Neuropsychologia ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neuropsychologia

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

Electrophysiological evidence for speech-specific audiovisual integration

Martijn Baart [a,b], Jeroen J. Stekelenburg [b], Jean Vroomen [b,*]

[a] Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd floor, 20009 Donostia, Spain
[b] Tilburg University, Department of Cognitive Neuropsychology, P.O. Box 90153, Warandelaan 2, 5000 LE, Tilburg, the Netherlands

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form
7 November 2013
Accepted 19 November 2013

Keywords:
N1
P2
Audiovisual speech
Sine-wave speech
Audiovisual integration

ABSTRACT

Lip-read speech is integrated with heard speech at various neural levels. Here, we investigated the extent to which lip-read induced modulations of the auditory N1 and P2 (measured with EEG) are indicative of speech-specific audiovisual integration, and we explored to what extent the ERPs were modulated by phonetic audiovisual congruency. In order to disentangle speech-specific (phonetic) integration from non-speech integration, we used Sine-Wave Speech (SWS) that was perceived as speech by half of the participants (they were in speech-mode), while the other half was in non-speech mode. Results showed that the N1 obtained with audiovisual stimuli peaked earlier than the N1 evoked by auditory-only stimuli. This lip-read induced speeding up of the N1 occurred for listeners in speech and non-speech mode. In contrast, if listeners were in speech-mode, lip-read speech also modulated the auditory P2, but not if listeners were in non-speech mode, thus revealing speech-specific audiovisual binding. Comparing ERPs for phonetically congruent audiovisual stimuli with ERPs for incongruent stimuli revealed an effect of phonetic stimulus congruency that started at ~200 ms after (in)congruence became apparent. Critically, akin to the P2 suppression, congruency effects were only observed if listeners were in speech mode, and not if they were in non-speech mode. Using identical stimuli, we thus confirm that audiovisual binding involves (partially) different neural mechanisms for sound processing in speech and non-speech mode.

1. Introduction

Like most natural events, human speech is multimodal in nature. When listening to a talker in a face-to-face conversation, the auditory speech signal is accompanied by visual articulatory gestures that, quite often, precede the onset of the auditory signal (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Although it is well-established that the brain integrates the auditory and lip-read signals into one coherent speech percept (e.g., McGurk & MacDonald, 1976; Sumby & Pollack, 1954) there is an ongoing debate whether audiovisual (AV) speech integration is qualitatively different from non-speech AV integration (e.g., Kuhl, Williams, & Meltzoff, 1991; Liberman & Mattingly, 1989; Massaro, 1998; Repp, 1982). A stimulus par excellence to contrast AV integration of speech versus non-speech is so-called sine-wave speech (hence SWS, see Remez, Rubin, Pisoni, & Carrell, 1981). In SWS, the natural spectral richness of the speech signal is reduced to sinusoids that follow the center frequency and the amplitude of the first three formants. Typically, naïve listeners perceive SWS as 'non-speech' sounds like whistles or computer bleeps. However, once listeners are told that these sounds are derived from speech, they cannot switch back to a non-speech mode again and continue to hear the SWS as speech (they are then in 'speech mode').

By using SWS, it has been demonstrated that AV integration has phonetic and non-phonetic behavioral consequences. A phonetic consequence is that lip-read information can change (or bias) the perceived identity of a speech sound (e.g., an auditory /b/ combined with lip-read /g/ is often perceived as /d/, McGurk & MacDonald, 1979). A non-phonetic consequence is that lip-read information can change the saliency, apparent origin, or loudness of a sound. As an example of the latter, Eskelund, Tuomainen, and Andersen (2011) investigated whether lip-read speech increases detection of auditory noised-masked SWS and observed that lip-read information induced an auditory detection benefit irrespective of whether the listeners were in speech- or non-speech mode. Audiovisual synchrony is another feature of audiovisual speech that appears not to be speech-specific. Vroomen and Stekelenburg (2011) introduced an AV temporal asynchrony between SWS stimuli and lip-read videos and demonstrated that sensitivity to temporal asynchrony was alike for listeners that heard SWS as speech or non-speech. In both studies, though, only listeners in speech mode showed phonetic AV integration (i.e., identification

* Corresponding author. Tel.: +31 13 466 2394.
E-mail address: J.Vroomen@uvt.nl (J. Vroomen).

0028-3932/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011

of the sound was biased by lip-read information (see Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Baart, 2009 for similar findings)).

Recently, Stekelenburg and Vroomen (2012b) showed that the electrophysiological mismatch negativity (i.e., MMN, e.g., Näätänen, Gaillard, & Mäntysalo, 1978), as induced by lip-read information that triggers an illusory phonetic auditory percept (Colin, Radeau, Soquet, & Deltenre, 2004; Colin et al., 2002; Kislyuk, Möttönen, & Sams, 2008; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Sams et al., 1991), only occurred when listeners were in speech mode. Taken together, these findings indicate that across experiments, listeners bound the AV information together into a coherent phonetic event only when in speech mode, but not when in non-speech mode.

Although it thus seems likely that AV phonetic integration of SWS with lip-read information takes place either at, or prior to, the time-point at which the McGurk-MMN is generated, the MMN only provides an indirect upper bound to the time-course that underlies phonetic AV integration (see also Besle, Fort, Delpuech, & Giard, 2004). Moreover, the McGurk-MMN can only be elicited by incongruent stimuli in which the deviant differs acoustically (Kislyuk et al., 2008) or visually (e.g., Colin et al., 2002) from the standard, and it therefore cannot disentangle detection of a 'change in AV congruency' from phonetic AV integration proper.

