Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events

Jeroen J. Stekelenburg and Jean Vroomen

Abstract

■ A question that has emerged over recent years is whether audiovisual (AV) speech perception is a special case of multisensory perception. Electrophysiological (ERP) studies have found that auditory neural activity (N1 component of the ERP) induced by speech is suppressed and speeded up when a speech sound is accompanied by concordant lip movements. In Experiment 1, we show that this AV interaction is not speech-specific. Ecologically valid nonspeech AV events (actions performed by an actor such as handclapping) were associated with a similar speeding-up and suppression of auditory N1 amplitude as AV speech (syllables). Experiment 2 demonstrated that these AV interactions were not influenced by whether A and V were congruent or incongruent. In Experiment 3 we show that the AV interaction on N1 was absent when there was no anticipatory visual motion, indicating that the AV interaction only occurred when visual anticipatory motion preceded the sound. These results demonstrate that the visually induced speeding-up and suppression of auditory N1 amplitude reflect multisensory integrative mechanisms of AV events that crucially depend on whether vision predicts when the sound occurs. ■

INTRODUCTION

Hearing and seeing someone speak evokes a chain of brain responses that has been of considerable interest to psychologists. Once visual and auditory signals reach the ears and the eyes, these sense organs transmit their information to dedicated sensory-specific brain areas. At some processing stage, the auditory and visual streams are then combined into a multisensory representation, as can be demonstrated by the so-called McGurk illusion (McGurk & MacDonald, 1976), where listeners report to ''hear'' /da/ when, in fact, auditory /ba/ is synchronized to a face articulating /ga/. A key issue for any behavioral, neuroscientific, and computational account of multisensory integration is to know when and where in the brain the sensory-specific information streams merge.

Hemodynamic studies have shown that multisensory cortices (superior temporal sulcus/gyrus) (Skipper, Nusbaum, & Small, 2005; Callan et al., 2004; Calvert, Campbell, & Brammer, 2000) and ''sensory-specific'' cortices (Von Kriegstein & Giraud, 2006; Gonzalo & Büchel, 2004; Callan et al., 2003; Calvert et al., 1999) are involved in audiovisual (AV) speech integration. Because of limited temporal resolution, these neuroimaging studies cannot address critical timing issues. Electrophysiological techniques, on the other hand, with their millisecond precision, provide an excellent tool to study the time course of multisensory integration. Electroencephalography (EEG) and magnetoencephalography (MEG) studies have shown that AV speech interactions occur in the auditory cortex between 150 and 250 msec using the mismatch negativity paradigm (Colin et al., 2002; Möttönen, Krause, Tiippana, & Sams, 2002; Sams et al., 1991). Others have reported that at as early as 100 msec the auditory N1 component is attenuated (van Wassenhove, Grant, & Poeppel, 2005; Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003) and speeded up (van Wassenhove et al., 2005) when auditory speech is accompanied by concordant lipread information. The observed cortical deactivation to bimodal speech reflects facilitation of auditory processing as it is associated with behavioral facilitation, that is, faster identification of bimodal syllables than auditory-alone syllables (Besle et al., 2004; Klucharev et al., 2003). The suppression and speeding-up of auditory brain potentials may occur because lipread information precedes auditory information due to natural coarticulatory anticipation, thereby reducing signal uncertainty and lowering computational demands for auditory brain areas (Besle et al., 2004; van Wassenhove et al., 2005). However, to date, it is unknown whether auditory facilitation is based on speech-specific mechanisms or more general multisensory integrative mechanisms, because AV integration of speech has hitherto not been compared with that of nonspeech events that share critical stimulus features with AV speech (e.g., natural and dynamic information with a meaningful relationship between auditory and visual elements, and with visual information preceding auditory information because of anticipatory motion). In the current study, we therefore compared

Tilburg University, Tilburg, The Netherlands

neural correlates of AV speech (the syllables /bi/ and /fu/ as produced by a Dutch speaker) with that of natural nonspeech stimuli (clapping of the hands and tapping a spoon against a cup) using event-related brain potentials (ERPs). The nonspeech stimuli were controlled so that the visual information allowed to predict, as in visual speech, both the informational content of the sound to be heard and its onset time. To investigate multisensory integration, neural activity evoked by auditory-only (A) stimuli was compared with that of audiovisual minus visual-only stimuli (AV − V). The difference between A and AV − V can be interpreted as integration effects between the two modalities (Besle et al., 2004; Klucharev et al., 2003; Fort, Delpuech, Pernier, & Giard, 2002; Molholm et al., 2002; Giard & Peronnet, 1999).

The first experiment demonstrated that the auditory-evoked ERPs N1/P2 were speeded up and reduced in amplitude by concordant visual information for speech and nonspeech stimuli alike. Two additional experiments explored which information in the visual stimulus—the content of which sound to be heard (''what'') or the potential to predict when the sound is to occur (''when'')—induced these effects. Experiment 2 tested the ''what''-question by presenting congruent (e.g., hearing /bi/ and seeing /bi/) and incongruent (e.g., hearing /bi/ and seeing /fu/) speech and nonspeech AV stimuli. If the AV interaction reflects a mechanism by which the content of V predicts the content of A, one expects incongruent AV combinations to be different from congruent ones. In Experiment 3 we tested the ''when''-question by using natural stimuli that did not contain anticipatory visual motion (i.e., moving a saw, tearing of a sheet of paper), in which case the visual information thus did not predict when the sound was to occur. If the AV interaction reflects visual prediction of auditory sound onset, one expects no AV effect from stimuli that lack visual anticipatory information.

EXPERIMENT 1

Methods

Participants

Sixteen healthy participants (11 men, 5 women) with normal hearing and normal or corrected-to-normal vision participated after giving written informed consent. Their age ranged from 18 to 25 years with a mean age of 21 years. The study was conducted with approval of the local ethics committee of Tilburg University.

