapping with MMN-like activity. An interesting perspective for further studies of audio-visual continuous speech would be to distinguish between the pre-attentive MMN-like audio-visual mismatch and the orienting complex. It should be noted that the word "one", which preceded the word "two" could be the major key of the auditory context for the brain. Thus, the possibility of this negativity in the equi-probable conditions remains as a perspective for the other paradigms.

Our results provide strong evidence for the involvement of audio-visual networks during the audio-visual condition. First, the observed negativity was not restricted to the occipital sites as in the studies using the visual oddball paradigm (Campanella et al. 2002; e.g., Tales et al. 1999; see Kimura 2012 for the recent review), but rather had a fronto-central distribution as has been found in the auditory MMN studies (Naatanen et al. 2007). Secondly, the CSD analysis confirms the implication of the posterior temporal regions, which are known to be involved in audio-visual integration during speech processing. This cortical region is involved, for example, in semantic decision under cross-modal influence (Kang et al. 2006) and in visually-based deciphering of ambiguous auditory phonemes (Kilian-Hutten et al. 2011, proposed that the posterior temporal region is involved in predicting forthcoming auditory phonemes on the basis of the visual information, which corresponds closely to the present study. Lastly, the absence of negativity at the same latency for the visual-only deviance further supports the involvement of audio-visual networks.

A Predictive Role of Visual Prosodic Information During Speech Processing

Using the incongruent audio-visual prosodic features, we observed a pre-auditory MMN-like activity which is not present during the same visual stimuli in silence. We hypothesize that this response anticipates the auditory changes that follow the visual emphasis. This hypothesis is consistent with the general predictive role of visual information when processing auditory speech (van Wassenhove et al. 2005; Arnal et al. 2009), simply because there is a systematic delay between the visual gestures and the sound. This predictive and facilitative role has been demonstrated using behavioral tests, where information from the lip movement allows participants to distinguish between phonemes (e.g. [gy gu dy du] vs. [ty tu ky ku] Schwartz et al. 2004). At the neuronal level, electrophysiological studies have clearly demonstrated that the auditory cortex is sensitive to visual information, with activations found during speech reading (Calvert et al. 1997), and with visually-induced modifications of auditory responses (Besle et al. 2009; Davis et al. 2008; Reale et al. 2007). These multisensory interactions at early stages of auditory processing have an important facilitatory role in preparing the auditory system for optimal responses (Lakatos et al. 2007). For example, early audio-visual interactions for syllables are manifested as a latency shortening of the N1/P2 responses, which relates to the salience of the visual input (van Wassenhove et al. 2005). This suggests that the visual input carries an important predictive value for the auditory utterance. Furthermore, the visual facilitation of auditory responses is positively correlated with the predictive proficiency of the lip movements (van Wassenhove et al. 2005; Arnal et al. 2009). Altogether, this demonstrates the clear impact of visual speech information on neuronal auditory responses. Our results suggest that the visual prosodic cues give rise to similar effects on auditory brain activations, but on a longer time-scale. As psychophysical studies of audiovisual prosody demonstrated, both the moth opening and eyebrow positions modify the subjective perception of the auditory features (Chandrasekaran

et al. 2009; Munhall et al. 2004). Though the separation of these influences could present a certain theoretical interest, we believe that given their combined variation in continuous speech, the greatest effect can be obtained only for the whole ecological set of visual influences on the auditory prosody. Considering Fig. 1, one can see that the amplitude of the eyebrow displacement is about two times higher at the peak and for the whole variation in the curve, thus this movement may be a predominant cue.

During auditory perception, it has been demonstrated that brain reactions to prosodic violations depend on the expectations derived from the general context, including the pitch contour and linguistic expectations (Colombo et al. 2011). In our study, the MMN-like response is dependent on the multisensory context, as it is only present during the audio-visual blocks. We attribute this to the predictive role of the prosodic visual cues within the context of audio-visual speech. These results are consistent with context-dependent audio-visual interactions that have been observed in ERP studies using non-speech protocols (Stekelenburg and Vroomen 2007, 2009). Specifically, it has been found that if the preceding visual stimuli do not provide anticipatory information, the auditory N1 component is not affected.

