Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence

David Hernández-Gutiérrez, Rasha Abdel Rahman, Manuel Martín-Loeches, Francisco Muñoz, Annekathrin Schacht, Werner Sommer

PII: S0010-9452(18)30122-9

DOI: 10.1016/j.cortex.2018.03.031

Reference: CORTEX 2295

To appear in: *Cortex*

Received Date: 15 August 2017

Revised Date: 23 March 2018

Accepted Date: 31 March 2018

Please cite this article as: Hernández-Gutiérrez D, Abdel Rahman R, Martín-Loeches M, Muñoz F, Schacht A, Sommer W, Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence, *CORTEX* (2018), doi: 10.1016/j.cortex.2018.03.031.

Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence

David Hernández-Gutiérrez[1*], Rasha Abdel Rahman[2], Manuel Martín-Loeches[1,3], Francisco Muñoz[1,3], Annekathrin Schacht[4,5], Werner Sommer[2]

[1] Center for Human Evolution and Behaviour UCM-ISCIII. Monforte de Lemos, 5, Pabellón 14, 28029 Madrid, Spain.

[2] Department of Psychology, Humboldt University of Berlin. Rudower Chaussee 18, 12489 Berlin, Germany

[3] Psychobiology Department, Complutense University of Madrid, Campus de Somosaguas. 28223 Pozuelo de Alarcón, Madrid, Spain

[4] Leibniz ScienceCampus Primate Cognition. Kellnerweg 4, 37077 Goettingen, Germany

[5] Affective Neuroscience and Psychophysiology Laboratory, University of Goettingen, Gosslerstrasse 14, 37073 Goettingen, Germany

[*]Correspondence should be addressed to David Hernández-Gutiérrez. Center for Human Evolution and Behaviour UCM-ISCIII. Monforte de Lemos, 5, Pabellón 14, 28029 Madrid, Spain.

E-mail address: dhernandez1@ucm.es

Highlights

Effects of speaker's face dynamics on semantic processing were investigated with ERPs

Expected and unexpected sentences were presented with either a video or a still frame

The speaker's face was either shown as a whole, with eyes covered, or mouth covered

N400 to unexpected words was unaffected by amount of facial information and motion

In dynamic conditions, expected words elicited a late posterior positivity

ABSTRACT

Face-to-face interactions characterize communication in social contexts. These situations are typically multimodal, requiring the integration of linguistic auditory input with facial information from the speaker. In particular, eye gaze and visual speech provide the listener with social and linguistic information, respectively. Despite the importance of this context for an ecological study of language, research on audiovisual integration has mainly focused on the phonological level, leaving aside effects on semantic comprehension. Here we used event-related potentials (ERPs) to investigate the influence of facial dynamic information on semantic processing of connected speech. Participants were presented with either a video or a still picture of the speaker, concomitant to auditory sentences. Along three experiments, we manipulated the presence or absence of the speaker's dynamic facial features (mouth and eyes) and compared the amplitudes of the semantic N400 elicited by unexpected words. Contrary to our predictions, the N400 was not modulated by dynamic facial information; therefore, semantic processing seems to be unaffected by the speaker's gaze and visual speech. Even though, during the processing of expected words, dynamic faces elicited a long-lasting late posterior positivity compared to the static condition. This effect was significantly reduced when the mouth of the speaker was covered. Our findings may indicate an increase of attentional processing to richer communicative contexts. The present findings also demonstrate that in natural communicative face-to-face encounters, perceiving the face of a speaker in motion provides supplementary information that is taken into account by the listener, especially when auditory comprehension is non-demanding.

Keywords: Language, Multimodal Processing, Social Neuroscience, Late Posterior Positivity, N400

Introduction

In human verbal communication, there is a natural prevalence for face-to-face interactions, involving the multimodal interplay of visual and auditory signals sent from the speaker to the listener. Though auditory information alone is sufficient for effective communication (Giraud & Poeppel, 2012), seeing the interlocutor's facial motions apparently provides further advantages (e.g., Crosse, Butler, & Lalor, 2015; Fort, Spinelli, Savariaux, Kandel, 2013; Peelle & Sommers, 2015; Rohr & Abdel Rahman, 2015; van Wassenhove, 2013). Some authors refer to this effect as *visual enhancement* (Peelle & Sommers, 2015), underscoring that human communication involves multisensory adaptation. Audiovisual integration in language processing is becoming, therefore, an area of growing interest.

Most of the literature on audiovisual integration in language processing has focused on the phonological level. Visual speech seems to increase the ability of a listener to correctly perceive utterances (Cotton, 1935; Sumby & Pollack, 1954), increase the speed at which phonemes are perceived (Soto-Faraco, Navarra & Alsius, 2004), and may even alter the perception of phonemes (McGurk & MacDonald, 1976). This multisensory gain depends on various factors, including spatial congruency, temporal coincidence, behavioral relevance, and experience (for review, see van Atteveldt, Murray, Thut & Schroeder, 2014).

Audiovisual integration has also been studied with electrophysiological measures like event-related brain potentials (ERPs). This technique is characterized by fine-grained temporal resolution and allows investigating the neural mechanisms underlying multisensory integration at different levels. At the phonological level, the facilitation provided by audiovisual integration is reflected in shorter latencies (Alsius, Möttönen, Sams, Soto-Faraco & Tiippana, 2014; Baart, Stekelenburg & Vroomen, 2014; Knowland, Mercure, Karmiloff-Smith, Dick & Thomas, 2014; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005) and smaller amplitudes (Hisanaga, Sekiyama, Igasaki & Murayama, 2016; Stekelenburg & Vroomen, 2007, 2012a; van Wassenhove et al., 2005) of the auditory N1 and P2 components of the ERP. Moreover, studies with functional magnetic resonance imaging and magnetic field potentials have reported that visual input about the speaker's lip positions or movements can modulate the activity of the primary auditory cortex (Calvert et al., 1997, Lakatos, Karmos, Mehta, Ulbert & Schroeder, 2008).

Available evidence suggests that audiovisual integration also facilitates lexical access at the semantic level as shown with *cross-modality priming.* In this paradigm, a silent video of a speaker uttering a (prime) word is followed by the auditory-only version of the critical word. Such priming by visual speech can improve semantic categorizations (Dodd, Oerlemens & Robinson, 1989), lexical decisions (Kim, Davis & Krins, 2004; Fort et al., 2013), and word recognition (Buchwald, Winters, & Pisoni, 2009) of critical words. These findings support an influence of visual speech on lexical or post-lexical processes and indicate that visual and auditory speech modalities share cognitive resources (Buchwald et al., 2009).

In typical audiovisual communication, facial gestures precede the auditory input by about 150 ms (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). A set of lexical candidates might therefore be available before the utterance can be heard (Fort et al., 2013) and, consequently, semantic processing might be easier if the visually pre-activated word matches the auditory input. Further, audiovisual presentation of complex texts improves performance compared to auditory-only conditions, as shown with comprehension questionnaires (Arnold & Hill, 2001; Reisberg, McLean & Goldfield, 1987). In sum, the perception of the speaker´s facial dynamics cannot only improve phonetic perception but also semantic comprehension.

Evidence regarding the neural correlates underlying audiovisual integration at the semantic level during sentence processing from on-line measures of brain activity, such as ERPs, is surprisingly scarce. To our knowledge, the only pertinent ERP study has been reported by Brunellière, Sánchez-García, Ikumi, and Soto-Faraco (2013). In a first experiment, these authors manipulated the semantic constraints (expectancy) of critical words within audiovisual sentences, as well as the articulatory saliency of lip movements. These variables interacted during the late part of the N400, an ERP component reflecting the access to semantic knowledge during language comprehension (Kutas & Federmeier, 2011). As compared to low visual articulatory saliency, high saliency increased the N400 amplitude for unexpected words and yielded a wide effect across the scalp. In a second experiment, Brunellière et al. (2013) compared the effect of visual articulatory saliency with respect to an audio-alone condition without manipulating semantic constraints. Words with high articulatory saliency yielded a significant N400 effect that was enhanced under audiovisual presentation relative to the audio-alone condition, which the authors interpreted in terms of late phonological effects.

The present study aimed to add further evidence to the scarce literature on audiovisual integration at the semantic level by comparing the neural processing of expected and unexpected words in spoken sentences. Sentences were presented either in a dynamic audiovisual mode, showing videos of the speaker, as compared to a still face mode, showing pictures of the speaker. This paradigm allowed exploring how the dynamics of speaker's facial movements impact the semantic processing of words during sentence comprehension. Our study therefore focused on semantic processing of connected speech while the speaker's face is seen in dynamic versus static mode.

The majority of studies on audiovisual processing of language did not consider that visual perception of face-to-face contexts is not restricted to oro-facial (i.e., mouth) speech movements. However, the perception of the eyes and their gaze direction strongly captures and directs attention (Conty, Tijus, Hugueville, Coelho & George, 2006; von Grünau & Anston, 1995; Senju & Hasegawa, 2005) and modulates the activity of neurons in auditory cortex (van Atteveld et al., 2014). Evolutionary evidence supports the importance of eyes in human communication, like the white sclera adaptation specific to humans (Kobayashi & Kohshima, 2001; Tomasello, Hare, Lehmann & Call, 2007). A study in macaques demonstrated enhanced activity of ventrolateral prefrontal neurons in response to combining vocalizations with pictures of direct-gaze faces (Romanski, 2012), demonstrating the role of this brain area in the integration of social-communicative information.