Stimuli and Procedure

The experiment took place in a dimly-lit, sound-attenuated, and electrically shielded room. Visual stimuli were presented on a 17-in. monitor positioned at eye level, 70 cm from the participant's head. The sounds came from a loudspeaker directly below the monitor. Speech stimuli were the syllables /bi/ and /fu/ pronounced by a Dutch female speaker whose entire face was visible on the screen (Figure 1). Nonspeech stimuli were two natural actions: clapping of the hands and tapping a spoon on a cup. The videos were presented at a rate of 25 frames/sec with an auditory sample rate of 44.1 kHz. The size of the video frames subtended 14° horizontal and 12° vertical visual angle. Peak intensity of the auditory stimuli was 70 dB(A). For each stimulus category, three exemplars were recorded, thus amounting to 12 unique recordings. Average duration of the video was 3 sec, including a 200-msec fade-in and fade-out, and a still image (400–1100 msec) at the start. The duration of the auditory sample was 306–325 msec for /bi/, 594–624 msec for /fu/, 292–305 msec for the spoon tapping on a cup, and 103–107 msec for the clapping hands. The time from the start of the articulatory movements until voice onset was, on average, 160 msec for /bi/ and 200 msec for /fu/. The time from the start of the movements of the arm(s) until sound onset in the nonspeech stimuli was 280 msec for the clapping hands and 320 msec for the tapping spoon.

The experimental conditions comprised audiovisual (AV), visual (V), and auditory (A) stimulus presentations. The AV condition showed the original video recording with the sound synchronized to the video; the V condition showed the same video, but without the sound track; the A condition presented the sound along with a static gray square with the same variable duration as the visual component of the AV and V conditions. Multisensory interactions were examined by comparing ERPs evoked by A stimuli with AV minus V (AV − V) ERPs. The additive model (A = AV − V) assumes that

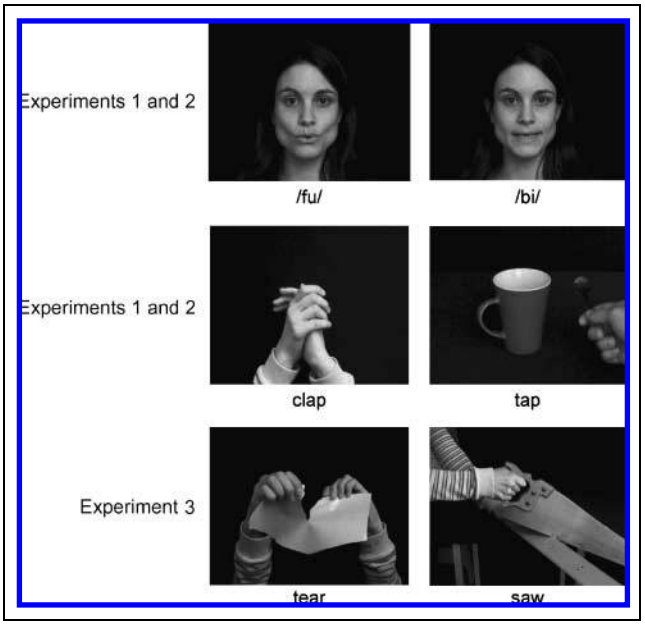

Figure 1. Stimuli used in Experiments 1 and 2 (syllables: /bi/ and /fu/, actions: tapping a spoon on a cup, handclapping) and Experiment 3 (sawing and tearing).

the neural activity evoked by AV stimuli is equal to the sum of activities of A and V if the unimodal signals are processed independently. This assumption is valid for extracellular media and is based on the law of superposition of electric fields (Barth, Goldberg, Brett, & Di, 1995). If 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. However, this additive model approach can lead to spurious interaction effects if common activity like anticipatory slow wave potentials (which continue for some time after stimulus onset) or N2 and P3 are found in all conditions, because this common activity will be present in A, but removed in the AV − V subtraction (Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002). To circumvent potential problems of late common activity, we therefore restricted our analysis to the early stimulus processing components (<300 msec). Furthermore, we added another control condition (C) to counteract spurious subtraction effects. In the C condition, the same gray square was shown as in A, but without sound. Attention paid to the stimuli (and the associated anticipatory slow wave potentials) was in C identical to the other conditions because participants were performing the same task (see below). In the additive model, ERPs of C were then subtracted from A (A − C), so that anticipatory slow waves (and visual ERP components common in A and C) were subtracted as in the AV − V comparison. AV interactions devoid of common activity could then be determined by comparing A − C with AV − V [i.e., (A − C) − (AV − V)].

For each condition (A, V, AV, and C), 96 randomized trials for each of the 12 exemplars were administered across 8 identical blocks. Testing lasted about 2 hr (including short breaks between the blocks). To ensure that participants were looking at the video during stimulus presentation, they had to detect, by keypress, the occasional occurrence of catch trials (7.7% of total number of trials). Catch trials occurred equally likely in all conditions. Catch trials contained a superimposed small white spot—either between the lips and nose for speech stimuli, or at collision site for the nonspeech stimuli—for 120 msec. The appearance of the spot varied quasi-randomly within 300 msec before or after the maximal opening of the mouth or the time of impact for nonspeech events. In the A and C conditions, the spot was presented on the gray square at about the same position and same time as in the AV and V conditions.

ERP Recording and Analysis

EEG was recorded at a sample rate of 512 Hz from 47 locations using active Ag–AgCl electrodes (BioSemi, Amsterdam, The Netherlands) mounted in an elastic cap and two mastoid electrodes. Electrodes were placed according the extended International 10-20 system. Two additional electrodes served as reference (Common Mode Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). EEG was referenced off-line to an average of left and right mastoids and band-pass filtered (0.5–30 Hz, 24 dB/octave). The raw data were segmented into epochs of 1000 msec, including a 200-msec prestimulus baseline. ERPs were time-locked to the sound onset in the AV and A conditions, and to the corresponding time stamp in the V and C conditions. After electrooculogram correction (Gratton, Coles, & Donchin, 1983), epochs with an amplitude change exceeding ±150 μV at any channel were rejected. ERPs of the non-catch trials were averaged per condition (AV, A, V, and C), separately for each speech and nonspeech stimulus. The first analysis focused on whether visual information in the AV condition suppressed and speeded up auditory-evoked responses by comparing the N1 and P2 of the audiovisual (AV − V) condition with the auditory-only (A − C) condition. Auditory N1 and P2 had a central maximum, and analyses were therefore conducted at the central electrode Cz. The N1 was scored in a window of 70–150 msec, P2 was scored in a window of 120–250 msec. Topographic analysis of N1 and P2 comprised vector-normalized amplitudes (McCarthy & Wood, 1985) of the electrodes surrounding Cz (FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2).[1] The second analysis explored the spatio-temporal dynamics of the AV interaction by conducting point-by-point two-tailed *t* tests on the (AV − V) − (A − C) difference wave at each electrode in a 1–300 msec window. Using a procedure to minimize type I errors (Guthrie & Buchwald, 1991), AV interactions were considered significant when at least 12 consecutive points (i.e., 24 msec when the signal was resampled at 500 Hz) were significantly different from zero. This analysis allowed for detection of the earliest time where AV interactions occurred.