Here, we used a different method to examine the time-course of AV phonetic integration by measuring ERPs from listeners in speech- or non-speech mode who were presented SWS and lip-read input. The Dutch pseudo-words /tabi/ and /tagi/ were presented in the auditory (A), visual (V) and audiovisual modality. In the AV stimuli, the sound and the video were either congruent or incongruent. Audiovisual integration can be examined by comparing ERPs evoked by the bimodal stimuli with the sum of the neural activity of the unimodal stimuli. This additive model assumes that the neural activity evoked by AV stimuli is equal to the sum of activities of the auditory (A) and visual (V) activity and its associated audiovisual interactions (AV=A+V+[A×V interactions], e.g., Giard & Peronnet, 1999). If the unimodal signals are processed independently, then the bimodal response equals the sum of unimodal responses (AV=A+V). If, however, the bimodal response differs (supra-additive or sub-additive) from the sum of the two unimodal responses, this is attributed to the interaction between the two modalities (Besle et al., 2004; Giard & Peronnet, 1999; Klucharev, Mottonen, & Sams, 2003; Molholm et al., 2002; Stekelenburg & Vroomen, 2007; Teder-Sälejärvi, Di Russo, McDonald, & Hillyard, 2005; Vroomen & Stekelenburg, 2010). Studies on AV speech integration using this additive model approach have demonstrated that the auditory N1 and P2 are attenuated (Arnal, Morillon, Kell, & Giraud, 2009; Besle et al., 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007, 2012a) and sped-up (Arnal et al., 2009; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005) when synchronized with lip-read speech.

The N1 and P2 are differentially sensitive to integration of audiovisual informational content on non-phonetic and phonetic levels. For instance, the visually induced suppression of the N1 is larger when the visual (speech) input is spatially aligned with the sound than when the AV event is spatially incongruent, whereas the auditory P2 is equally modulated by spatially congruent and incongruent visual information (Stekelenburg & Vroomen, 2012a). Moreover, the N1 is suppressed whenever the visual signal reliably predicts onset of the AV event, irrespective of whether the stimuli are speech, ecologically valid stimuli (Stekelenburg & Vroomen, 2007) or artificial AV events (Vroomen & Stekelenburg, 2010) whereas visually induced suppression of the P2 seems independent of the predictability of the leading visual input (Vroomen & Stekelenburg, 2010).

More specific for AV speech are studies that show that lip-read information with emotional content (i.e., displaying fear or anger) modulates N1 and P2 amplitudes when compared to emotionally neutral AV stimuli (Jessen & Kotz, 2011) whereas phonetic AV congruency (e.g., hearing /ba/ while lip-reading /fu/) affected the P2 component but not the interaction effects at the N1 (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007).

The latter finding indicates that AV integration at the P2 may reflect a processing stage at which the unimodal signals are bound together on a phonetic level. If so, we expected to observe lip-read induced P2-suppression only for listeners in speech mode, but not for listeners in non-speech mode. In contrast, we hypothesized that the auditory N1 would be modulated alike for listeners in speech- and non-speech mode because it is sensitive to low-level features like anticipatory visual motion and spatial AV congruency (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007, 2012a; Vroomen & Stekelenburg, 2010), but likely does not depend on the speech mode of the listeners.

Our design also allowed us to disentangle AV phonetic binding from processes dedicated to congruency detection because the phonetic incongruency in the AV pseudo-words /tabi/ and /tagi/ only became apparent 270 ms after sound onset. Phonetic incongruency was thus absent at the time that the initial N1 and P2 were generated (as measured from sound onset). Effects of phonetic incongruency thus were expected to occur late (i.e., >270 ms), and we expected these to observe only for listeners in the speech mode.

2. Material and methods

2.1. Participants

Twenty-eight first-year students (20 females, all native speakers of Dutch) from Tilburg University participated in return for course credits. Half of them were randomly assigned to the speech mode group, the other half to the non-speech mode group. Participants' age ranged in between 18 and 26 years (mean=21) and did not differ across groups; $t_{26}=1.17$, $p=.252$, $d=.459$. All reported normal hearing, had normal/corrected to normal vision, and gave their written informed consent prior to testing. All testing was conducted in accordance with the Declaration of Helsinki.

2.2. Stimuli

Stimulus material was based on audiovisual recordings (recorded at a rate of 25 frames per second) of a Dutch male speaker pronouncing the Dutch pseudo-words /tabi/ and /tagi/. The audio was converted into sine-wave speech via a script provided by C. Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS) in the Praat software (Boersma & Weenink, 2005). The audio files were 627 ms (/tabi/) and 706 ms (/tagi/) in duration. Vowel (/a/) to consonant (/b/ vs /g/) transitions started at 270 and 300 ms, respectively and onsets of the critical /b/ and /g/ consonants were 372 ms (/b/) and 428 ms (/g/) (see Fig. 1). Videos displayed the speaker's face from shoulders upward. Eight different stimuli were created from these recordings; auditory-only SWS /tabi/ and /tagi/ (i.e. Ab and Ag), visual-only videos (Vb and Vg), AV congruent SWS (AbVb and AgVg), and AV incongruent SWS (AbVg and AgVb). All AV stimuli looked and sounded naturally timed.

2.3. Procedure and design

The experiment took place in a dimly lit and sound-attenuated booth in which participants sat at approximately 70 cm from a 17-inch CRT-monitor. The audio was delivered at ~65 dBa (ear level) via a computer speaker placed directly below the monitor. Size of the videos subtended 14° horizontal and 12° vertical visual angle.