Gaze perception seems to influence social interactions and communication among humans. For instance, conversations typically begin with eye contact between individuals

(Schilbach, 2015) and communicative intent is usually signalled by direct gaze (Farroni, Csibra, Simion & Johnson, 2002; Gallagher, 2014). Hence, eyes are extremely informative both about the mental state of the interaction partner and for ascertaining what a speaker demands from a listener (Myllyneva & Hietanen, 2015). Eye contact between persons can modulate concurrent cognitive and behavioural activities, a phenomenon known as the "eye contact effect" (Senju & Johnson, 2009), which are mediated by the social brain network, including areas such as fusiform gyrus, superior temporal sulcus, medial prefrontal and orbitofrontal cortex, and amygdala. Hence, eyes are evidently "special" visual stimuli for humans.

As reviewed above, facial information can be relevant for language comprehension, and is not restricted to oro-facial movements, but also includes other cues such as eye gaze. The present ERP study explored the effects of facial dynamic information on semantic processing by manipulating the presence of two main sources of information in the face, mouth (visual speech) and eyes (gaze). To this aim, we compared the amplitude of the semantic N400 component elicited by unexpected words within audiovisual connected speech. Expectancy of critical words within a given sentence was manipulated by a preceding context sentence. This allowed comparing exactly the same stimulus material across conditions with a maximum degree of experimental control, by merely exchanging the preceding context sentence. In parallel to the auditory material, participants were presented with two kinds of visual information, consisting in either a video of the speaker's face and upper torso (dynamic conditions) or stills of the speaker taken from these videos (static control conditions). In three experiments, we investigated the effects of visual speech and gaze on the semantic processing of connected speech on ERPs, focusing on the N400 component: In Experiment 1, participants were presented with the whole speaker's face in dynamic and static versions. In Experiments 2 and 3, the specific contributions of facial information to the observed effects were studied by concealing either the eyes or the mouth, respectively.

In Experiment 1, we expected modulations of the N400 component by dynamic information provided by the whole face. According to the literature reviewed above, visibility of lip movements should affect lexical access and semantic processing of unexpected words, leading to increased amplitudes of the N400 component (Brunellière et al., 2013). As direct eye contact has been shown to attract attention and to activate a broad network of social brain areas, resources for semantic processes might be diminished, presumably reflected in reduced N400 amplitudes. Therefore, we expected an increase of the N400 amplitude when the speaker's eyes are occluded and only visual speech is available (Experiment 2) and the N400 might be reduced when the eyes are available and the mouth is covered (Experiment 3).

Experiment 1. Whole face presentation

This experiment investigated audiovisual processing of words in spoken sentences in an ecological context, that is, perceiving the whole face of the speaker while listening to connected speech. In the video condition, participants could therefore focus on both eye gaze and visual speech available from the whole face of the speaker while concurrently listening to auditory speech.

Method

Participants

Twenty-one native German speakers participated in this experiment. One of them was excluded because of poor EEG quality. The remaining 20 individuals (15 females, 5 males; age range 20 - 35 [*M* = 27.3]) were all right-handed (Mean Oldfield scores +88) and declared normal or corrected-to-normal vision and normal hearing. Participants gave written informed consent and received monetary reimbursement. The study was approved by the ethics committee of the Humboldt-Universität at Berlin (application number 2013-43 R) and conducted in accordance with the declaration of Helsinki.

Materials and procedure

The linguistic material used in this experiment was taken from the Postdam Sentence Corpus 3 (Dambacher et al., 2012). Stimuli consisted of 144 sentence units, each containing two different context sentences and two target sentences. A context sentence defined the expectancy of a critical word. Depending on the context sentence a critical word of the exact same target sentence could be expected (cloze probability = .84 –high frequency– and .83 –

low frequency–; SD = .13 in both cases) or unexpected (cloze probability = .01, SD = .02, regardless of frequency). Therefore, the material allowed comparing expected and unexpected critical words by using exactly the same word stimuli in both conditions (for further details on the material, see Dambacher et al., 2012). An example (in English translation from the original German) is given below:

1. Expected combinations:

• *The man on the picture fiddled around with models of Columbus' fleet. (context 1). In his right hand he held a* ***ship*** *of considerable length (target 1).*

• *The man on the picture wore a golden crown and sat stately on a throne (context 2). In his right hand he held a* ***scepter*** *of considerable length (target2).*

2. Unexpected combinations:

• *The man on the picture fiddled around with models of Columbus' fleet (context 1). In his right hand he held a* ***scepter*** *of considerable length (target 2).*

• *The man on the picture wore a golden crown and sat stately on a throne (context 2). In his right hand he held a* ***ship*** *of considerable length (target 1).*

Four sets of sentences were prepared for this experiment, each containing only one of the four possible sentence combinations per sentence unit. These four stimulus sets were presented to different participants in balanced fashion, such that none of the sentences were repeated to a given participant. Moreover, all sentence combinations of a given sentence unit were presented equally often across participants. Each set of sentences contained the same number of expected and unexpected critical words. According to these manipulations, we recorded four videos for each of the 144 sentence units. Sentences were spoken by a male speaker with neutral prosody, neutral emotional expression, and direct eye gaze. There was an SOA of 500 ms between the start of the video and the start of the first vocalization. Auditory speech was synchronized with visual speech.

In the dynamic condition, participants were presented with the videos together with the audio tracks of the sentences. For the static condition, static pictures of the speaker – with mouth closed – were shown while participants listened to the audio tracks (Fig. 1). Because the audio files were taken from the video recordings, the only difference between dynamic and static conditions was the dynamics of the visual stimuli. The audio files of the sentences in all conditions were matched in acoustic intensity, with mean intensity values of 68.4 dB.

Videos and static pictures were displayed on a computer screen (1280 x 1024 pt) at a viewing distance of 80 cm. Auditory sentences were presented through a pair of speakers placed at both sides of the screen. Sound levels were kept constant for all participants. Every participant completed 144 trials (36 dynamic/unexpected, 36 dynamic/expected, 36 static/unexpected, 36 static/expected). The assignment of the sets of sentences to participants was counterbalanced. The setting of the triggers at the onset of critical words was done with GoldWave software by three independent persons. The decisions about trigger settings were based on both the visual pattern of the sound wave and the auditory onset of the initial phoneme of the critical word. The three trigger setting values were averaged to obtain an objective time-point. Since sentences in the expected and unexpected conditions were the same, trigger time-points were identical. Consequently, any differences between conditions cannot be due to differences in trigger positions or voice onsets~~.~~

EEG-recordings

The EEG was recorded from 62 electrodes placed according to the international 10-20 system. Impedances were kept below 5 kΩ. The vertical electrooculogram (VEOG) was recorded from below versus above the left eye, and the horizontal electrooculogram (HEOG) was recorded from the outer canthus of each eye. Except for the electrode below the left eye all other were placed within an elastic cap. The signals were recorded continuously with a bandpass from 0.1 to 100 Hz and a sampling rate of 250 Hz. The EEG recordings were initially referenced to the left mastoid (M1); offline, the EEG was re-referenced to average mastoids, and a low-pass filter of 15 Hz was applied.

Data Analysis

EEG epochs of 1150 ms were extracted, starting 150 ms before critical word onset. Ocular correction for blinks and eye movements was performed by Independent Component Analysis (ICA). Trials with remaining artifacts, exceeding a range of 100 μV were semi-automatically rejected. Overall, the mean rate of rejected segments was 13%. Each experimental condition was averaged individually. The average amplitude during the first 150 ms served as pre-stimulus baseline.

Visual inspection of the ERPs confirmed the expected effects on the N400 in centro-parietal areas in all participants (Fig. 2). Therefore, repeated-measures ANOVAs were performed using a region of interest (ROI) comprising electrode sites Cz, CP1, CPz, CP2, P3, Pz, P4, PO3, POz, and PO4. To reduce the number of statistical comparisons, the individual values at the electrodes within this ROI were collapsed to their mean. The analysis included the following factors: Expectancy (expected vs. unexpected), and Presentation Mode (dynamic vs. static). Separate ANOVAs were performed on the average amplitudes within the following time windows, based upon visual inspection: 300-450, 450-600, 600-750 and 750-900 ms after critical word onset. Huynh-Feldt corrections were applied when appropriate.