Results of Experiment 1

Participants detected 99.5% of the catch trials, indicating that they indeed watched the video. Figure 2 shows that the amplitudes of N1 and P2 were attenuated and speeded up in the AV − V condition compared to the A − C condition for both speech and nonspeech stimuli, with larger effects on P2 for nonspeech stimuli. In the analyses, ERPs were pooled across the two syllables and actions because there were no significant differences or interactions within these categories. Latency and amplitude difference scores (AV − V) − (A − C) of speech and nonspeech stimuli at electrode Cz were submitted to a multivariate analysis of variance for repeated measures (multivariate analysis of variance [MANOVA]).[2] N1 amplitude in the AV condition was significantly reduced by 1.9 μV compared to the auditory-only condition [$F(1, 15) = 21.21$, $p < .001$]. N1 latency was speeded up by 12 msec [$F(1, 15) = 15.35$, $p < .01$], with no difference between speech and nonspeech stimuli (*F* values < 1).

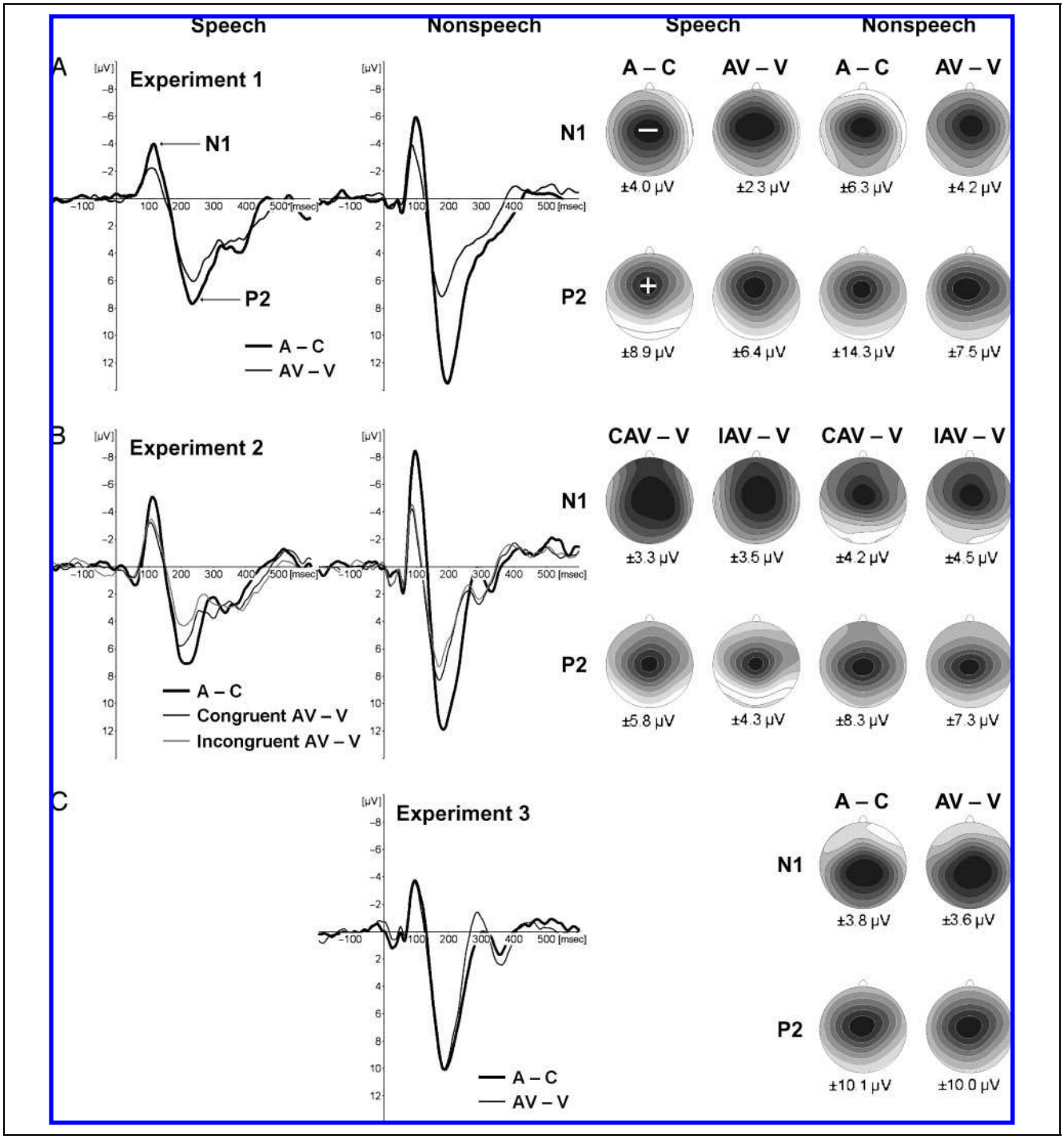

Figure 2. Event-related potentials (ERPs) at electrode Cz (left panel) and the scalp topography of auditory peak N1 and P2 (right panel). The range of the voltage maps in microvolts (µV) are displayed below each map. ERPs for speech and nonspeech were pooled across syllables and actions, respectively. (A) Experiment 1: Auditory-only minus control (A − C) and audiovisual minus visual-only (AV − V) ERPs. (B) Experiment 2: Auditory-only minus control (A − C), congruent audiovisual minus visual-only (Congruent AV − V), and incongruent audiovisual minus visual-only (Incongruent AV − V) ERPs. (C) Experiment 3: Auditory-only minus control (A − C) and audiovisual minus visual-only ERPs (AV − V) of nonspeech events containing no visual anticipatory motion.