The experiment started with a short training session. Participants in the speech mode learned to perceive the SWS stimuli as speech by alternating the original audio recordings with the corresponding SWS tokens (12 presentations of each stimulus). Listeners in non-speech mode only heard the SWS sounds (also 12 times for each sound) while under the impression they were hearing two different arbitrary computer sounds. After training, participants in non-speech mode were asked whether the SWS resembled any type of sound they were familiar with and none of the participants reported to have heard the sounds as speech-like. Next, ERPs were recorded during six ~10 min-blocks with short breaks in between. One experimental

Please cite this article as: Baart, M., et al. Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* (2013), http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011

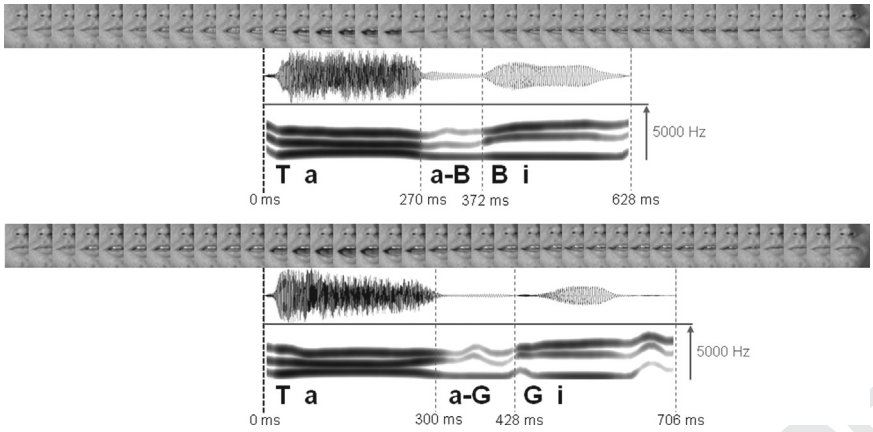

Fig. 1. Overview of the audiovisual /tabi/ (upper panels) and /tagi/ (lower panels) stimuli. Timing of the SWS is indicated in the oscillograms (upper sections) and spectrograms displaying the formants F1, F2 and F3 (lower sections). Please note that the lip-read stimuli comprised the entire face; the cropped bitmaps are created for displaying purpose only.

block comprised 112 trials (of which 96 were experimental trials and 16 catch trials) delivered in random order. The inter-trial interval was 1–2 s. Half of the experimental trials per block were unimodal and the other half were audiovisual. Half of the unimodal trials were auditory-only (i.e. 12 Ab and 12 Ag trials) and the other half were visual-only trials (12 Vb and 12 Vg trials). Of the audiovisual trials, 24 were congruent (12 AbVb and 12 AgVg trials) and 24 were incongruent (12 AbVg and 12 AgVb trials). Participants were engaged in an unrelated visual detection task: They were instructed to attend the monitor and press a button whenever an occasional small white square appeared for 120 ms on the upper-lip of the speaker (or on the black screen during auditory-only trials). There were 16 of these catch-trials in each block, 2 in each of the 8 different stimuli. After the experiment, participants in non-speech mode were specifically asked whether they had noticed that the sounds were derived from speech, which was never the case.

2.4. EEG recording and analyses

The electroencephalogram (EEG) was recorded at a sampling rate of 512 Hz from 64 locations corresponding to the extended International 10–20 system. Electrodes were active Ag–AgCl electrodes (BioSemi, Amsterdam, the Netherlands) with an output impedance of less than 1 Ω (Metting van Rijn, Peper, & Grimbergen, 1990), which were mounted in an elastic cap. Two additional electrodes were used to reference the EEG on-line; the active Common Mode Sense electrode (CMS) and ground (Driven Right Leg passive electrode; DRL). Four additional electrodes (2 on the orbital ridge above and below the right eye and 2 on the lateral junctions of both eyes) recorded the vertical- and horizontal electrooculogram (EOG) and two electrodes were placed on the left and right mastoids. The EEG was referenced offline to an average of these mastoids and band-pass filtered (Butterworth Zero Phase Filter, 0.5–30 Hz, 24 dB/octave). The 50 Hz interference was removed by a 50 Hz notch filter. ERPs were time-locked to auditory onset and the raw data were segmented into epochs of 1100 ms, including a 100 ms pre-stimulus baseline. After EOG correction (Gratton, Coles, & Donchin, 1983), we excluded data obtained at electrodes P9 and P10 due to excessive noise in almost all participants. Next, we excluded all epochs in which the amplitude difference at any channel exceeded 120 μV (in the entire 1100 ms epoch), which led to rejection of 13.4% of the data (SD = 19.3%), which corresponds to ~90 trials out of the total of 672.

The ERPs of the experimental trials were averaged per modality (A, V and AV) for the speech- and non-speech modes separately. We subtracted the visual-only ERPs from the audiovisual ERPs (i.e., $AV_{congruent}$ – V and $AV_{incongruent}$ – V, henceforth; AVC – V and AVI – V), compared the residual with the A-Only ERPs and interpreted the difference as an effect of audiovisual interaction (e.g. Besle et al., 2004; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007, 2012a; Vroomen & Stekelenburg, 2010) which aligns with the additive model in which, as mentioned, AV interactions are explained by the difference between AV activity and the sum of the unimodal activities. The peak amplitudes of the auditory N1 and P2 were scored in, and extracted from, time-windows of 50–150 ms (N1) and 150–250 ms (P2) after sound onset respectively. To investigate effects of audiovisual stimulus congruency, we compared the ERPs between the AVC and AVI conditions. Analyses comprised analyses of variance (ANOVAs) that were corrected (Greenhouse–Geisser) whenever sphericity was violated, simple effects tests and post hoc *t*-tests that were backwards Bonferroni corrected for multiple comparisons (i.e., reported *p*-values are the observed *p*-values multiplied by the number of comparisons, which facilitates interpretation with respect to the .05 alpha threshold).

3. Results

3.1. Catch-trials

Participants were almost flawless on catch-trial detection (99% in the speech group versus 100% in the non-speech group, $t(26) = 1.13$, $p = .27$, $d = .44$), indicating that they were looking at the screen as instructed.