Results and Discussion

Figure 2 shows the ERPs to expected and unexpected words. While unexpected words yielded an N400-like negative-going modulation between 300 and 600 ms, expected words showed a parietal positivity between 300 and 900 ms. The ANOVA revealed significant effects of Expectancy between 300 and 900 ms (300-450 ms: $F(1,19) = 43.07$; $p = .000$; $\eta^2 = .694$; $\pi = 1$; 450-600 ms: $F(1,19) = 51.02$; $p = .000$; $\eta^2 = .729$; $\pi = 1$; 600-750 ms: $F(1,19) = 40.15$; $p = .000$; $\eta^2 = .679$; $\pi = 1$; 750-900 ms: $F(1,19) = 31.41$; $p = .000$; $\eta^2 = .623$; $\pi = 1$). The main effect of Presentation Mode was statistically significant between 300-900 ms except for the interval 600-750 ms (300-450 ms: $F(1,19) = 10.7$; $p = .004$; $\eta^2 = .362$; $\pi = .875$; 450-600 ms: $F(1,19) = 9.52$; $p = .06$; $\eta^2 = .334$; $\pi = .833$; 600-750 ms: $F(1,19) = 2.7$; $p = .114$; $\eta^2 = .126$; $\pi = .350$; 750-900 ms: $F(1,19) = 8$; $p = .01$; $\eta^2 = .298$; $\pi = .769$). When comparing dynamic and static modes (Fig. 3) for expected words a long-lasting posterior positivity emerged specifically in the dynamic mode. For unexpected words, by contrast, ERPs in both dynamic and static modes appeared to be largely identical, displaying very similar N400 deflections between 300 and 900 ms. Supporting these impressions ANOVA of ERPs amplitudes yielded a significant interaction Expectancy by Presentation Mode (300-450 ms: $F(1,19) = 7.49$; $p = .013$; $\eta^2 = .283$; $\pi = .738$; 450-600 ms: $F(1,19) =$

9.08; p = .007; η^2 = .323; π = .816; 600-750 ms: F(1,19) = 9.1; p = .007; η^2 = .323; π = .816; 750-900 ms: F(1,19) = 8.1; p = .01; η^2 = .299; π = .771).

To further investigate these results, and given the apparent differences between expected and unexpected critical words, likely involving two components of opposing polarities over similar regions (N400 versus late positivity), separate ANOVAs were performed for the two conditions. Presentation Mode was therefore the only factor included in the following ANOVAs. On the one hand, Presentation Mode was not significant at all in the unexpected words between 300 and 900 ms (300-450 ms: F(1,19) = 0.21; p = .887; η^2 = .001; π = .052; 450-600 ms: F(1,19) = 0.00; p = .993; η^2 = .00; π = .00; 600-750 ms: F(1,19) = 0.21; p = .566; η^2 = .018; π = .086; 750-900 ms: F(1,19) = 0.422; p = .524; η^2 = .022; π = .095), indicating that the N400 was insensitive to this factor. On the other hand, within ERPs to expected words Presentation Mode effects were significant during the 300-900 ms interval (300-450 ms: F(1,19) = 14.6; p = .001; η^2 = .953; 450-600 ms: F(1,19) = 13.99; p = .001; η^2 = .424; π = .944; 600-750 ms: F(1,19) = 14.5; p = .001; η^2 = .433; π = .951; 750-900 ms: F(1,19) = 12.38; p = .002; η^2 = .395; π = .916), substantiating the long-lasting positivity in the dynamic condition.

Overall, and of main interest, the N400 to unexpected words was insensitive to the type of presentation since no amplitude modulation could be observed in the spoken word-elicited N400 accompanied by the dynamic video presentation as compared to a static picture of the speaker. Therefore, neither mouth movements (visual speech) nor eye gaze seemed to have an impact on the N400 component, reflecting the efforts of semantic processing when an unexpected word occurs (Kutas & Federmeier, 2011). It seems therefore that activation of resources to deal with unexpected lexical information is prioritized over other types of information, such as dynamic social cues depicted in the eyes or mouth of the speaker.

Interestingly, however, when spoken words were expected, social cues seem to play a relevant role. Dynamic cues of the speaker's face likely elicited a long-lasting late centro-parietal positivity. Although the dynamics of the stimuli presented during the baseline differed among conditions, it cannot explain the emergence of this component. In this case, the effect of different baselines should have impacted the N400 as well, but this was not at all the case. Therefore, the late posterior positive component may reflect an increase in motivated attention to the dynamic relative to the static face. Such an interpretation was suggested for comparable positivities in a study by Schindler, Wegrzyn, Steppacher and Kissler (2015; c.f. General Discussion), which presumably occurs to words emitted by a dynamic face. Alternatively, the posterior positivity observed here might indicate that during easy-to-process verbal conditions, that is in case of expected words, the participants had larger resources available for attentively inspecting the video (Rohr & Abdel Rahman, 2015).

Experiments 2 and 3 were designed to determine whether the increased posterior positivity during dynamic as compared to static visual face input while listening to expected words are specifically related to eye gaze, dynamic mouth information (visual speech), or to a combination of both. In addition, we wanted to test whether the stability of the N400 holds after removing information from the eyes or mouth.

- Insert Figure 2 about here -

- Insert Figure 3 about here -

Experiment 2 – Eyes covered

The aim of this experiment was to explore the influence of dynamic mouth movements (visual speech) on semantic processing of connected speech comprehension. The procedure of Experiment 2 was the same as in Experiment 1, except that the eyes of the speaker were covered (Fig. 1).

Method

Participants

Twenty native German speakers (9 females, 11 males; age range 18 - 28 [M = 25.04]), different from those in Experiment 1, participated in this experiment. All declared to have normal or corrected-to-normal vision, normal hearing, and were right-handed (mean

Oldfield score +79). Participants gave written informed consent and received monetary reimbursement.

Materials and procedure

Stimulus materials and their presentation were identical to Experiment 1, except that a skin-coloured opaque bar covering the speaker's eyes was digitally added to both videos and pictures.

EEG recordings and data analysis

The EEG-recording settings were as described for Experiment 1. Due to the presence of artifacts, on average 11.5 % of the recorded EEG segments had to be excluded from data analysis.

Results and Discussion

A large N400 was obtained in the ERPs to unexpected words (Fig. 4, left), very similar to the results of Experiment 1. In turn, for the expected words (Fig. 4, right), the long-lasting positivity in the dynamic mode was again observed, though apparently smaller than in Experiment 1. The main ANOVA, including both Expectancy and Presentation Mode as factors, showed significant effects of Expectancy between 300 and 750 ms (300-450ms: $F(1,19) = 57.43$; $p < .001$; $\eta^2 = .751$; $\pi = 1$; 450-600 ms: $F(1,19) = 39.72$; $p < .001$; $\eta^2 = .676$; $\pi = 39.7$; 600-750 ms: $F(1,19) = 12.66$; $p = .002$; $\eta^2 = .4$; $\pi = .921$), and as a trend between 750 and 900 ms ($F(1,19) = 3.31$; $p = .084$; $\eta^2 = .149$; $\pi = .409$). Presentation Mode had significant effects between 450-600 ms ($F(1,19) = 11.72$; $p = .03$; $\eta^2 = .382$; $\pi = .901$), but not in the other time windows analyzed (300-450 ms: $F(1,19) = 1.37$; $p = .255$; $\eta^2 = .068$; $\pi = .2$; 600-750 ms: $F(1,19) = 0.75$; $p = .395$; $\eta^2 = .038$; $\pi = .131$; 750-900 ms: $F(1,19) = 2.02$; $p = .171$; $\eta^2 = .096$; $\pi = .272$). The Expectancy by Presentation Mode interaction reached a trend for the interval 450-600 ms ($F(1,19) = 4.14$; $p = .056$; $\eta^2 = .179$; $\pi = .489$), and no other time segment reached or approached significance (300-450ms: $F(1,19) = 0.25$; $p = .62$; $\eta^2 = .013$; $\pi = .077$; 600-750 ms: $F(1,19) = 0.27$; $p = .609$; $\eta^2 = .014$; $\pi = .079$; 750-900 ms: $F(1,19) = 0.01$; $p = .912$; $\eta^2 = .001$; $\pi = .051$).

For the same reasons as in Experiment 1, ANOVAs were performed for expected and unexpected words separately. Since the interaction Expectancy by Presentation Mode was a very strong trend between 450 and 600 ms, and in order to focus on a pure measure of the late positivity, separate analyses were conducted on the late posterior positivity for expected words. The ANOVA for unexpected words revealed no differences in the N400 between dynamic and static modes (300-450 ms: $F(1,19) = 0.09$; $p = .757$; $\eta^2 = .005$; $\pi = .06$; 450-600 ms: $F(1,19) = 0.33$; $p = .572$; $\eta^2 = .017$; $\pi = .085$; 600-750 ms: $F(1,19) = 0.02$; $p = .88$; $\eta^2 = .001$; $\pi = .052$; 750-900 ms: $F(1,19) = 0.91$; $p = .35$; $\eta^2 = .0146$; $\pi = .149$). Both conditions also showed the same scalp distribution (Fig. 4). In the ANOVA for expected words Presentation Mode was significant as a main effect in the 450-600 ms window ($F(1,19) = 15.14$; $p = .001$; $\eta^2 = .444$; $\pi = .958$) but failed significance in the other time windows analyzed (300-450 ms: $F(1,19) = 1.2$; $p = .286$; $\eta^2 = .06$; $\pi = .181$; 600-750 ms: $F(1,19) = 0.8$; $p = .381$; $\eta^2 = .041$; $\pi = .136$; 750-900 ms: $F(1,19) = 1.26$; $p = .275$; $\eta^2 = .062$; $\pi = .187$).