Current theories claim that the brain generates predictions about the sensory environment, which are then compared to the actual incoming signal (Friston and Kiebel 2009; Strelnikov 2010, 2007). In the case of speech processing, lip movements provide predictive information which facilitates auditory processing, and this involves a network linking the visual and auditory areas (Arnal et al. 2009, 2011). For syllable perception, there is functional connectivity between the visual motion and auditory areas, which relates to the degree of visual predictability. It has been suggested that a fast direct cortico–cortical pathway conveys visual motion parameters to the auditory cortex, and that a slower indirect feedback pathway signals the error between the visual prediction and the auditory input (Arnal et al. 2009). There is evidence that the comparison of the signals involves the superior temporal sulcus (Arnal et al. 2009), see also Ghazanfar et al. 2008. Our analysis of the MMN-like component in continuous speech may suggest that negativity originates in the occipito-parietal electrodes, propagates forward to the fronto-temporal sites and then returning back to the occipital electrodes. Similar occipito–temporo–frontal loops have previously been shown for lip-reading syllables in silence (Arnal et al. 2009). This indicates that the visual information gives rise to auditory predictions in a fronto-temporal network, which are then compared to the next incoming auditory information. Prosodic cues do not transfer exactly the same meaning across languages in particular in what concerns tonal languages where specific MMN findings were demonstrated for native speakers (Chandrasekaran et al. 2007). Thus, one could also expect variations between the coupling of the visual and auditory prosodic cues across the speakers of different languages. An interesting perspective would be to study whether prosodic cues elicit the same type of modulation in native speakers versus non-native speakers.

A striking result is that, although the MMN-like response is probably driven by the visual prosodic cues, it can be observed only during the bimodal AV condition; no such negativity is observed at the same electrode location and latency in the V-only condition. Such an observation could suggest that the negativity reflects a certain process of visuo-auditory interaction. Indeed, it is now widely accepted that visual perception can be enhanced by the simultaneous presentation of an auditory cue as indicated by the decrease in visual perceptive threshold and the decrease in the reaction times to visual stimuli (Stein et al. 1996). Could it be possible that the MMN-like response we observed reflects a simple improvement to detect the visual emphasis in the context of the auditory continuous speech? First, in our video recording the visual emphasis on the word “two” was easily detected by all the subjects even in the V-only condition (tested during the pilot study). Thus, the perception of the visual only emphasis is already optimal and might not benefit from the auditory context. Besides, in the V-only condition we observed a small visual negativity located in the occipital visual cortex indicating that the brain reacted to the visual prosodic changes. In the case of an enhancement of the visual reaction in the visuo-auditory context, we might expect this visual negativity to be enhanced in its amplitude as observed during AV oddball protocols (Li et al. 2009). No such effect is observed in the present data suggesting that the MMN-like response in the continuous AV speech does not correspond to a simple enhancement of the visual perception but rather reflects a separate cognitive process linked to the auditory stimulus. To further investigate this phenomenon, one can vary the emphasis in the visual prosodic cues and assess the changes in both the illusion effect and the MMN-like response.

Conclusions

Speech comprehension constitutes what is probably a unique process of multisensory integration. In addition to the redundant information carried by visual lip movements, which are crucial in degraded auditory situations, the visual cues play a predictive role for the acoustic signals. Here we furthered our understanding of the facilitative role of visual information during continuous audio-visual speech processing. The MMN-like activity we observed suggests that

facial movements, which convey the prosodic stress of a word in continuous speech, act as a predictive representation for the forthcoming auditory inputs. These mechanisms of audio-visual interactions may be particularly important in patients with auditory deficits preventing prosody perception. For example, we have shown that cochlear implanted deaf patients present a strong deficit in detecting auditory prosodic cues (Marx et al. 2013). As these patients present strong skills in visuo-auditory integration (Barone and Deguine 2011), we might expect that they can develop mechanisms of audio-visual integration as a strategy to maintain prosody comprehension in noisy auditory environments. Indeed, cross-modal prosody processing can not only enhance the comprehension of speech in general (Munhall et al. 2004), but could also compensate for the reduced pitch perception as in cochlear implanted deaf patients (Chatterjee and Peng 2008; Donnelly et al. 2009). In addition, there are numerous observations that patients with focal brain lesions can suffer from aprosodia—a deficit in detecting prosodic cues (Ross and Monnot 2008). Like CI deaf patients, patients with aprosodia can also benefit from the visual support and visuo-auditory protocols could be included in their rehabilitation programs.

Acknowledgments We thank E. Barbeau for help in the first pilot study and C. Marlot for the bibliography. This study was supported by the Human Frontier Science Program (to JMF), the DRCI Toulouse (Direction de la Recherche Clinique et de l'Innovation to KS and MM), the ANR (ANR *Plasmody* ANR-11-BSHS2-0008 (to BP), and the recurrent funding of the CNRS (to BP).