3.2. N1 and P2

Recorded N1 and P2 amplitudes were largest at electrode Cz, which is indicated in Fig. 2 that shows the scalp topography of both peaks. Furthermore, the distribution of activity related to the N1 was more wide-spread than for the P2. In an initial set of comparisons, we tested N1 amplitude at Cz against all other mid-central and fronto-central electrodes. Different approaches (i.e., with/without correcting for multiple comparisons and testing either the auditory N1 or the average N1 across auditory, AVC – V and AVI – V conditions) revealed 4 clusters in which the N1 at Cz was not statistically different from the other electrodes. However, since ANOVAs on these data (that included 'Modality; A-Only, AVC – V, AVI – V' and 'Group: speech mode/non-speech mode' as factors) yielded the same pattern of results as the ANOVA on Cz-only (i.e., besides a main effect of 'Electrode' that varied across analyses, no significant effects were observed), we confined our analyses to Cz only (see also Stekelenburg, Maes, van Gool, Sitskoorn, & Vroomen, 2013). From Fig. 3, it can be deduced that AV integration effects at the N1 were similar for both groups, whereas interactions at the P2 only occurred in the speech group. To test these observations, amplitude and latencies scores were subjected to 2 (Group; Speech mode vs. Non-speech mode) × 3 (Modality; A-Only, AVC – V, AVI – V) mixed-effects repeated measures ANOVAs with Group as a between-subjects factor and Modality as a within-subjects factor.

For N1 amplitude there was no main effect of Modality, $F(2,52) = 1.10$, $p = .33$, $\eta_p^2 = .04$, no main effect of Group, $F(1,26) = .08$, $p = .78$, $\eta_p^2 < .01$, and no interaction between the two factors, $F(2,52) = .25$, $p = .72$, $\eta_p^2 = .01$. For N1 latency the ANOVA showed a main effect of Modality, $F(2,52) = 7.21$, $p < .01$, $\eta_p^2 = .22$. Post-hoc analysis revealed that N1 latencies of AVC – V and AVI – V were shorter than for A-Only, namely ~6 ms, $t(27) = 2.83$, $p = .03$, $d = .36$ and ~7 ms, $I(27) = 2.97$, $p = 0.02$, $d = .40$, respectively. Congruent and incongruent N1 latency did not differ from each other, $t(27) = .43$, $p > .99$, $d = .04$. There was no Modality × Group interaction,

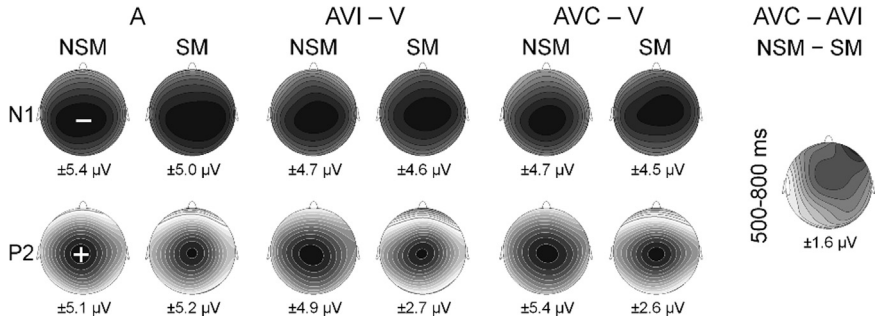

Fig. 2. Scalp topography of the N1 and P2 peaks for auditory-only (A), the audiovisual congruent – visual (AVC – V) and audiovisual incongruent – visual (AVI – V) conditions for the speech mode (SM) and non-speech mode (NSM). The rightmost topography map displays the mean difference between speech- and non-speech mode in a 500–800 ms window for the difference between audiovisual congruent- and incongruent stimuli.

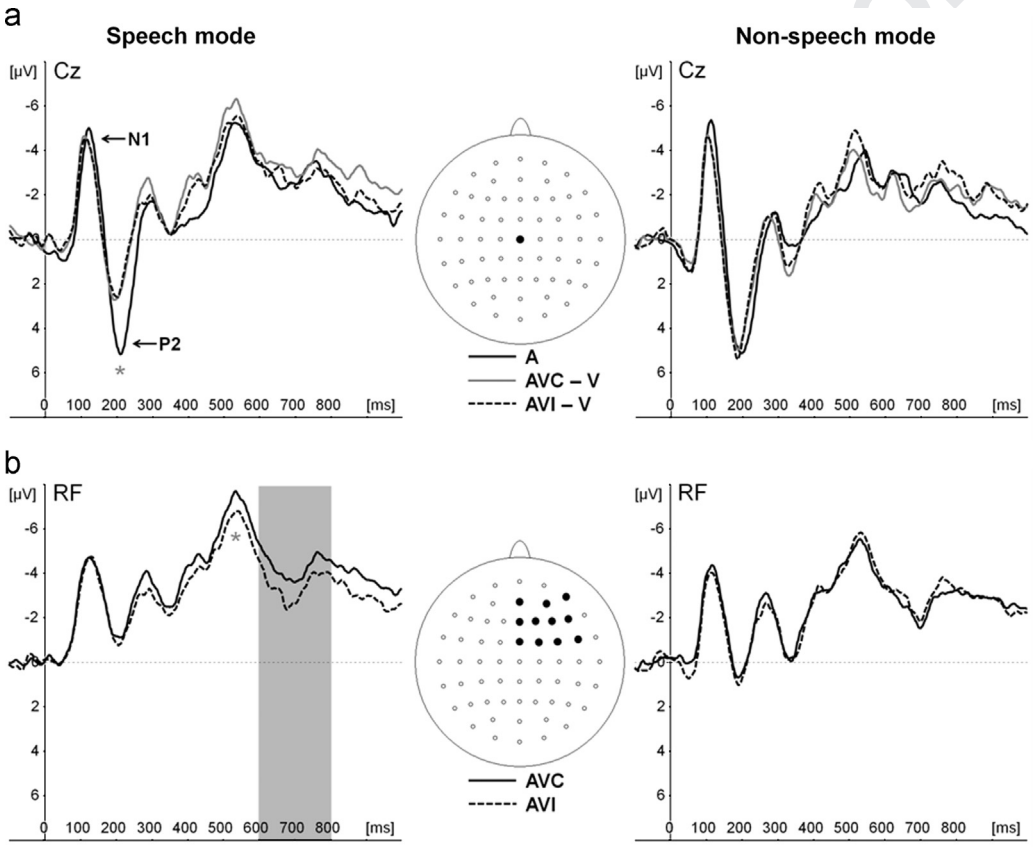

Fig. 3. (a) Averaged event-related potentials (ERPs) at Cz for auditory-only (A) and the audiovisual congruent – visual (AVC – V) and incongruent – visual (AVI – V) conditions for the speech mode (left) and non-speech mode (right). (b) Displays the ERPs averaged across right-frontal electrodes for AV congruent- (AVC) and incongruent (AVI) stimuli. The asterisks and shaded area indicate significant peak- and mean amplitude differences between speech- and non-speech mode conditions.