- Insert Figure 4 about here -

In sum, a late posterior positive component for expected words in dynamic conditions appeared when eyes of the speaker were covered, though attenuated in amplitude and temporally more restricted as compared to Experiment 1, where the whole face was visible. This result appeared obscured in the main ANOVA including both expected and unexpected words, maybe because of the presence of components with opposite polarities for the different Expectancy conditions (N400 and late posterior positivity). Visibility of the eyes seems therefore to be one of the elements contributing to the effects obtained in Experiment 1. However, since the late posterior positivity was again elicited, the visibility of the mouth (or other parts of the face) might also be relevant for this effect. To directly test this possibility, we conducted Experiment 3, where the mouth of the speaker was covered.

Experiment 3. Mouth covered

By occluding the mouth region (Fig. 1), Experiment 3 investigated whether information other than lip movements – leaving the eyes as most important facial features visible - contributes to the effect of dynamic versus static presentation mode, observed in Experiment 1.

Method

Participants

Twenty native German speakers, other those of Experiments 1 or 2, participated in this experiment (13 females, 7 males; age range 18 - 34 [*M* = 25.6]). All declared to have normal or corrected-to-normal vision, normal hearing, and were right-handed (Mean Oldfield scores +75). Participants gave written informed consent and received reimbursement for their participation.

Materials and procedure

Materials and procedure were the same as in Experiment 1 except that a skin-coloured opaque bar covering the speaker's mouth was added to both videos and pictures.

EEG-recordings and data analysis

The EEG-recording settings were as described for Experiment 1. Due to the presence of artifacts, on average 7.4 % of the recorded EEG segments had to be excluded from data analysis.

Results and Discussion

Again, a large N400 was obtained in the ERPs to unexpected words (Fig. 5, right), highly resembling the corresponding results of Experiments 1 and 2. For the expected words (Fig. 5, left), the long-lasting posterior positivity in the dynamic mode emerged again, smaller, however, than in Experiment 1. The main ANOVA, including the factors Expectancy and Presentation Mode, showed significant effects of Expectancy between 300 and 750 ms (300-450ms: $F(1,19) = 26.08$; $p < .001$; $\eta^2 = .57$; $\pi = .998$; 450-600 ms: $F(1,19) = 34.03$; $p < .001$; $\eta^2 = .642$; $\pi = 1$; 600-750 ms: $F(1,19) = 14.2$; $p = .001$; $\eta^2 = .428$; $\pi = .947$) but not between 750 and 900 ms ($F(1,19) = 2.93$; $p = .103$; $\eta^2 = .134$; $\pi = .370$). The Presentation Mode effect was statistically significant between 450 and 750 ms (450-600 ms: $F(1,19) = 1.6$; p = .221; $\eta^2 = .078$; $\pi = .225$; 600-750 ms: $F(1,19) = 5.76$; $p = .027$; $\eta^2 = .233$; $\pi = .625$). However it did not reach statistical significance between 300 and 400 ms ($F(1,19) = 1.35$; p = .259; $\eta^2 = .066$; $\pi = .197$), and between 750 and 900 ms ($F(1,19) = 1.72$; $p =$

.204; η^2 = .083; π = .239). The Expectancy by Presentation Mode interaction did not yield significant effects (300-450 ms: *F*(1,19) = 0.31 ; *p* = .584; η^2 = .016; π = .083; 450-600 ms: *F*(1,19) = 1.16; *p* = .293; η^2 = .058; π = .177; 600-750 ms: *F*(1,19) = 0.68; *p* = .419; η^2 = .035; π =.123; 750-900 ms: *F*(1,19) = 0.00; *p* = .951; η^2 = .000; π = .05).

As before, separate ANOVAs for expected and unexpected words were performed. Although the interaction of Expectancy by Presentation Mode failed significance in the main ANOVA, both Expectancy conditions were analysed separately to assure that the co-occurrence of an N400 for unexpected words would not affect the detection of a simultaneous posterior positivity for expected words. Thus, consistent analyses were performed throughout all three experiments. The ANOVA for unexpected words revealed no differences in the N400 between dynamic and static presentation modes (300-450 ms: *F*(1,19) = 0.42; *p* = .84; η^2 = .002; π = .054; 450-600 ms: *F*(1,19) = 0.02; *p* = .966; η^2 = .00; π = .05; 600-750 ms: *F*(1,19) = 0.57; *p* = .458; η^2 = .029; π = .111; 750-900 ms: *F*(1,19) = 0.37; *p* = .547; η^2 = .019; π = .09). In contrast, the ANOVA to expected words revealed a significant effect of Presentation Mode between 600 and 750 ms, (*F*(1,19) = 4.84; *p* = .04; η^2 = .203; π = .551) but not in the other time windows (300-450 ms: *F*(1,19) = 1.20; *p* = .287; η^2 = .059; π = .18; 450-600 ms: *F*(1,19) = 3.14; *p* = .92; η^2 = .142; π = .391; 750-900 ms: *F*(1,19) = 0.37; *p* = .547; η^2 = .019; π = .09).

- Insert Figure 5 about here –

Summarizing Experiment 3, a late posterior positivity for expected words in dynamic mode still appeared when the mouth was covered, though apparently weaker and temporally more restricted as compared to Experiment 1 (whole face) but similar to that in Experiment 2 (eyes covered). Therefore, lip movements (retained in the videos) seem to also contribute to the effects observed in Experiment 1.

On the other hand, the N400 in response to unexpected words was again unaffected by the dynamic or static presentation mode of the speaker’s face, very similar to the results of Experiments 1 and 2. Hence, the N400 to unexpected words does not seem to be modulated by any of the facial cues manipulated here. This will be discussed in more detail below.

Comparison of Experiments 1, 2, and 3

Data Analysis

In order to directly compare the potential impact of different face areas visible during speech processing, we conducted analyses of the effects of type of presentation and expectancy of words across all three facial feature presentation conditions: whole face (Exp.1), eyes covered (Exp. 2), and mouth covered (Exp. 3). To this aim, a mixed ANOVA was first performed including the factors Facial Feature (Experiment) as group factor and Expectancy and Presentation Mode as within-subject factors. To further compare the N400 and the late posterior positivity across facial features (i.e., experiments) devoid of the conceivable confounds caused by the inclusion of two components of opposite polarity over similar regions – possibly inducing type II errors – , separate mixed ANOVAs were performed for the expected and unexpected conditions, with Facial Feature as group factor and Presentation Mode as within-subject factor. Bonferoni-corrected post-hoc pairwise comparisons were applied to significant interactions; Huynh-Feld corrections were applied where appropriate.

Results and Discussion

The main ANOVA on ERPs including the factors Facial Feature, Expectancy and Presentation Mode showed significant main effects of Expectancy between 300 and 900 ms (300-450 ms: $F(1,57) = 121.41$; $p < .001$; $\eta^2 = .681$; $\pi = 1$; 450-600 ms: $F(1,57) = 124.31$; $p < .001$; $\eta^2 = .686$; $\pi = 1$; 600-750 ms: $F(1,57) = 59.25$; $p < .001$; $\eta^2 = .51$; $\pi = 1$; 750-900 ms: $F(1,57) = 23.53$; $p < .001$; $\eta^2 = .292$; $\pi = .998$). The interaction of Expectancy with Facial Feature reached statistical significance between 300 and 450 ms ($F(2,57) = 98.36$; $p = .041$; $\eta^2 = .106$; $\pi = .6134$) and trends between 450 and 600 ms ($F(2,57) = 2.48$; $p =.093$; $\eta^2 = .08$; $\pi = .479$) and between 750 and 900 ms ($F(2,57) = 2.67$; $p = .078$; $\eta^2 = .086$; $\pi = .51$), but it did not even approach statistical significance in the 600-700 ms time window ($F(1,57) = 1.54$; $p = .223$; $\eta^2 = .051$; $\pi = .314$). Presentation Mode reached significance between 300 and 900 ms (300-450 ms: $F(1,57) = 148.46$; $p = .001$; $\eta^2 = .165$; $\pi = .909$; 450-600 ms:

$F(1,57) = 19.45$; $p < .001$; $\eta^2 = .254$; $\pi = .991$; 600-750 ms: $F(1,57) = 7.96$; $p = .007$; $\eta^2 = .123$; $\pi = .792$; 750-900 ms: $F(1,57) = 10.55$; $p = .002$; $\eta^2 = .156$; $\pi = .891$), but no interval showed significant effects for the Presentation Mode by Facial Feature interaction (300-450 ms: $F(2,57) = 27.58$; $p = .133$; $\eta^2 = .068$; $\pi = .412$; 450-600 ms: $F(2,57) = 1.17$; $p = .316$; $\eta^2 = .04$; $\pi = .248$; 600-750 ms: $F(2,57) = 0.7$; $p = .5$; $\eta^2 = .024$; $\pi = .163$; 750-900 ms: $F(2,57) = 0.84$; $p = .435$; $\eta^2 = .029$; $\pi = .188$). The ANOVA also yielded a significant interaction Expectancy by Presentation Mode between 300 and 750 ms (300-450 ms: $F(2,57) = 5.07$; $p = .028$; $\eta^2 = .82$; $\pi = .601$; 450-600 ms: $F(2,57) = 12.38$; $p = .001$; $\eta^2 = .179$; $\pi = .933$; 600-750 ms: $F(2,57) = 6.54$; $p = .013$; $\eta^2 = .103$; $\pi = .711$), but the interval 750-900 ms did not reach statistical significance ($F(2,57) = 1.82$; $p = .182$; $\eta^2 = .031$; $\pi = .264$). The Expectancy by Presentation Mode by Facial Feature interaction did not reach significance in any interval (300-450 ms: $F(2,57) = 1.84$ $p = .167$; $\eta^2 = .061$; $\pi = .370$; 450-600 ms: $F(2,57) = .903$; $p = .411$; $\eta^2 = .031$; $\pi = .198$; 600-750 ms: $F(2,57) = 1.95$; $p = .151$; $\eta^2 = .064$; $\pi = .389$; 750-900 ms: $F(2,57) = 1.42$; $p = .249$; $\eta^2 = .048$; $\pi = .293$).