References

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

Arnal LH, Wyart V, Giraud AL (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14(6):797–801

Barker J, Berthommier F (1999) Evidence of correlation between acoustic and visual features of speech. In: Ohala JJ, Hasegawa Y, Ohala M, Granville D, Bailey AC (eds) 14th international congress of phonetic sciences, San Francisco, USA, 1999. the congress organizers at the Linguistics Department, University of California, Berkeley, pp 199–202

Barkhuysen P, Krahmer E, Swerts M (2008) The interplay between the auditory and visual modality for end-of-utterance detection. J Acoust Soc Am 123(1):354–365

Barone P, Deguine O (2011) Multisensory processing in cochlear implant listeners. In: Zeng FG, Fay R, Popper A (eds) Springer handbook of auditory research. auditory prostheses: cochlear implants and beyond. Springer, New York, pp 365–382

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 20(8):2225–2234

Besle J, Fort A, Giard MH (2005) Is the auditory sensory memory sensitive to visual information? Exp Brain Res 166(3–4):337–344

Besle J, Fischer C, Bidet-Caulet A, Lecaignard F, Bertrand O, Giard MH (2008) Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. J Neurosci 28(52):14301–14310

Besle J, Bertrand O, Giard MH (2009) Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. Hear Res 258(1–2):143–151

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276(5312):593–596

Campanella S, Gaspard C, Debatisse D, Bruyer R, Crommelinck M, Guerit JM (2002) Discrimination of emotional facial expressions in a visual oddball task: an ERP study. Biol Psychol 59(3): 171–186

Campbell R (2008) The processing of audio-visual speech: empirical and neural bases. Philos Trans R Soc Lond B Biol Sci 363(1493):1001–1010

Cappe C, Thut G, Romei V, Murray MM (2010) Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. J Neurosci 30(38):12572–12580

Carpenter J, Bithell J (2000) Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. Stat Med 19(9):1141–1164

Cavé C, Guaïtella I, Bertrand R, Santi S, Harlay F (1996) Espesser R about the relationship between eyebrow movements and Fo variations. ICSLP, Philadelphia, pp 2175–2178

Chandrasekaran B, Krishnan A, Gandour JT (2007) Mismatch negativity to pitch contours is influenced by language experience. Brain Res 1128(1):148–156. doi:10.1016/j.brainres.2006.10.064

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5(7):e1000436

Chatterjee M, Peng SC (2008) Processing F0 with cochlear implants: modulation frequency discrimination and speech intonation recognition. Hear Res 235(1–2):143–156

Cohen J (1992) A power primer. psychol Bull 112(1):155–159

Colin C, Radeau M, Soquet A, Demolin D, Colin F, Deltenre P (2002) Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clin Neurophysiol 113(4):495–506

Colombo L, Deguchi C, Boureux M, Sarlo M, Besson M (2011) Detection of pitch violations depends upon the familiarity of intonational contour of sentences. Cortex 47(5):557–568

Davis C, Kislyuk D, Kim J, Sams M (2008) The effect of viewing speech on auditory speech processing is different in the left and right hemispheres. Brain Res 1242:151–161

de Gelder B, Bocker KB, Tuomainen J, Hensen M, Vroomen J (1999) The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. Neurosci Lett 260(2):133–136

Donnelly PJ, Guo BZ, Limb CJ (2009) Perceptual fusion of polyphonic pitch in cochlear implant users. J Acoust Soc Am 126(5):EL128–EL133

Foxton JM, Riviere LD, Barone P (2010) Cross-modal facilitation in speech prosody. Cognition 115(1):71–78

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364(1521):1211–1221

Friston K, Harrison L, Daunizeau J, Kiebel S, Phillips C, Trujillo-Barreto N, Henson R, Flandin G, Mattout J (2008) Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39(3):1104–1120. doi:10.1016/j.neuroimage.2007.09.048

Ghazanfar AA, Chandrasekaran C, Logothetis NK (2008) Interactions between the superior temporal sulcus and auditory cortex