The same analysis on the P2 revealed a greater amplitude [$F(1, 15) = 4.89$, $p < .05$] and latency reduction [$F(1, 15) = 38.33$, $p < .001$] for nonspeech stimuli than speech stimuli (speech: 1.8 µV, 2.9 msec; nonspeech: 6.5 µV, 12.8 msec). Post hoc analysis on the P2 of speech stimuli showed a significant amplitude reduction [$t(15) = 3.19$, $p < .01$], but no latency effect. Figure 2 shows that the scalp distribution of N1 and P2 in the bimodal condition (AV − V) resembled N1 and P2 in the auditory-only (A − C) condition. Topographic analysis confirmed that

for both speech and nonspeech N1 and P2, there was no interaction between electrode (FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2) and modality (AV − V vs. A − C).

The second analysis concerned the time course of AV interactions across all electrode positions (Figure 3) using pointwise *t* test. Reliable AV interactions started at about 90 msec for nonspeech stimuli and at 100 msec for speech stimuli, both lasting approximately 50 msec. For both stimulus categories, the effect was maximal at the fronto-central electrodes. These early AV interactions were followed by longer-lasting interactions starting at 160 msec for nonspeech stimuli and 180 msec for speech stimuli. These later AV interactions were confined to frontal, fronto-central, and central electrodes for speech stimuli, whereas a more widespread topography was found for nonspeech stimuli ranging from anterior–frontal to parietal regions. The timing and the location of the AV interactions corresponded to the modulation of both the auditory N1 and P2. To conclude, then, there was no hint that the speeding-up and suppression of

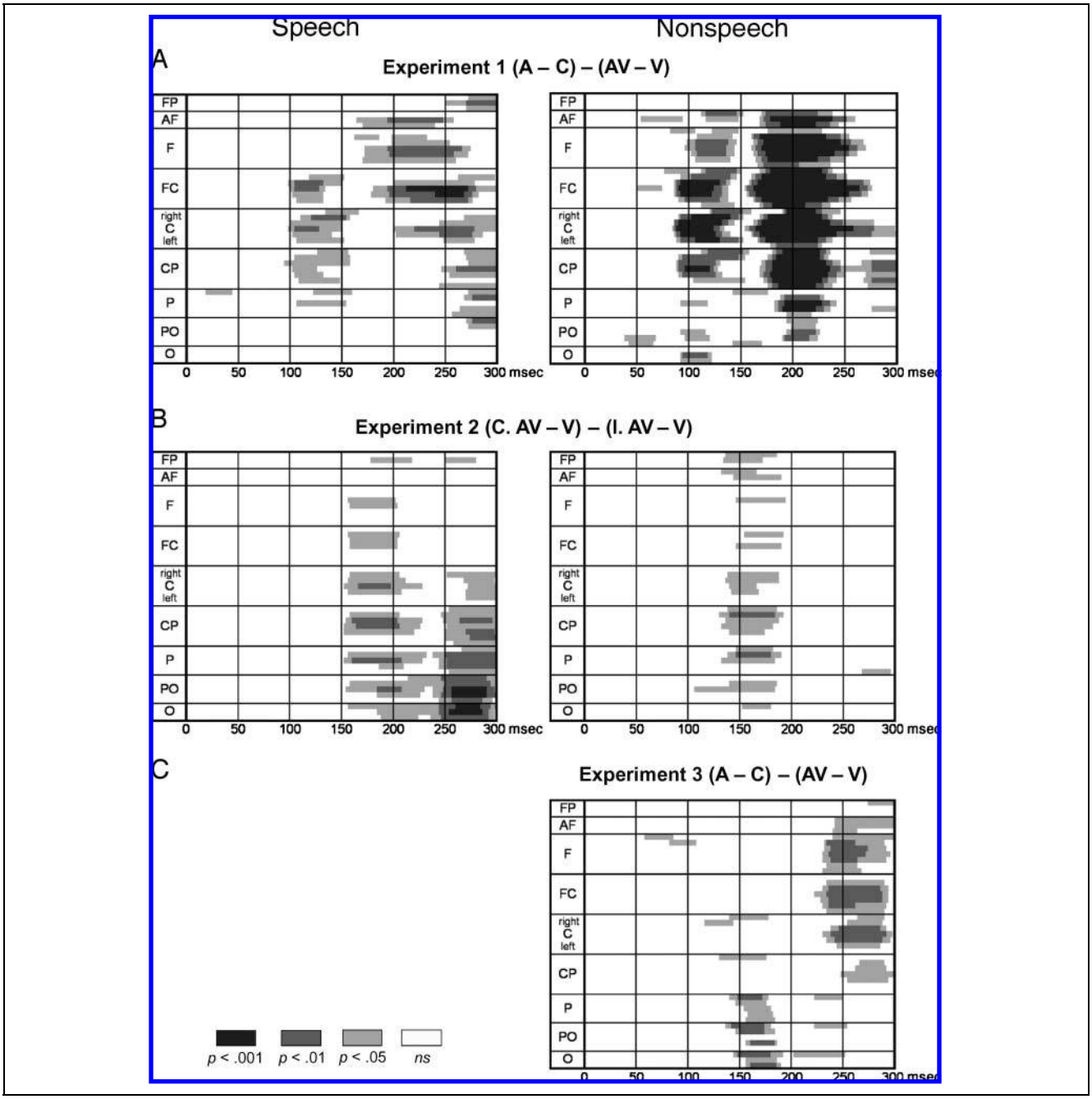

Figure 3. Time course of AV interactions using pointwise *t* tests at every electrode. (A) and (C) Experiments 1 and 3: Pointwise *t* tests on the difference wave (AV − V) − (A − C) evaluating interactions between A and V. (B) Experiment 2: Pointwise *t* tests on the difference wave between congruent and incongruent audiovisual ERPs (C. AV − V) − (I. AV − V) examining congruency effects between A and V.

auditory potentials only occurred for speech stimuli, as AV interactions of nonspeech stimuli started somewhat earlier, were stronger, and more widespread over the scalp than those of speech stimuli.