The significance of main effects of Presentation Mode clearly conflicts with visual inspection of the data as well as with separate ANOVAs performed for the three Experiments individually, which systematically showed that unexpected words were blind to presentation mode. On the other hand, the absence of Expectancy by Presentation Mode by Facial Feature interaction would also be at odds with main ANOVAs performed for the three Experiments in sequence, as the latter indicated that no significant effects were observed for Expectancy by Presentation Mode in Experiment 3 and informed about a trend in Experiment 2. All these features endorse that mixing expected and unexpected words in the same analyses is being prone to statistical inaccuracies probably as a consequence of conflating two different components of opposite polarities over similar scalp areas (cf. Brower & Crocker, 2017; Luck, 2005) and, therefore, that separate ANOVAs are the most appropriate approach to analyse the present results.

The separate ANOVA to expected words revealed significant main effects of Presentation Mode between 300 and 900 ms (300-450 ms: $F(1,57) = 13.33$; $p = .001$; $\eta^2 = .190$; $\pi = .948$; 450-600 ms: $F(1,57) = 29.55$; $p < .001$; $\eta^2 = .341$; $\pi = .1$; 600-750 ms:

$F(1,57) = 16.5$; $p < .001$; $\eta^2 = .225$; $\pi = .979$; 750-900 ms: $F(1,57) = 11.54$; $p = .001$; $\eta^2 = .168$; $\pi = .916$). These results confirm the presence of the long-lasting late posterior positivity for dynamic as compared to static modes across the three experiments. The interaction of Presentation Mode and Facial Feature was significant in the time window 300-450 ms ($F(2,57) = 3.49$; $p = .037$; $\eta^2 = .109$; $\pi = .631$), and re-appeared as trends between 600 and 900 ms (600-750 ms: $F(1,19) = 2.43$; $p = .096$; $\eta^2 = .079$; $\pi = .472$; 750-900 ms: $F(1,19) = 2.46$; $p = .094$; $\eta^2 = .08$; $\pi = .476$).

Post-hoc pairwise comparisons for the 300 and 450 ms interval revealed that the posterior positivity in the dynamic mode was significantly larger for Experiment 1 than Experiment 3 ($\Delta = 2.106$ µV, $p = .004$) (Fig. 6). This difference was present also in the time window 600-750 ms though only as a trend ($\Delta = 1.37$ µV, $p = .084$) but was significant again later between 750-900 ms ($\Delta = 1.7$ µV, $p = .008$). No other significant comparison was found.

The ANOVA of ERP amplitudes to unexpected words did not yield any significant differences for Presentation Mode (300-450 ms: $F(1,57) = 0.04$; $p = .828$; $\eta^2 = .001$; $\pi = .055$; 450-600 ms: $F(1,57) = 0.123$; $p = .727$; $\eta^2 = .002$; $\pi = .064$; 600-750 ms: $F(1,57) = 0.00$; $p = .944$; $\eta^2 = .00$; $\pi = .051$; 750-900 ms: $F(1,57) = 1.53$; $p = .22$; $\eta^2 = .026$; $\pi = .23$), or for the interaction of Facial Feature and Presentation Mode (300-450 ms: $F(1,57) = 0.05$; $p = .942$; $\eta^2 = .002$; $\pi = .059$; 450-600 ms: $F(1,57) = 0.1$; $p = .905$; $\eta^2 = .004$; $\pi = .065$; 600-750 ms: $F(1,57) = 0.53$; $p = .589$; $\eta^2 = .018$; $\pi = .134$; 750-900 ms: $F(1,57) = 0.048$; $p = .953$; $\eta^2 = .002$; $\pi = .057$).

- Insert Figure 6 about here -

In sum, the long-lasting late posterior positivity for expected words appeared whenever the stimulus was presented in a dynamic context. Interestingly, this positivity was significantly reduced when the mouth was covered. Statistical analyses showed a significant interaction in the 300-450 ms window, while post-hoc analyses of this segment were significant only when comparing Experiments 1 (whole face) and 3 (mouth covered). In contrast, the N400 amplitude in ERPs to unexpected words was robust and insensitive against all our manipulations; it was of similar amplitude regardless of the dynamic or static presentation mode of the face and whether the face was shown at full view or whether mouth or eyes of the speaker were occluded.

General Discussion

In the present study, we investigated whether dynamic features of a speaker's face can facilitate the semantic processing of connected speech, as compared to a static condition. We recorded ERPs to expected and unexpected words, embedded in spoken sentences, in two different conditions. In one condition, the spoken utterances appeared together with a video of the speaker; in the other condition, a static picture of the face was shown. Importantly, the same critical words embedded in the same sentences were either expected or unexpected, depending on a preceding context sentence (cf. Dambacher et al., 2012). This manipulation ruled out any confounds due to different (acoustic) properties of the stimulus material. As expected, a large N400 component appeared to the unexpected compared to expected critical words (Kutas & Federmeier, 2011). Contrary to our predictions, this ERP deflection was not modulated at all by dynamic versus static presentation modes in any of the three experiments. A consistent finding across all three experiments, however, was a long-lasting posterior positivity in the condition with expected words in the dynamic relative to the static condition; such an effect was not seen in the condition with unexpected critical words. Experiments 2 and 3 were conducted to disentangle the contribution of different facial features of the speaker, which is eyes and mouth, to these effects.

No modulation of the N400

As mentioned, presentation mode did not have any impact on the semantic processing of sentences, reflected in the N400. If we had focused on the N400 *effect*, that is, the difference between unexpected and expected words, differences between dynamic and static conditions in this ERP fluctuation would have emerged, but merely as a consequence of the late posterior positivity in the expected word condition (Fig. 2). For this reason, we focused on the N400 *component*. According to our results, online semantic word processing in sentences seems unaffected by the perception of facial dynamics (visual speech and eye gaze). Articulatory visual saliency of speaker lip movements has been shown to impact the N400 in dynamic situations (Brunellière et al., 2013; van Wassenhove et al., 2005). We did not control the articulatory visual saliency of our critical words because our main manipulation was of a different kind. In this regard, our dynamic condition might be considered as more salient than the static one, but no difference in the N400 was present between these two situations.

Other factors might be in play in accounting for our results. In the light of limited resources, the human cognitive system focuses on highly relevant stimuli while ignoring other information via selective attention (Lakatos et al., 2008; Li et al., 2016; Schroeder & Lakatos, 2009). Unexpected words might make listeners devote more resources to the semantic processing of the auditory stimulus and ignore visual stimuli; in this case, effects may be similar irrespective of the dynamic versus static information in the face. On the contrary, listening to expected words does not require much effort and might liberate resources otherwise devoted to semantic processing, permitting attention to be directed to the visual information in the face. According to our results, only when semantic processing is not cognitively demanding (expected words), there is an influence of visual dynamic information provided in parallel to the auditory input.

It should also be considered that facilitating effects of multisensory integration might be weak or absent when the input coming from both sensory streams is unambiguous and can be perceived clearly (Colonius & Diederich, 2004; Hong & Shim, 2016; Stanford, Quessy & Stein, 2005). A few studies demonstrated visual speech to be advantageous over phonological and lexical processes when using noisy contexts (Buchwald et al., 2009; Fort et al., 2010; Sumby & Pollack, 1954). Under noise-free situations, auditory information alone may be enough to complete a lexical decision task, rendering both dynamic and static modes redundant (Fort et al., 2010; 2013). As the present experiment used clearly audible spoken sentences, audiovisual integration might not have been required to semantically process the unexpected words.