$F(2,52)=.63$, $p=0.50$, $\eta_p^2=.02$, but the N1 had peaked ~13 ms earlier in the non-speech group than in the speech group, $F(1,26)=5.27$, $p=.03$, $\eta_p^2=.17$. We ran an additional analysis on the A-only N1 to exclude the possibility that the difference in N1 latency between the two groups was due to differences in timing in the visual ERP. A-only N1 latency was shorter for the non-speech mode group, $t(26)=2.09$, $p<.05$, $d=0.82$.

For P2 amplitude there was no main effect of Group, $F(1,26)=1.80$, $p=.20$, $\eta_p^2=.06$. There was a main effect of Modality, $F(2,52)=10.02$, $p<.001$, $\eta_p^2=.28$as overall, the auditory P2 was larger than the AVC – V P2, $t(27)=3.68$, $p<.01$, $d=.51$ whereas the auditory P2 did not statistically differ from the AVI – V P2, $t(27)=2.48$, $p=.06$, $d=.37$, and there also was no difference between both congruent and incongruent AV – V P2 amplitudes, $t(27)=2.06$, $p=.147$, $d=.192$. Most importantly however, as depicted in Fig. 3, there was an interaction between Modality and Group, $F(2,52)=3.95$, $p=0.04$, $\eta_p^2=.13$ because P2 amplitude was reduced for listeners in speech mode, but not for listeners in non-speech mode. This was indeed confirmed by simple effect tests that revealed a main effect of Modality for the speech mode, $F(2,26)=15.43$, $p<.001$, $\eta_p^2=.54$, but not for the non-speech mode, $F(2,62)=.84$, $p=.44$, $\eta_p^2=.06$. Further post-hoc tests in the speech mode group showed that P2 amplitude was 2.9 μV smaller for AVC – V than for A, $t(13)=4.42$, $p<.01$, $d=.84$, and 2.3 μV smaller for AVI – V than for A, $t(13)=3.85$, $p<.01$, $d=.70$. AV congruent and incongruent P2 amplitudes did not differ from each other, $t(13)=1.67$, $p=.36$, $d=.20$, which was to be expected since SWS /tabi/ and /tagi/ only started to deviate at 270 ms after sound onset (i.e., at the transition of the second consonant). For P2 latency there was no main effect of Group, $F(1,26)=.53$, $p=.47$, $\eta_p^2=.02$ and no

Please cite this article as: Baart, M., et al. Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* (2013), http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011

interaction between Group and Modality, $F(2,52)=2.04$, $p=.16$, $\eta_p^2=.07$. Although there was a main effect of Modality, $F(2,52)=5.30$, $p=0.02$, $\eta_p^2=.17$, none of the post hoc t-tests reached significance, $t(27)=2.30$, $p=.09$, $d=.46$ for A versus AVC – V, $t(27)=2.43$, $p=.07$, $d=.45$ for A versus AVI – V and $t(27)=.09$, $p>.99$, $d=.01$ for AVC – V vs. AVI – V.

Topographic analysis of N1 and P2 – testing differences in scalp distribution between conditions – comprised vector-normalized amplitudes (McCarthy & Wood, 1985) of the 61 electrodes. For both N1 and P2 there were no interactions with Electrode and Modality or Group (p-values >0.290) suggesting that topography of the N1 and P2 did not differ between groups and stimulus conditions.

3.3. Congruency effects

As an exploratory analysis to investigate the time-course of stimulus congruency we subtracted AVI from AVC separately per group. Subsequently, the difference waves were subjected to point-by-point t-tests at each electrode between speech and non-speech mode groups. Using a procedure to minimize type I errors (Guthrie & Buchwald, 1991), differences in audiovisual interactions were considered significant when at least 12 consecutive points (i.e., 24 ms when the signal was re-sampled at 500 Hz[1]) were significantly different between groups. Significant differences in congruency effects between groups were found at the right frontal electrodes in a window of approximately 500–800 ms from sound onset, which corresponds to 200–530 ms after auditory onset of the /a/ to /b/ transition and to 200–500 ms after onset of the transition from /a/ to /g/. As depicted in Fig. 3b, in that window, the averaged amplitudes for AVC and AVI stimuli across right-frontal electrodes (AFz, AF4, AF8, Fz, F2, F4, F6, FCz, FC2, FC4 and FC6) showed effects of stimulus congruency only for listeners that heard the SWS sounds as speech. A 2 (Group; Speech mode vs. Non-speech mode) × 2 (Congruency; AVC vs. AVI) × 11 (Electrodes; AFz, AF4, AF8, Fz, F2, F4, F6, FCz, FC2, FC4 and FC6) mixed-effects repeated measures ANOVA on the negative peak at 500 ms (individually scored within a window of 400–700 ms post stimulus onset) showed no main effects of Group, $F(1,26)=1.81$, $p=.19$, $\eta_p^2=.07$ or Congruency, $F(1,26)=.18$, $p=.67$, $\eta_p^2=.01$. There was a significant interaction between Congruency and Group, $F(1,26)=5.59$, $p=.03$, $\eta_p^2=.18$, and separate simple effect tests for the speech and non-speech mode yielded a significant effect of Congruency for the speech mode, $F(1,13)=8.22$, $p=.01$, $\eta_p^2=.39$, but not for the non-speech mode, $F(1,13)=1.23$, $p=.29$, $\eta_p^2=.09$.The same ANOVA on the mean activity within a 600–800 ms time-window showed no main effects of Congruency or Group, $F(1,26)=1.30$, $p=.26$, $\eta_p^2=.05$ and $F(1,26)=1.76$, $p=.20$, $\eta_p^2=.06$, respectively, but did show a significant interaction between these two variables, $F(1,26)=7.30$, $p=.01$, $\eta_p^2=.02$. Again, follow-up simple effect tests revealed a congruency effect for listeners in speech mode, $F(1,13)=6.82$, $p=.02$, $\eta_p^2=.34$, but not for listeners in non-speech mode, $F(1,13)=1.33$, $p=.27$, $\eta_p^2=.09$.