EXPERIMENT 2

In Experiment 2 we varied whether the sound was congruent or incongruent with the content of the video so as to determine whether a match in informational content was crucial for the AV interactions to occur. Seventeen new healthy women and two men (17–24 years, mean = 19 years) participated. Stimulus materials, procedure, number of stimuli per condition, task, and recording were identical to those in Experiment 1, except that incongruent AV pairings were added (auditory /fu/ combined with visual /bi/, auditory /bi/ combined with visual /fu/, auditory handclapping combined with visual tapping of a spoon, and auditory tapping of a spoon combined with visual clapping of hands). The onset of the sounds of the incongruent stimuli was synchronized to the onset of the sound in the original recordings so that the video accurately predicted sound onset.

Results of Experiment 2

Participants detected 99.7% of the catch trials. Latency and amplitude difference scores (AV − V) − (A − C) at electrode Cz were computed for congruent and incongruent AV stimuli and submitted to a MANOVA with category (speech vs. nonspeech) and congruency (congruent vs. incongruent) as factors. Addition of the visual signal significantly reduced auditory N1 amplitude with 2.9 μV [$F(1, 16) = 63.06$, $p < .001$], and this reduction was greater for nonspeech stimuli (4.2 μV) than for speech stimuli (1.7 μV) [$F(1, 16) = 13.73$, $p < .01$]. Separate test showed that the amplitude reduction in speech and nonspeech stimuli was both significantly bigger than zero [speech: $F(1, 16) = 20.19$, $p < .01$; nonspeech: $F(1, 16) = 49.34$, $p < .001$]. There was no effect of congruency on the attenuation of N1 amplitude, nor was there an interaction between category and congruency (Figure 2). Peak latency of AV − V N1 was 7 msec shortened compared to A − C N1 [$F(1, 16) = 44.23$, $p < .001$], with no difference between speech and nonspeech stimuli ($F < 1$). Shortening of N1 peak latency to incongruent pairs was not significantly different from congruent pairings. Although there was Category × Congruency interaction for N1 latency [$F(1, 16) = 5.58$, $p < .05$], simple-effect tests showed that shortening of N1 latency was significant for each of the four AV stimuli (t values > 2.48). P2 amplitude in the AV condition was reduced by 2.9 μV compared to A − C P2 [$F(1, 16) = 38.68$, $p < .001$]. P2 amplitude reduction did not differ between speech and nonspeech stimuli, but was larger for incongruent pairings (3.4 μV) than for congruent ones (2.3 μV) [$F(1, 16) = 15.62$, $p < .01$]. There was a main effect of shortening of P2 latency of 10 msec [$F(1, 16) = 11.79$, $p < .01$]. As observed in Experiment 1, latency facilitation of P2 was greater for nonspeech (16 msec) than for speech stimuli (3 msec) [$F(1, 16) = 5.05$, $p < .01$], but did not differ between congruent and incongruent pairings. Post hoc analysis showed no shortening of P2 latency in speech stimuli ($F < 1$). For both P2 amplitude and latency scores, there were no Category × Congruency interactions. Topographic analysis of N1 and P2 amplitudes revealed that the scalp distribution for congruent and incongruent AV pairings did not differ between speech and nonspeech stimuli.

Pointwise t tests at each electrode on the difference wave between AV − V ERPs to congruent and incongruent AV stimuli (Congruent AV − V) − (Incongruent AV − V) showed that congruency effects did not take place before the onset of auditory P2 (Figure 3). For speech stimuli, the earliest congruency effect started around 150 msec and lasted until 230 msec. Congruency in nonspeech stimuli affected the ERP from 140 to 190 msec. Both epochs correspond to the auditory P2. Figure 3 also shows that, for speech stimuli, the effect of congruency was prolonged compared to nonspeech stimuli in a time window of 250–300 msec at occipito-parietal electrodes. The results of Experiment 2 thus demonstrated that the early AV interactions occurring at around 100 msec were unlikely to be caused by the informational content of the video, as both congruent and incongruent AV pairings showed a speeding-up and suppression of N1.

EXPERIMENT 3

To further explore the basis of the AV interaction, new stimuli were created that did not contain visual anticipatory motion. The visual information did not, in this case, allow to predict when the sound was to occur. If temporal prediction of the sound by the visual information is crucial, then the robust N1 effect observed before should disappear with these stimuli.

Sixteen new healthy women and three men (17–27 years, mean = 21 years) participated in Experiment 3. Stimuli were clips of two different actions performed by the same actor as used before. In the first clip, two hands held a paper sheet which was subsequently torn apart. In the second clip, the actor held a saw resting on a plastic plank and, subsequently, made one forward stroke. Of each action, three different exemplars were selected resulting in six unique video clips. Note that the onsets of the visual and auditory information were synchronized as before, but unlike Experiments 1 and 2, there was no anticipatory visual motion. All other experimental details were identical to those in Experiments 1 and 2.

Results of Experiment 3

Participants detected 99% of the catch trials. Latency and amplitude of N1 and P2 at electrode Cz, pooled across

the two actions, of the A − C condition were compared to those of the AV − V condition. Unlike in Experiments 1 and 2, AV − V N1 and P2 amplitude and latency did not differ from A − C N1 and P2 (t values < 1.25, p values > .23) (Figure 2). Scalp distributions of N1 [$F(8, 8) = 1.68$, $p = .24$] and P2 ($F < 1$) also did not differ between A − C and AV − V. Pointwise t-test analysis confirmed that at N1 latency there was no AV interaction (Figure 3). AV interactions started at approximately 150 msec at the posterior sites. Late interactions were found at the fronto-central N2.

Discussion

In line with previous studies on AV speech perception, we found that the auditory-evoked N1 and P2 potentials were smaller (van Wassenhove et al., 2005; Besle et al., 2004; Klucharev et al., 2003) and occurred earlier (van Wassenhove et al., 2005) when visual information accompanied the sound. The novel finding is that these effects were not restricted to speech, but they also occurred with nonspeech events like clapping hands, in which case the effects were actually stronger. There were no topographical differences between the AV and auditory-evoked N1, which suggests that AV integration modulates the neural generators of the auditory N1 (Besle et al., 2004; Oray, Lu, & Dawson, 2002; Adler et al., 1982). We also observed a qualitative distinction between the early N1 effect and the later occurring P2 effects. Suppression and speeding-up of the N1 was unaffected by whether the auditory and visual information were congruent or incongruent. Instead, the N1 effect crucially depended on whether the visual information contained anticipatory motion. When there was no anticipatory visual motion, the cross-modal effect on the N1 disappeared. This indicates that it is the temporal information in the visual stimulus rather than the content of the sound that is key to the AV interaction.