Posterior positivity to expected words in dynamic presentation

The long-lasting late posterior positivity to expected words in the dynamic relative to the static condition was surprising because there seem to be no previous reports of such an effect. Its amplitude may relate to both visual speech and direct dynamic gaze perception during a simulated communicative situation. In fact, we observed the largest and longest (300-900 ms) effect when the whole face was visible (Exp. 1). A visual inspection of Figure 6 revealed apparent differences between the three experiments in the amplitude of the late posterior positivity, being most pronounced when the whole face accompanied speech in dynamic motion. Covering the eyes or the mouth seemed to attenuate this modulation, though only the post-hoc comparison between Experiments 1 and 3 (mouth-covered) yielded a significant reduction of the effect relative to the whole face.

Our late posterior positivity is similar to the ERP waves consistently reported by Schindler and colleagues in the frame of social contexts (Schindler et al., 2015; Schindler & Kissler, 2016; 2017). They asked participants to describe themselves on a video to be watched by another participant next door. Immediately afterwards, participants were presented a feedback consisting of evaluative adjectives on a screen (e.g. “happy”, “weak”). These adjectives were claimed to either to stem from the other participant or to be random generated by a computer. This manipulation resulted in a continuous long-lasting posterior positivity, interpreted as a P3 component followed by an LPP (late positive potential). Interestingly, the observed long-lasting posterior positivity was larger when participants were informed that feedback was given by a human sender rather than by a computer. This increased posterior positivity seems to resemble our effect of dynamic versus static presentation. It is important to note, however, that Schindler and colleagues did not use face stimuli, and therefore their results cannot be related to specific facial features. Schindler and colleagues suggested their results as an influence of sender information on language processing due to different communicative contexts and interpreted them as effects of higher motivated attention directed to a human sender compared to the computer. Similarly, communicative situations, in which a talking face is presented, have recently been found to enhance (emotional) word processing (Rohr & Abdel Rahman, 2015). In our study, the social context of the dynamic mode was richer than the static mode. Seeing the whole face of a speaker in motion appears to be the condition that most resembles a natural communicative context.

Schindler and colleagues framed their data within the motivated attention model (Lang, Bradley, & Cuthbert, 1997), adding that context manipulations and not only the emotional content could amplify the motivational relevance of written words. Our findings are consistent with this idea, indicating that motivated attention can be manipulated by contextual factors, possibly enhancing the late processing of words. Further, they extend the effect to connected speech in a multisensory context, similar to face-to-face communication.

The posterior positivity diminished when covering the mouth of the speaker, but not significantly when covering the eyes. In social interactions, direct gaze is important because it allows the beholder to make communicative inferences (Gallagher, 2014). It seems necessary to underline, however, that the characteristics of our stimuli prevent us from firmly attributing the late posterior positivity effect to the importance of the mouth over gaze. The main reason is that on the videos the speaker’s mouth was more dynamic than his eyes, which gazed at the participant rather constantly. The difference between covering the eyes and the mouth might not be related to any particular facial feature, but mainly to differences on overall dynamics and perception of motion. More research is necessary to confirm our results. It would be useful to compare specific contributions of eyes and mouth motions in similar dynamic contexts.

Possibly, verbal communication cues from the mouth dominate over the social cues from eye gaze. In fact, mouth dynamics are directly related to the auditory stimuli; hence the interplay and coherence between mouth movements and utterances may enhance language comprehension and be advantageous in communication contexts. In line with this thought, Western people (our sample) relative to Eastern seem to focus their attention on the mouth of the speaker more than on her gaze during audiovisual speech perception (Hisanaga, Sekiyama, Igasaki & Murayama, 2016). Consequently, dynamic presentation of both speaker’s eyes and mouth importantly contribute to modulate the processing of connected speech, with a larger contribution of the mouth.

Conclusions

The main finding of the present study is the higher attentional processing to contexts that resemble most strongly natural communicative situations, as long as semantic speech processing is not very demanding. Contrary to our predictions, we could not find any modulation of the N400 semantic effect by the concomitant dynamic facial information. The speaker's dynamic facial features did not affect semantic processing. When semantic comprehension is more demanding (i.e., when an unpredicted or unexpected word is presented), visual cues are not critical or disregarded for language processing. By contrast, material that is predictable in a linguistic stream generates, when accompanied by the dynamic face of the speaker, a long-lasting late posterior positivity. This positivity is reminiscent of the communicative effect described by Schindler and colleagues (Schindler et al., 2015; Schindler & Kissler, 2016; 2017), interpreted as an increase of motivated attention in richer social contexts. The present study has provided new information about the interaction of audiovisual social-contexts and word processing of connected speech. Hence, it is a step forward towards the study of language comprehension in real-life situations.

Acknowledgments

The authors thank Johannes Rost for stimulus preparation, Hossein Sabri for data collection, Anna Eiserbeck for her help in data analysis, and Guido Kiecker for technical support. This research was supported by the structured graduate program "Self Regulation Dynamics Across Adulthood and Old Age: Potentials and Limits". It was funded by the DFG excellence initiative “Talking heads”, from the DAAD, [grant code 57049661], and the Spanish Ministerio de Economía y Competitividad [grant code PSI2013-43107-P] (MINECO, Programa Estatal de Investigación Científica y Técnica de Excelencia, Subprograma Estatal de Generación de Conocimiento y Proyectos de I+D).

Conflict of Interest Statement

All the authors declare no conflict of interest.

References

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

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.*, *92*(2), 339-355.

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*, 115-121.

Brouwer, H., & Crocker, M. W. (2017). On the Proper Treatment of the N400 and P600 in Language Comprehension. *Frontiers in Psychology, 8,* 1327.

Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *Int. J. Psychophysiol*. *89*(1), 136-147.

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Lang. Cogn. Process*, *24*(4), 580-610.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P.W.R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593-596.

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

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cogn. Neurosci.*, *16*(6), 1000-1009.

Conty, L., Tijus, C., Hugueville, L., Coelho, E., & George, N. (2006). Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: a behavioural study. *Spat. Vis.*, *19*(6), 529-545.

Cotton, J. C. (1935). Normal “visual hearing". *Science*, 82, 592-593

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.*, *35*(42), 14195-14204.

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852-1870.

Dodd, B., Oerlemens, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Lang.*, 22, 59–77.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proc. Nat. Acad. Sci.*, *99*(14), 9602-9605.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Lang. Cogn. Proc.*, *28*(8), 1207-1223.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Commun.*, *52*(6), 525-532.

Gallagher, S. (2014). In your face: transcendence in embodied interaction. *Front. Hum. Neurosci.*, *8*.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.*, *15*(4), 511-517.

Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/Culture Modulates Brain and Gaze Processes in Audiovisual Speech Perception. *Sci. Rep.*, *6*.

Hong, S. W., & Shim, W. M. (2016). When audiovisual correspondence disturbs visual processing. *Exp. Brain. Res.*, *234*(5), 1325-1332.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1), B39-B47.

Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. (2014). Audio-visual speech perception: a developmental ERP investigation. *Dev. Sci.*, *17*(1), 110-124.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *J. Hum. Evol.*, *40*(5), 419-435.

Kutas, M., & Federmeier, K.D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annu. Rev. Psychol., 62*, 621-647.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110-113.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In P. J. Lang, R. F. Simons, & M. T. Balaban (Eds.), *Attention and Orienting: Sensory and Motivational Processes* (pp. 97-135). Mahwah, NJ: LEA.

Li, Y., Long, J., Huang, B., Yu, T., Wu, W., Li, P., Fang, F., & Sun, P. (2016). Selective Audiovisual Semantic Integration Enabled by Feature-Selective Attention. *Sci. Rep.*, *6*.

Luck, S. J. (2005). *An introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976) Hearing lips and seeing voices. *Nature,* 264(5588): 746–748.

Myllyneva, A., & Hietanen, J. K. (2015). There is more to eye contact than meets the eye. *Cognition*, *134*, 100-109.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp 97-113). London: Erlbaum.

Rohr, L., & Abdel Rahman, R. (2015). Affective responses to emotional words are boosted in communicative situations. *NeuroImage*, *109*, 273-282.

Romanski, L. M. (2012). Integration of faces and vocalizations in ventral prefrontal cortex: implications for the evolution of audiovisual speech. *Proc. Nat. Acad. Sci.*, *109* (Suppl. 1), 10717-10724.

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Curr. Opin. Behav. Sci.*, *3*, 130-135.

Schindler, S., & Kissler, J. (2016). People matter: Perceived sender identity modulates cerebral processing of socio-emotional language feedback. *NeuroImage*, *134*, 160-169.

Schindler, S., & Kissler, J. (2017). Language-based social feedback processing with randomized 'senders': An ERP study. *Soc. Neurosci.*, 1-12.

Schindler, S., Wegrzyn, M., Steppacher, I., & Kissler, J. (2015). Perceived communicative context and emotional content amplify visual word processing in the fusiform gyrus. *J. Neurosci.*, *35*(15), 6010-6019.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trend. Neurosci.*, *32*(1), 9-18.

Senju, A., & Hasegawa, T. (2005). Direct gaze captures visuospatial attention. *Vis. Cogn.*, *12*(1), 127-144.

Senju, A., & Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trend. Cogn. Sci.*, *13*(3), 127-134.

Soto-Faraco, S., Navarra, J., Alsius, A. (2004) Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, 92(3):B13–B23.