4. Discussion

The main finding was that lip-read information suppressed the auditory-evoked P2 amplitude only when listeners were in speech mode, whereas AV interactions at N1 latency were independent of the listeners' speech mode. In a number of studies showing parallel suppression of N1 and P2, no distinction could be made between AV interaction based on temporal prediction or phonetic/semantic binding (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007, 2012a; van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010). Our data do not support the notion that P2 suppression occurs by default whenever the visual input leads the auditory signal, and instead, show that AV interactions at N1 can be dissociated from interactions at P2. Since temporal prediction of sound onset is identical for both speech modes, AV interactions at the P2 are unlikely to reflect the neural consequences of temporal prediction. The question then remains whether AV interactions at the P2 reflect phonetic binding or processes dedicated to detect AV congruency. As mentioned in Section 1, in previous MMN studies (e.g., Colin et al., 2004, 2002; Kislyuk et al., 2008; Stekelenburg & Vroomen, 2012b) both processes were intertwined because of the used experimental design. In the current study, this issue was resolved by changing the phonetic congruency in the stimuli well after generation of P2. We therefore obtained rather direct evidence that the suppression of the audiovisual P2 reflects AV binding, beyond the level of AV phonetic congruency detection. To be more precise, it appears that the auditory P2 is modulated by lip-read input, only when both unimodal inputs are attributed to the same distal speech event. However, the P2 is not an unitary response and there are many active neural generators within a 150–250 ms window (Crowley & Colrain, 2004). Moreover, visually-induced P2 suppressions are certainly not specific to speech (e.g., Vroomen & Stekelenburg, 2010) so we certainly would not characterize the P2 as a speech-specific component, but it does appear to be modulated by AV integration that is specific to inputs originating from the same phonetic source.

As anticipated, lip-read information induced modulations of the auditory N1 for listeners in speech- and non-speech mode, indicating that the N1 latency shift is not affected by whether the auditory input was interpreted as speech or not. This fits previous reports (e.g., Stekelenburg & Vroomen, 2007) in which we observed a speeding up of the N1 was found. We could not replicate, however, the visually induced suppression of auditory N1 amplitude. This confirms reports showing that both N1 latency and amplitude are not modulated in parallel but that both effects can be dissociated (Arnal et al., 2009) and may reflect different processes in AV integration (van Wassenhove et al., 2005). It has been hypothesized that suppression of N1 depends on the effectiveness with which the leading visual signal can predict the temporal onset of the sound, which was high in other studies (Stekelenburg & Vroomen, 2007, 2012a) as stimuli with labial place of articulation were used or stimuli like the video of a hand clap that allow precise timing of sound onset. In the current study, the alveolar place of articulation may have been less effective in the temporal prediction of sound onset and consequently had less effect on N1 amplitude.

As expected, we only observed effects of AV phonetic congruency when the SWS sounds were perceived as speech. Interestingly, these effects became significant at around 200 ms after the AV incongruency had become apparent, in-line with reports that show that AV incongruency at stimulus onset also modulates the ERPs at around 200 ms (i.e., at the P2; e.g., Klucharev et al., 2003; Stekelenburg & Vroomen, 2007). More specifically, we observed a congruency effect at the negative ERP peak at ~500 ms and the area in between 600 and 800 ms (~300–500 ms after onset of AV congruency) in which ERPs were more negative for congruent – than incongruent stimuli. It is most likely that, in the AbVg stimuli, auditory /b/ preceded lip-read /g/ by > 30 ms (see Fig. 1). Although this temporal asynchrony may have possibly been detected (van Wassenhove, Grant, & Poeppel, 2007), it is highly unlikely that our ERP differences can be explained by group differences in asynchrony detection because

[1] Please note that only the point-by-point t-tests were conducted on the re-sampled data (see also Stekelenburg & Vroomen, 2007, 2012a, 2012b; Vroomen & Stekelenburg, 2010). All other analyses were conducted on the original data sampled at 512 Hz.

Please cite this article as: Baart, M., et al. Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* (2013), http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011

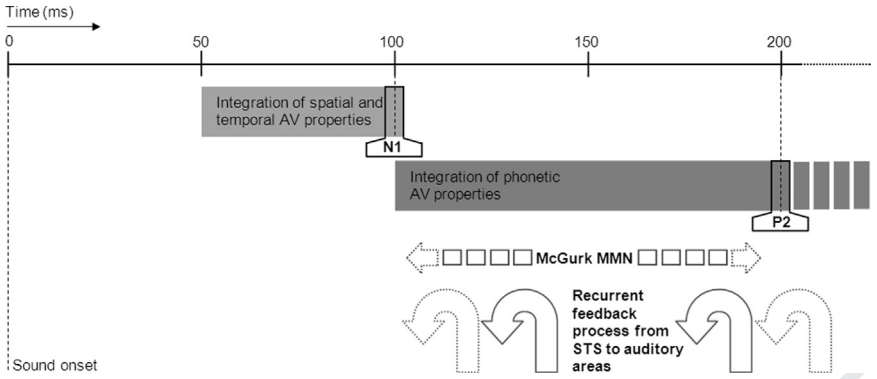

Fig. 4. Proposed time-course of AV speech integration in a 200 ms time-window from sound onset.

participants in speech- and non-speech mode are equally sensitive to AV asynchronies (Vroomen & Stekelenburg, 2011).