In contrast to this early AV interaction, the later occurring effect on P2 was content-dependent because the amplitude reduction of P2 was bigger for incongruent than congruent AV stimuli. Whereas congruency effects for nonspeech stimuli were mainly confined to auditory P2, pointwise t tests revealed an additional late congruency effect for speech stimuli (Figure 3). The fact that this speech-specific interaction was found at different (occipito-parietal) electrodes—similar to Klucharev et al. (2003)—than the more centrally distributed congruency effect in nonspeech events may indicate a dissociation between AV integration at the phonetic level versus the associative or semantic level. Our data, therefore, demonstrate that there are two qualitatively different integrative mechanisms at work with different underlying time courses. The early N1 interactions are unaffected by informational congruency and crucially depend on the temporal relationship between visual and auditory signals, whereas the mid-latency and late interactions are susceptible to informational congruency and possibly indicate multisensory integration at the associative, semantic, or phonetic level.

Others have argued previously that the suppression of auditory N1 is exclusively related to the integration of AV speech, because this was not found in simplified AV combinations such as pure tones and geometrical shapes (Fort et al., 2002; Giard & Peronnet, 1999), or spoken and written forms (Raij, Uutela, & Hari, 2000). These comparisons, though, have so far left unexplained what the unique properties of AV speech are that cause the effect. It might, among others, be the ecological validity of AV speech, the meaningful relationship between A and V, the fact that visual speech provides phonetically relevant information, or the dominance of the auditory modality in AV speech (van Wassenhove et al., 2005; Besle et al., 2004; Klucharev et al., 2003). Our results demonstrate that (the lack of) visual anticipatory motion is crucial. We observed striking similarities between the neural correlates of AV integration of speech and nonspeech events provided that the nonspeech events contained visual anticipatory information. Most likely, therefore, early AV interactions in the auditory cortex are not speech-specific, but reflect anticipatory visual motion whether present in speech or nonspeech events.

What are the neural mechanisms involved in multisensory processing of AV speech and nonspeech events? Neuroimaging and electrophysiological studies of AV speech and nonspeech objects have found multisensory interactions in multimodal areas such as the superior temporal sulcus (STS) and sensory-specific areas including the auditory and visual cortices (van Wassenhove et al., 2005; Beauchamp, Lee, Argall, & Martin, 2004; Besle et al., 2004; Callan et al., 2003, 2004; Möttönen, Schürmann, & Sams, 2004; Klucharev et al., 2003; Calvert et al., 1999; Giard & Peronnet, 1999). It has been proposed that the unisensory signals of multisensory objects are initially integrated in the STS, and that interactions in the auditory cortex reflect feedback inputs from the STS (Calvert et al., 1999). On this account, one expects the suppressive effects in the auditory cortex in our Experiments 1 and 2 to be mediated by the STS via backward projections (Besle et al., 2004). The function of this feedback might be to facilitate and speed up auditory processing. As concerns this speeding-up interpretation, it should be noted, however, that although visual anticipatory information induced a shortening of the N1 *peak*, it did not affect the *onset* and the *slope* of the N1, as they were similar for A and AV stimuli (see Figure 2, and also van Wassenhove et al., 2005). The speeding-up of the peak of N1 may therefore be an artifact due to the effect that visual information reduces the amplitude itself.

Recently, the feedback interpretation from the STS has been challenged by an MEG study in which it was demonstrated that interactions in the auditory cortex (150–200 msec) *preceded* activation in the STS region

(250–600 msec) (Möttönen et al., 2004). In addition, an ERP study demonstrated that visual speech input may affect auditory-evoked responses via subcortical (brainstem) structures (Musacchia, Sams, Nicol, & Kraus, 2006). These very early AV interactions at the level of the brainstem (~11 msec) may only become understandable if one realizes that the visual input in AV speech can precede the auditory signal by tens, if not hundreds, of milliseconds. Based on our findings, we therefore conjecture that such early interactions may also be found with nonspeech stimuli, provided that the visual signal contains anticipatory information about sound onset.

Another link that possibly mediates AV interactions is that, besides the STS, motor regions of planning and execution (Broca's area, premotor cortex, and anterior insula) are involved via so-called mirror neurons (Ojanen et al., 2005; Skipper et al., 2005; Callan et al., 2003, 2004; Calvert & Campbell, 2003). Broca's area has been suggested to be the homologue of the macaque inferior premotor cortex (area F5) where mirror neurons reside that discharge upon action and perception of goal-directed hand or mouth movements (Rizzolatti & Craighero, 2004). The presumed function of these mirror neurons is to mediate imitation and aid action and understanding (Rizzolatti & Craighero, 2004). Broca's area is not only involved in speech production (Heim, Opitz, Muller, & Friederici, 2003) but is also activated during silent lipreading (Campbell et al., 2001) and passive listening of auditory speech (Wilson, Saygin, Sereno, & Iacoboni, 2004). Activation of mirror neurons in Broca's area may, on this view, thus constitute a link between auditory and visual speech inputs and the corresponding motor representations. On this motor account of AV speech, vision affects auditory processing via articulatory motor programs of the observed speech acts (Callan et al., 2003). Interestingly, Broca's area is not only active during AV speech but is also responsive to perception and imitation of meaningful goal-directed hand movements (Koski et al., 2002; Grezes, Costes, & Decety, 1999; Iacoboni et al., 1999). It may therefore be the case that the AV interactions of our nonspeech events were mediated by mirror neurons in Broca's area. If so, it becomes interesting to test whether artificial AV stimuli that lack an action component evoke similar AV integration effects.