Stanford, T. R., Quessy, S., & Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J. Cogn. Neurosci.*, *25*(28), 6499-6508.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.*, *19*(12), 1964-1973.

Stekelenburg, J. J., & Vroomen, J. (2012a). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Front. Integr. Neurosci.*, *6*.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, *26*(2), 212-215.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *J. Hum. Evol.*, *52*(3), 314-320.

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron*, *81*(6), 1240-1253.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Nat. Acad. Sci.*, *102*(4), 1181-1186.

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Front. Psychol.*, 4, 388.

von Grünau, M., & Anston, C. (1995). The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, *24*(11), 1297-1313.

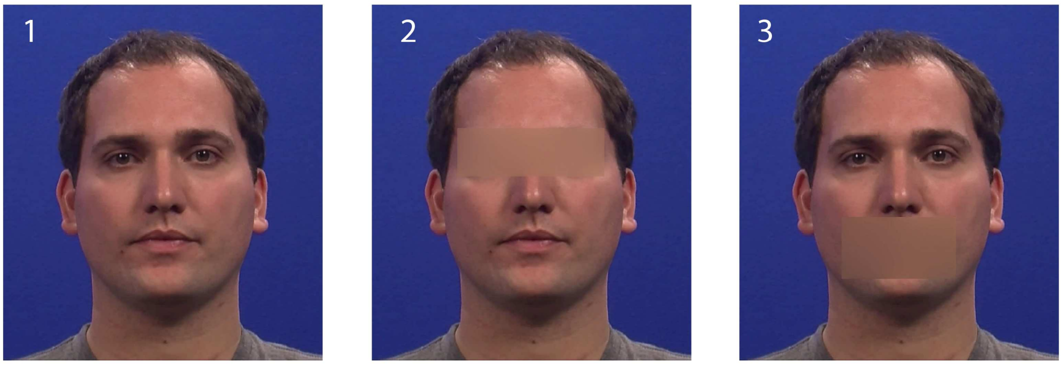

Figure 1. Visual stimuli of the static condition. (1) Whole face -Experiment 1-, (2) Eyes covered -Experiment 2-, (3) Mouth covered -Experiment 3-.

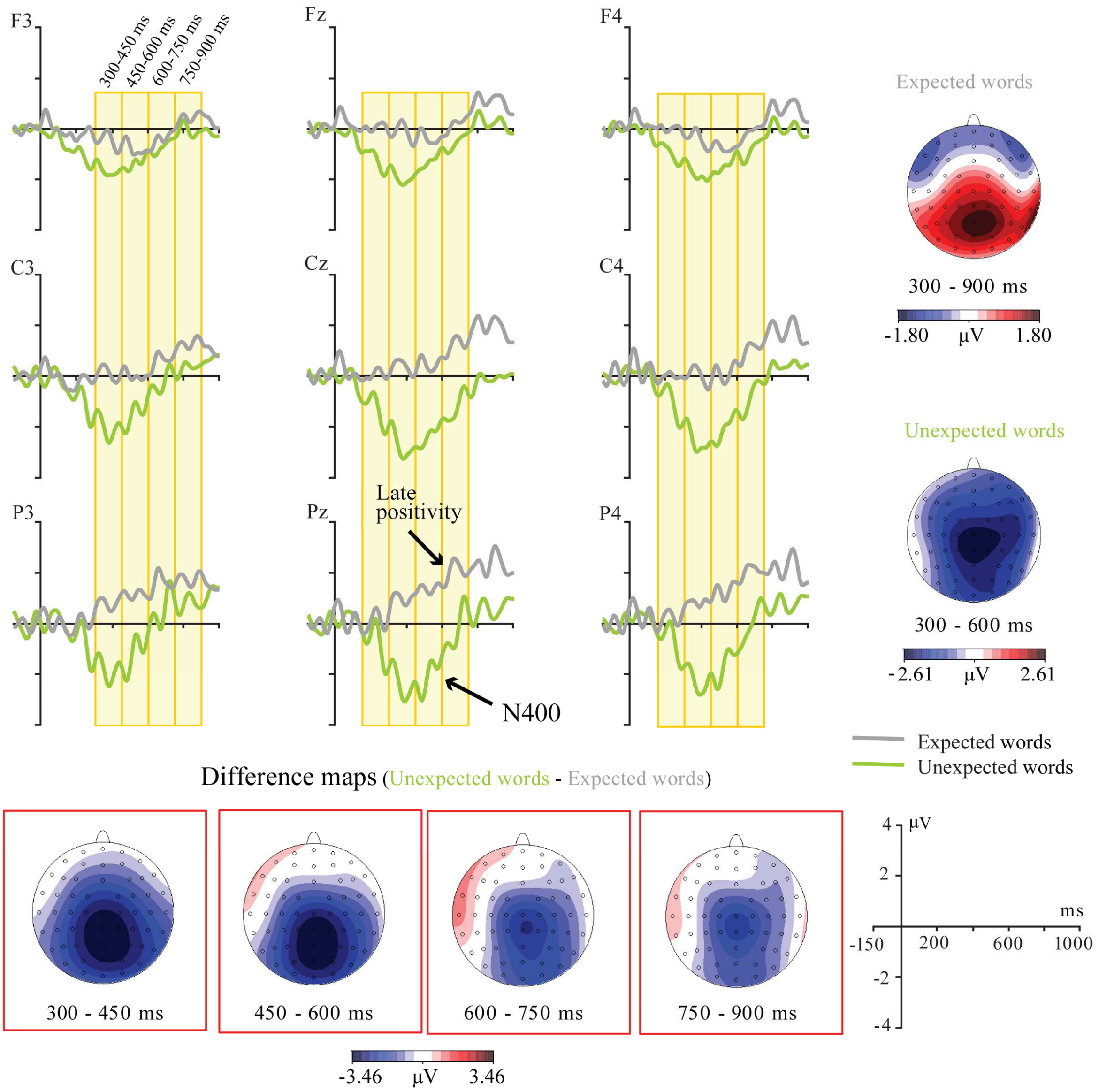

Figure 2. ERPs to expected (grey) and unexpected words (green), in the whole face condition (experiment 1). The ERP topographies for each condition correspond to a late posterior positivity for expected words (300-900 ms) and an N400 for unexpected words (300-600 ms). Therefore, each condition yielded a different component. The difference maps depict an N400 effect.

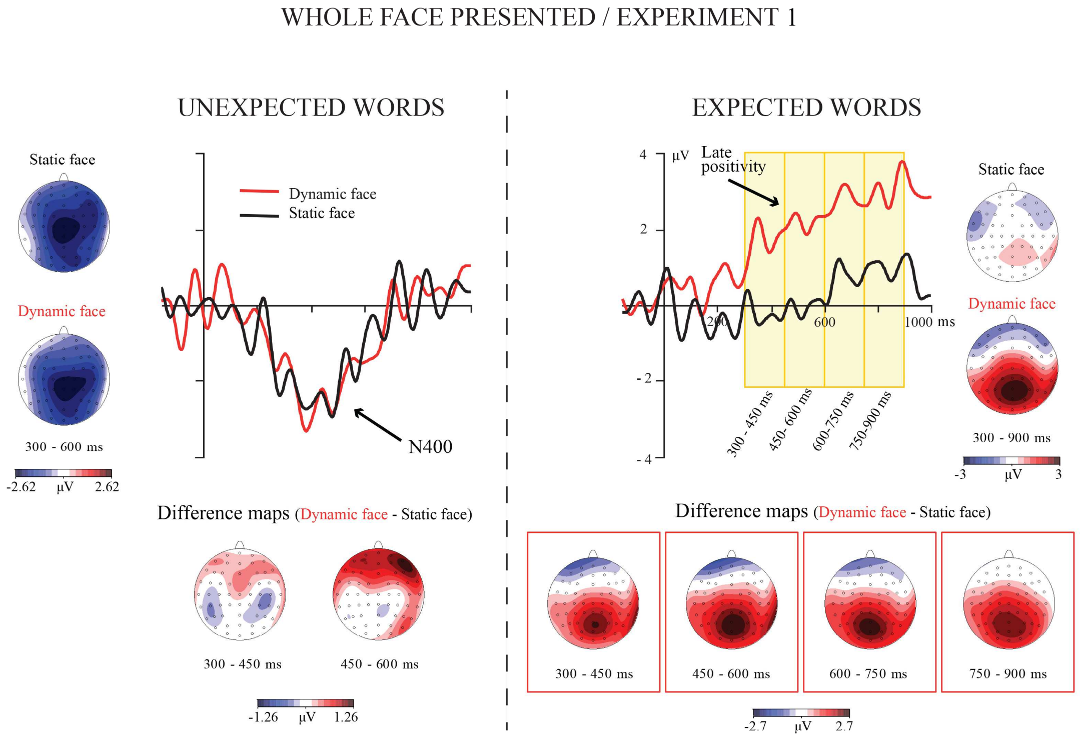

Figure 3. Effects of Presentation Mode on expected and unexpected words in the whole face condition (Experiment 1). ERPs of the ROI electrodes are pooled. The ERP topographies represent each component individually for each presentation mode. The difference maps represent the Presentation Mode effect (dynamic minus static) for unexpected (left) and expected words (right).