Speech mode had mixed effects on the neural correlates (at N1 and P2) of audiovisual speech integration, supporting the hypothesis that AV speech may be integrated by both speech-specific and more general multisensory integrative mechanism (e.g., Eskelund et al., 2011; Stekelenburg & Vroomen, 2012a; van Wassenhove et al., 2007; Vroomen & Stekelenburg, 2011). The finding that expectations about the nature of auditory stimuli strongly affects audiovisual speech integration at the P2 component is in accordance with behavioral studies and an EEG study using SWS stimuli, where McGurk-like effects (Eskelund et al., 2011; Tuomainen et al., 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011) and the McGurk MMN (Stekelenburg & Vroomen, 2012b) were elicited only when listeners were in speech mode. By contrast, speech mode had no effect on the early interactions at the N1 component. This absence of speech specificity for AV speech integration may relate to the study of Eskelund et al. (2011) who observed that lip-read information induced an auditory detection benefit for SWS, irrespective of whether the listeners were in speech- or non-speech mode. Furthermore, Vroomen and Stekelenburg (2011) found that sensitivity to temporal asynchrony between the SWS stimuli and the lip-read videos was alike for listeners that heard SWS as speech or non-speech. These behavioral findings support the hypothesis that AV speech perception is constituted through a multi-staged process (e.g. Schwartz, Berthommier, & Savariaux, 2004; Soto-Faraco & Alsius, 2009). An fMRI study (Lee & Noppeney, 2011) on audiovisual SWS speech perception support this hypothesis as it specified that the left anterior mid-STS depended on higher-order linguistic information whereas bilateral posterior and left mid-STS integrated audiovisual speech inputs on the basis of physical factors. According to this view, different features of the AV stimulus are integrated at different stages in the processing chain, which also aligns with current findings. Given the above, we propose the following time-course that underlies the different stages in AV speech integration (see also Fig. 4). At a first stage, general (i.e., spatial and temporal) AV properties are integrated. This process starts ~50 ms after sound onset (e.g., Stekelenburg & Vroomen, 2012a; Talsma, Doty, & Woldorff, 2007; Talsma & Woldorff, 2005). After the auditory N1, AV phonetic binding is realized (current study) either before, or at, the same time that a recurrent feedback process from the superior temporal sulcus (i.e., STS) to auditory areas regarding stimulus congruency is initiated (Arnal et al., 2009). The (partial) outcome(s) of this feedback process may be observed in McGurk MMN-response (e.g., Colin et al., 2002; Sams et al., 1991, of relevance here is that the first part of the MMN is argued to reflect a release-from-refractoriness of N1 neurons, see Näätänen, Kujala, & Winkler, 2011), at the P2 (Arnal et al., 2009; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007) and later during processing (current study; Arnal et al., 2009; Magnée, de Gelder, van Engeland, & Kemner, 2008).

Besides differences in AV integration as function of speech mode we also found differences in auditory-only processing. Across modality, auditory N1 peaked earlier in the non-speech mode than in the speech-mode. Although it is known that the N1 peaks earlier after hearing tones than after hearing vowels (Tiitinen, Sivonen, Alku, Virtanen, & Näätänen, 1999), our data suggest that the processes involved in speech perception delay the N1 as compared to non-speech processes under the exact same acoustical conditions and suggest more computational demands for processing speech. These differences between the latency in neural processing of SWS in speech and non-speech mode may be the consequence of different neural networks involved in processing speech and non-speech sounds. For instance, speech sounds may elicit stronger activation of long-term memory traces than unfamiliar sounds (Jaramillo et al. 2001). Furthermore, it is well-established that some areas within the STS are explicitly involved in processing phonetic acoustic information (see Hickok & Poeppel, 2007 for a review) as activation in STS is higher for SWS speech stimuli than for sine-wave non-speech tokens with the same spectral-temporal acoustic properties (Benson, Richardson, Whalen, & Lai, 2006; Benson et al., 2001). It also appears that the (posterior) STS and supramarginal gyrus are critical for decoding the phonetic information in SWS because there is significantly less activity in these areas when listeners are in non-speech mode (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006).

To conclude, we examined the neural correlates of AV speech-versus non-speech processing and controlled the acoustical variability by using SWS. The results unequivocally indicate that neural activity underlying the P2 reflects (amongst other processes) binding of congruency-independent phonetic features in AV speech. Furthermore, the current data support the view that AV speech integration is a multistage process. Finally, our observation that the N1 peaked later when SWS was perceived as speech (as opposed to non-speech), provides further support for the notion that speech processing (partially) relies on different mechanisms than non-speech processing.

References

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, *29*(43), 13445–-13453, http://dx.doi.org/10.1523/JNEUROSCI.3194-09.2009.

Benson, R. R., Richardson, M., Whalen, D. H., & Lai, S. (2006). Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. *Neuroimage*, *31*(1), 342–353, http://dx.doi.org/10.1016/j.neuroimage.2005.11.029.

Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., et al. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and Language*, *78*, 364–396, http://dx.doi.org/10.1006/brln.2001.2484.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). *Bimodal speech: Early suppressive visual effects in human auditory cortex*. *European Journalof Neuroscience*, *20*, http://dx.doi.org/10.1111/j.1460-9568.2004.03670.x.

Boersma, P., & Weenink, K. (2005). Praat: Doing phonetics by computer. Retrieved from ⟨http://www.fon.hum.uva.nl/praat⟩.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436, http://dx.doi.org/10.1371/journal.pcbi.1000436.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology*, *115*(9), 1989–2000, http://dx.doi.org/10.1016/j.clinph.2004.03.027.