Besides "facilitating" auditory processing (van Wassenhove et al., 2005; Besle et al., 2004) or "mediating actions" (Skipper et al., 2005; Callan et al., 2003), there are yet other functional interpretations of the AV interaction effect. One alternative is that visual anticipatory motion evokes sensory "gating" of auditory processing (Musacchia et al., 2006; Lebib, Papo, de Bode, & Baudonniere, 2003). Sensory gating refers to blocking or filtering out redundant information or stimuli (Adler et al., 1982). In the auditory domain, sensory gating takes place when a sound is preceded by the same sound within 1 sec and is reflected by the suppression of auditory potentials (P50, N1, P2) (Kizkin, Karlidag, Ozcan, & Ozisik, 2006; Johannesen et al., 2005; Arnfred, Chen, Eder, Glenthoj, & Hemmingsen, 2001; Nagamoto, Adler, Waldo, & Freedman, 1989; Adler et al., 1982). Along with suppression of auditory ERP components, a number of studies report shortening of N1 latency as well (Kizkin et al., 2006; Johannesen et al., 2005; Croft, Dimoska, Gonsalvez, & Clarke, 2004; Arnfred, Chen, et al., 2001; Arnfred, Eder, Hemmingsen, Glenthoj, & Chen, 2001). Importantly, sensory gating can be observed cross-modally as auditory N1 and P2 are suppressed when a click is paired with a leading flash (Oray et al., 2002). The suppression and speeding-up of auditory activity in speech and nonspeech events might therefore be interpreted as the neural correlate of cross-modal sensory gating. Our study and other AV speech studies (Jääskeläinen et al., 2004; Klucharev et al., 2003) have also shown that cross-modal sensory gating of N1 does not depend on the informational congruency between A and V, but crucially depends on the temporal relation. That is, auditory processing is only suppressed when the visual signal is leading, and thus, predicts sound onset. Consistent with this interpretation, there are no AV effects on N1 when there are no visible lip movement preceding the utterance of a vowel (Miki, Watanabe, & Kakigi, 2004). Likewise, in the absence of anticipatory visual motion, pictures of animals and their vocalization (Molholm, Ritter, Javitt, & Foxe, 2004) or artificial auditory and visual objects (geometric figures and beeps) do not suppress auditory potentials (Fort et al., 2002; Giard & Peronnet, 1999).

It might further be reasoned, arguably, that an attentional account explains the current findings. For example, one may conjecture that visual anticipatory information serves as a warning signal (a cue) that directs attention to the auditory channel. The AV interaction effect on the auditory N1 would then, in essence, reflect the difference between attended versus unattended auditory information. Such an attentional account, though, seems unlikely because directing attention to the auditory modality generally results in an amplitude increase rather than decrease of ERP components in the time window of auditory N1 (Besle et al., 2004; Näätänen, 1992). One could also ask whether the visual task, as used in the present study (the detection of a spot in catch trials), had an effect on the observed AV interaction. For example, would similar results be obtained if participants were engaged in an auditory rather than visual task? We conjecture that such task-related effects will only be secondary because, at least with AV speech stimuli, similar results (depression of auditory N1) were obtained when attention was focused on the auditory modality (Besle et al., 2004) rather than on the visual modality (our study). Furthermore, van Wassenhove et al. (2005) manipulated the attended modality (focus on either the visual or auditory modality) and found no effect of this manipulation.

There are two potential reasons why depression and latency facilitation of N1 and P2 were more pronounced

for nonspeech events than for speech events. Firstly, the nonspeech stimuli contained more anticipatory visual motion (280–320 msec) than the speech stimuli (160–200 msec), which may be a more optimal temporal window for prediction of sound onset. Secondly, nonspeech events were also more precisely defined in time because of their relatively sharp visual and auditory onsets, whereas the rise time of our speech stimuli was more gradual. The temporal precision of a subset of our stimuli (the /bi/ and the handclap) was further determined in a control experiment, wherein 15 participants performed a cross-modal temporal order judgment task between A and V. The onset asynchronies of A and V varied from −320 msec (A first) to +320 msec (V first) in 40-msec steps. Participants judged whether A or V was presented first. The just noticeable difference (JND), which reflects the time between A and V needed to accurately judge in 75% of the cases which stimulus appeared first, was smaller (i.e., better performance) for nonspeech events (64.5 msec) than for speech (105.8 msec), $t(14) = 4.65$, $p < .001$, thus confirming that the temporal relation between A and V of the nonspeech events was more precisely defined.

To conclude, our results demonstrate that the neural correlates underlying integration of A and V are not speech-specific because they are also found in nonspeech AV events, provided that there is visual anticipatory motion. These results bear importance to the question whether the processes underlying multisensory integration of AV speech are unique to speech or can be generalized to nonspeech events (Tuomainen, Andersen, Tiippana, & Sams, 2005; van Wassenhove et al., 2005; Besle et al., 2004; Massaro, 1998). We conjecture that when critical stimulus features are controlled for, especially the temporal dynamics between A and V, there is no difference in early AV integration effects between speech and nonspeech. Rather, speeding-up and suppression of auditory potentials are induced by visual anticipatory motion, which can be inherent to both speech and nonspeech events. Whether the AV ERP effects reflect a general or a more specific human action-related multisensory integrative mechanism is open to debate. Further evidence would come from studies in which visual predictability in nonspeech and nonaction-related AV events is manipulated.

Reprint requests should be sent to Jeroen J. Stekelenburg, Psychonomics Laboratory, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands, or via e-mail: J.J.Stekelenburg@uvt.nl.

Notes

1. Using a univariate analysis of variance with Greenhouse–Geisser corrections, we additionally tested in all experiments the N1 and P2 distributions incorporating all 47 electrodes. No differences were found between this approach and the one using a limited number of electrode positions.

2. Similar results were obtained when the analyses were performed without the control condition C, and thus, comparing directly AV − V − A.

REFERENCES

Adler, L. E., Pachtman, E., Franks, R. D., Pecevich, M., Waldo, M. C., & Freedman, R. (1982). Neurophysiological evidence for a defect in neuronal mechanisms involved in sensory gating in schizophrenia. *Biological Psychiatry, 17,* 639–654.

Arnfred, S. M., Chen, A. C., Eder, D. N., Glenthoj, B. Y., & Hemmingsen, R. P. (2001). A mixed modality paradigm for recording somatosensory and auditory p50 gating. *Psychiatry Research, 105,* 79–86.

Arnfred, S. M., Eder, D. N., Hemmingsen, R. P., Glenthoj, B. Y., & Chen, A. C. (2001). Gating of the vertex somatosensory and auditory evoked potential p50 and the correlation to skin conductance orienting response in healthy men. *Psychiatry Research, 101,* 221–235.