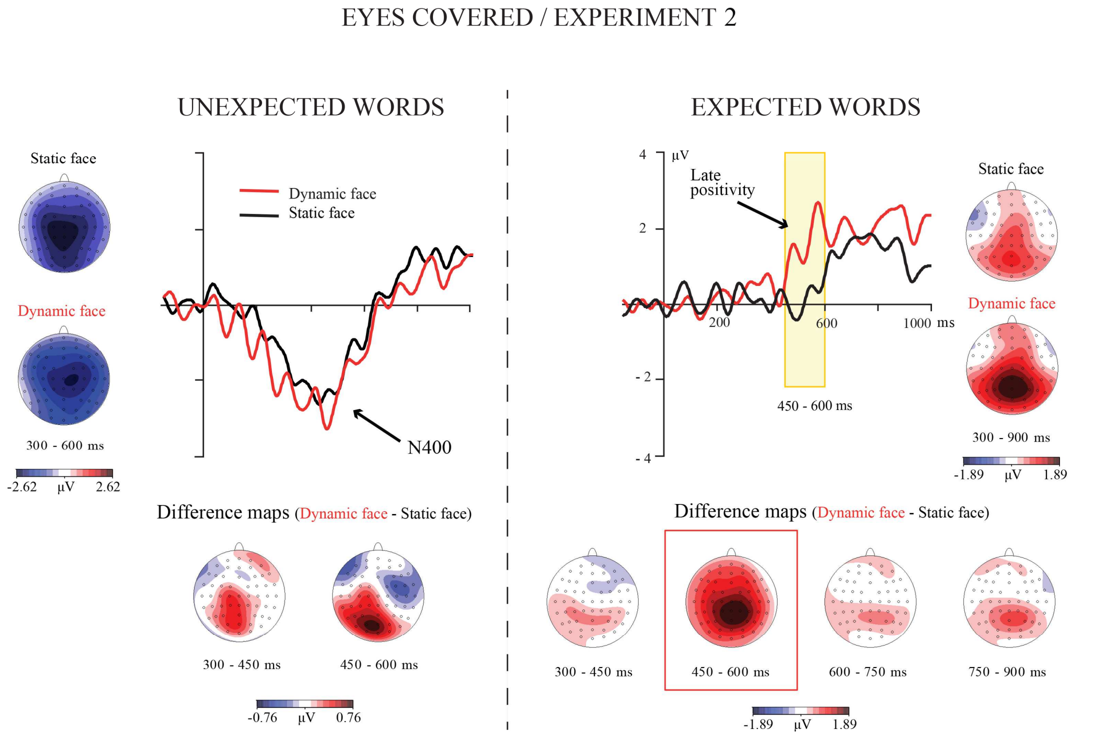

Figure 4. Effects of Presentation Mode on expected and unexpected words in the eyes covered condition (Experiment 2). ERPs of the ROI electrodes are pooled. The ERP topographies represent each component individually for each presentation mode. The difference maps represent the Presentation Mode effect (dynamic minus static) for unexpected (left) and expected words (right).

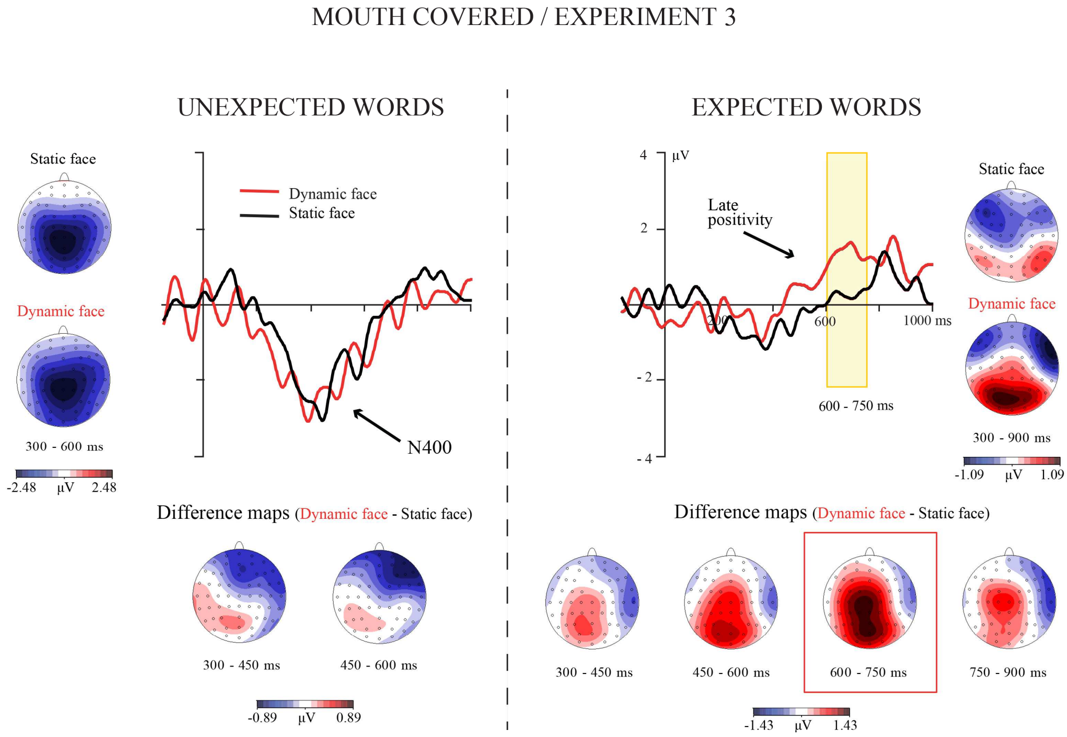

Figure 5. Effects of Presentation Mode on expected and unexpected words in the mouth covered condition (Experiment 3). ERPs of the ROI electrodes are pooled. The ERP topographies represent each component individually for each presentation mode. The difference maps represent the Presentation Mode effect (dynamic minus static) for unexpected (left) and expected words (right).

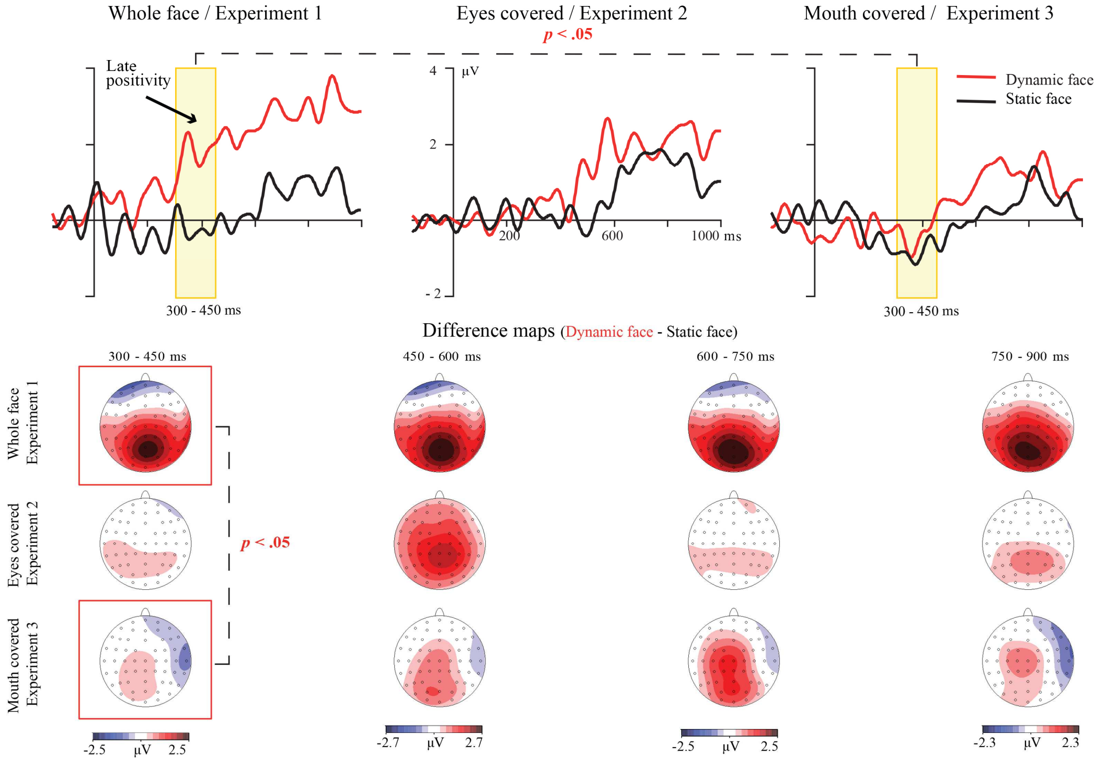

Figure 6. Effects of Presentation Mode on expected words in each experiment. ERPs of the ROI electrodes are pooled. The difference maps represent the Presentation Mode effect (dynamic minus static) for each Facial Feature.