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. *Clinical Neurophysiology*, *113*(4), 495–506, http://dx.doi.org/10.1016/S1388-2457(02)00024-X.

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology*, *115*(4), 732–744, http://dx.doi.org/10.1016/j.clinph.2003.11.021.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage*, *24*(1), 21–33, http://dx.doi.org/10.1016/j.neuroimage.2004.09.039.

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research*, *208*(3), 447–457, http://dx.doi.org/10.1007/s00221-010-2495-9.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Early auditory–visual interactions in human cortex during nonredundant target identification. *Cognitive Brain Research*, *14*, 20–30, http://dx.doi.org/10.1016/S0926-6410(02)00058-7.

Giard, M. H., & Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490, http://dx.doi.org/10.1162/089892999563544.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484, http://dx.doi.org/10.1016/0013-4694(83)90135-9.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*(2), 240–244, http://dx.doi.org/10.1111/j.1469-8986.1991.tb00417.x.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402, http://dx.doi.org/10.1038/nrn2113.

Jaramillo, M., Ilvonen, T., Kujala, T., Alku, P., Tervaniemi, M., & Alho, K. (2001). Are different kinds of acoustic features processed differently for speech and non-speech sounds? *Cognitive Brain Research*, *12*, 459–466, http://dx.doi.org/10.1016/S0926-6410(01)00081-7.

Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage*, *58*(2), 665–674, http://dx.doi.org/10.1016/j.neuroimage.2011.06.035.

Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience*, *20*, 1–10, http://dx.doi.org/10.1162/jocn.2008.20152.

Klucharev, V., Mottonen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65–75, http://dx.doi.org/10.1016/j.cogbrainres.2003.09.004.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 829–840, http://dx.doi.org/10.1037/0096-1523.17.3.829.

Lee, H., & Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *Journal of Neuroscience*, *31*(31), 11338–11350, http://dx.doi.org/10.1523/Jneurosci.6510-10.2011.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, *243*(4890), 489–494, http://dx.doi.org/10.1126/science.2643163.

Magnée, M. J., de Gelder, B., van Engeland, H., & Kemner, C. (2008). Audiovisual speech integration in pervasive developmental disorder: evidence from event-related potentials. *Journal of Child Psychology and Psychiatry*, *49*(9), 995–1000, http://dx.doi.org/10.1111/j.1469-7610.2008.01902.x.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: The MIT Press.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology*, *62*(3), 203–208, http://dx.doi.org/10.1016/0168-5597(85)90015-2.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748, http://dx.doi.org/10.1038/264746a0.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research*, *14*(1), 115–128, http://dx.doi.org/10.1016/S0926-6410(02)00066-6.

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage*, *30*(2), 563–569, http://dx.doi.org/10.1016/j.neuroimage.2005.10.002.

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect in evoked potential reinterpreted. *Acta Psychologica*, *42*, 313–329, http://dx.doi.org/10.1016/0001-6918(78)90006-9.

Näätänen, R., Kujala, T., & Winkler, I. (2011). Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. *Psychophysiology*, *48*(1), 4–22, http://dx.doi.org/10.1111/j.1469-8986.2010.01114.x.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–949, http://dx.doi.org/10.1126/science.7233191.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110, http://dx.doi.org/10.1037/0033-2909.92.1.81.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (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, http://dx.doi.org/10.1016/j.neuropsychologia.2006.03.036.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–145, http://dx.doi.org/10.1016/0304-3940(91)90914-F.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio–visual interactions in speech identification. *Cognition*, *93*(2), B69–78, http://dx.doi.org/10.1016/j.cognition.2004.01.006.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580–587, http://dx.doi.org/10.1037/a0013483.

Stekelenburg, J. J., Maes, J. P., van Gool, A. R., Sitskoorn, M., & Vroomen, J. (2013). Deficient multisensory integration in schizophrenia: An event-related potential study. *Schizophrenia Research*, *147*, 253–261, http://dx.doi.org/10.1016/j.schres.2013.04.038.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973, http://dx.doi.org/10.1162/jocn.2007.19.12.1964.

Stekelenburg, J. J., & Vroomen, J. (2012a). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, 6, http://dx.doi.org/10.3389/fnint.2012.00026.

Stekelenburg, J. J., & Vroomen, J. (2012b). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia*, *50*, 1425-–1431, http://dx.doi.org/10.1016/j.neuropsychologia.2012.02.027.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215, http://dx.doi.org/10.1121/1.1907309.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*(3), 679–690, http://dx.doi.org/10.1093/cercor/bhk016.

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–1114, http://dx.doi.org/10.1162/0898929054475172.

Teder-Sälejärvi, W. A., Di Russo, F., McDonald, J. J., & Hillyard, S. A. (2005). Effects of spatial congruity on audio–visual multimodal integration. *Journal of Cognitive Neuroscience*, *17*(9), 1396–1409, http://dx.doi.org/10.1162/0898929054985383.

Tiitinen, H., Sivonen, P., Alku, P., Virtanen, J., & Näätänen, R. (1999). Electromagnetic recordings reveal latency differences in speech and tone processing in humans. *Cognitive Brain Research*, *8*(3), 355–363, http://dx.doi.org/10.1016/S0926-6410(99)00028-2.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*(1), B13–22, http://dx.doi.org/10.1016/j.cognition.2004.10.004.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186, http://dx.doi.org/10.1073/pnas.0408949102.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia*, *45*(3), 598–607, http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.001.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*(2), 254–259, http://dx.doi.org/10.1016/j.cognition.2008.10.015.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596, http://dx.doi.org/10.1162/jocn.2009.21308.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*(1), 75–83, http://dx.doi.org/10.1016/j.cognition.2010.10.002.

Please cite this article as: Baart, M., et al. Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* (2013), http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.011