Barth, D. S., Goldberg, N., Brett, B., & Di, S. (1995). The spatio-temporal organization of auditory, visual, and auditory–visual evoked potentials in rat cortex. *Brain Research, 678,* 177–190.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20,* 2225–2234.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14,* 2213–2218.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16,* 805–816.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport, 10,* 2619–2623.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57–70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Research, Cognitive Brain Research, 12,* 233–243.

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,* 495–506.

Croft, R. J., Dimoska, A., Gonsalvez, C. J., & Clarke, A. R. (2004). Suppression of p50 evoked potential component, schizotypal beliefs and smoking. *Psychiatry Research, 128,* 53–62.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Early auditory–visual interactions in human cortex during nonredundant target identification. *Brain Research, Cognitive Brain Research, 14,* 20–30.

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,* 473–490.

Gonzalo, D., & Büchel, C. (2004). Audio-visual associative learning enhances responses to auditory stimuli in visual cortex. In N. Kanwisher & J. Duncan (Eds.), *Functional neuroimaging of visual cognition: Attention and performance XX* (pp. 225–240). New York: Oxford University Press.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484.

Grezes, J., Costes, N., & Decety, J. (1999). The effects of learning and intention on the neural network involved in the perception of meaningless actions. *Brain, 122,* 1875–1887.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology, 28,* 240–244.

Heim, S., Opitz, B., Muller, K., & Friederici, A. D. (2003). Phonological processing during language production: fMRI evidence for a shared production–comprehension network. *Brain Research, Cognitive Brain Research, 16,* 285–296.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286,* 2526–2528.

Jääskeläinen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., et al. (2004). Adaptation of neuromagnetic n1 responses to phonetic stimuli by visual speech in humans. *NeuroReport, 15,* 2741–2744.

Johannesen, J. K., Kieffaber, P. D., O'Donnell, B. F., Shekhar, A., Evans, J. D., & Hetrick, W. P. (2005). Contributions of subtype and spectral frequency analyses to the study of p50 ERP amplitude and suppression in schizophrenia. *Schizophrenia Research, 78,* 269–284.

Kizkin, S., Karlidag, R., Ozcan, C., & Ozisik, H. I. (2006). Reduced p50 auditory sensory gating response in professional musicians. *Brain and Cognition, 61,* 249–254.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research, Cognitive Brain Research, 18,* 65–75.

Koski, L., Wohlschlager, A., Bekkering, H., Woods, R. P., Dubeau, M. C., Mazziotta, J. C., et al. (2002). Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex, 12,* 847–855.

Lebib, R., Papo, D., de Bode, S., & Baudonniere, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human p50 event-related brain potential modulation. *Neuroscience Letters, 341,* 185–188.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: 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,* 203–208.

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

Miki, K., Watanabe, S., & Kakigi, R. (2004). Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans: A magnetoencephalographic study. *Neuroscience Letters, 357,* 199–202.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14,* 452–465.

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. *Brain Research, Cognitive Brain Research, 14,* 115–128.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Research, Cognitive Brain Research, 13,* 417–425.

Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. *Neuroscience Letters, 363,* 112–115.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research, 168,* 1–10.

Näätänen, R. (1992). *Attention and brain function*. Hillsdale, NJ: Erlbaum.

Nagamoto, H. T., Adler, L. E., Waldo, M. C., & Freedman, R. (1989). Sensory gating in schizophrenics and normal controls: Effects of changing stimulation interval. *Biological Psychiatry, 25,* 549–561.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25,* 333–338.

Oray, S., Lu, Z. L., & Dawson, M. E. (2002). Modification of sudden onset auditory ERP by involuntary attention to visual stimuli. *International Journal of Psychophysiology, 43,* 213–224.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron, 28,* 617–625.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.

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,* 141–145.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25,* 76–89.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Brain Research, Cognitive Brain Research, 14,* 106–114.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96,* B13–B22.

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, U.S.A., 102,* 1181–1186.

Von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PloS Biology, 4,* 1809–1820.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.

This article has been cited by:

1. Jean Vroomen, Jeroen J. Stekelenburg. 2010. Visual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual StimuliVisual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual Stimuli. *Journal of Cognitive Neuroscience* **22**:7, 1583-1596. [Abstract] [Full Text] [PDF] [PDF Plus]
2. U. Zimmer, S. Itthipanyanan, T. Grent-'t-Jong, M.G. Woldorff. 2010. The electrophysiological time course of the interaction of stimulus conflict and the multisensory spread of attention. *European Journal of Neuroscience* **31**:10, 1744-1754. [CrossRef]
3. Gabriella Musacchia, Laurie Arum, Trent Nicol, Dean Garstecki, Nina Kraus. 2009. Audiovisual Deficits in Older Adults with Hearing Loss: Biological Evidence. *Ear and Hearing* **30**:5, 505-514. [CrossRef]
4. Mikhail Zvyagintsev, Andrey R. Nikolaev, Heike Thönnessen, Olga Sachs, Jürgen Dammers, Klaus Mathiak. 2009. Spatially congruent visual motion modulates activity of the primary auditory cortex. *Experimental Brain Research* **198**:2-3, 391-402. [CrossRef]
5. Daniel Senkowski, Till R. Schneider, Frithjof Tandler, Andreas K. Engel. 2009. Gamma-band activity reflects multisensory matching in working memory. *Experimental Brain Research* **198**:2-3, 363-372. [CrossRef]
6. J. Navarra, J. Hartcher-O'Brien, E. Piazza, C. Spence. 2009. Adaptation to audiovisual asynchrony modulates the speeded detection of sound. *Proceedings of the National Academy of Sciences* **106**:23, 9169-9173. [CrossRef]
7. Julie Brefczynski-Lewis, Svenja Lowitszch, Michael Parsons, Susan Lemieux, Aina Puce. 2009. Audiovisual Non-Verbal Dynamic Faces Elicit Converging fMRI and ERP Responses. *Brain Topography* **21**:3-4, 193-206. [CrossRef]
8. Maurice J.C.M. Magnée, Beatrice de Gelder, Herman van Engeland, Chantal Kemner. 2008. Audiovisual speech integration in pervasive developmental disorder: evidence from event-related potentials. *Journal of Child Psychology and Psychiatry* **49**:9, 995-1000. [CrossRef]