mediate dynamic face/voice integration in rhesus monkeys. J Neurosci 28(17):4457–4469
Guaitella I, Santi S, Lagrue B, Cave C (2009) Are eyebrow movements linked to voice variations and turn-taking in dialogue? an experimental investigation. Lang Speech 52(Pt 2–3): 207–222
Hadar U, Steiner TJ, Rose FC (1984) Involvement of head movement in speech production and its implications for language pathology. Adv Neurol 42:247–261
Halgren E, Baudena P, Clarke JM, Heit G, Marinkovic K, Devaux B, Vignal JP, Biraben A (1995) Intracerebral potentials to rare target and distractor auditory and visual stimuli. II. medial, lateral and posterior temporal lobe. Electroencephalogr Clin Neurophysiol 94(4):229–250
Hertrich I, Mathiak K, Lutzenberger W, Menning H, Ackermann H (2007) Sequential audiovisual interactions during speech perception: a whole-head MEG study. Neuropsychologia 45(6):1342–1354
Jiang J, Alwan A, Keating PA, Auer ET, Bernstein LE (2002) On the relationship between face movements, tongue movements and speech acoustics. EURASIP J Appl Signal Process 11:1174–1188
Kang E, Lee DS, Kang H, Hwang CH, Oh SH, Kim CS, Chung JK, Lee MC (2006) The neural correlates of cross-modal interaction in speech perception during a semantic decision task on sentences: a PET study. Neuroimage 32(1):423–431. doi:10.1016/j.neuroimage.2006.03.016
Kilian-Hutten N, Vroomen J, Formisano E (2011) Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. Neuroimage 57(4):1601–1607. doi:10.1016/j.neuroimage.2011.05.043
Kimura M (2012) Visual mismatch negativity and unintentional temporal-context-based prediction in vision. Int J Psychophysiol 83(2):144–155. doi:10.1016/j.ijpsycho.2011.11.010
Kislyuk DS, Mottonen R, Sams M (2008) Visual processing affects the neural basis of auditory discrimination. J Cogn Neurosci 20(12):2175–2184
Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53(2):279–292
Li X, Yang Y, Ren G (2009) Immediate integration of prosodic information from speech and visual information from pictures in the absence of focused attention: a mismatch negativity study. Neuroscience 161(1):59–66
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164(1):177–190
Marx M, James C, Foxton J, Capber A, Fraysse B, Barone P, Deguine O (2013) Prosodic cues in cochlear implant users. In: 20th IFOS world congress, Seoul, June 2013
McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264(5588):746–748
Mottonen R, Krause CM, Tiippana K, Sams M (2002) Processing of changes in visual speech in the human auditory cortex. Brain Res Cogn Brain Res 13(3):417–425
Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E (2004) Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol Sci 15(2):133–137
Naatanen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118(12):2544–2590
Ponton CW, Bernstein LE, Auer ET Jr (2009) Mismatch negativity with visual-only and audiovisual speech. Brain Topogr 21(3–4): 207–215
Proverbio AM, Riva F (2009) RP and N400 ERP components reflect semantic violations in visual processing of human actions. Neurosci Lett 459(3):142–146
Quian Quiroga R, Garcia H (2003) Single-trial event-related potentials with wavelet denoising. Clin Neurophysiol 114(2):376–390
Reale RA, Calvert GA, Thesen T, Jenison RL, Kawasaki H, Oya H, Howard MA, Brugge JF (2007) Auditory-visual processing represented in the human superior temporal gyrus. Neuroscience 145(1):162–184
Ross ED, Monnot M (2008) Neurology of affective prosody and its functional-anatomic organization in right hemisphere. Brain Lang 104(1):51–74. doi:10.1016/j.bandl.2007.04.007
Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17(5):1147–1153
Saint-Amour D, De Sanctis P, Molholm S, Ritter W, Foxe JJ (2007) Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. Neuropsychologia 45(3):587–597
Sams M, Alho K, Naatanen R (1984) Short-term habituation and dishabituation of the mismatch negativity of the ERP. Psychophysiology 21(4):434–441
Sams M, Aulanko R, Hamalainen M, Hari R, Lounasmaa OV, Lu ST, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127(1):141–145
Scarborough R, Keating P, Mattys SL, Cho T, Alwan A (2009) Optical phonetics and visual perception of lexical and phrasal stress in English. Lang Speech 52(Pt 2–3):135–175
Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. Curr Opin Neurobiol 15(4):454–458
Schwartz JL, Savariaux C (2013) Data and simulations about audiovisual asynchrony and predictability in speech perception. the 12th international conference on auditory-visual speech processing, Annecy, France, 2013
Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93(2):B69–B78
Stein BE, London N, Wilkinson LK, Price DD (1996) Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. J Cogn Neurosci 8(6):497–506. doi:10.1162/jocn.1996.8.6.497
Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19(12):1964–1973
Stekelenburg JJ, Vroomen J (2009) Neural correlates of audiovisual motion capture. Exp Brain Res 198(2–3):383–390
Strelnikov K (2007) Can mismatch negativity be linked to synaptic processes? a glutamatergic approach to deviance detection. Brain Cogn 65(3):244–251
Strelnikov K (2008) Activation-verification in continuous speech processing. interaction of cognitive strategies as a possible theoretical approach. J Neurolinguist 21:1–17
Strelnikov K (2010) Neuroimaging and neuroenergetics: brain activations as information-driven reorganization of energy flows. Brain Cogn 72(3):449–456
Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26(2):212–215
Tales A, Newton P, Troscianko T, Butler S (1999) Mismatch negativity in the visual modality. Neuroreport 10(16):3363–3367
Ullsperger P, Erdmann U, Freude G, Dehoff W (2006) When sound and picture do not fit: mismatch negativity and sensory interaction. Int J Psychophysiol 59(1):3–7
van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA 102(4):1181–1186