

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

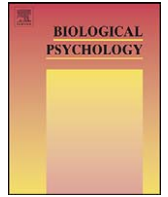


This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Deviance detection in congruent audiovisual speech: Evidence for implicit integrated audiovisual memory representations

István Winkler^{a,b,*}, János Horváth^a, Júlia Weisz^a, Leonard J. Trejo^c

^a Institute for Psychology, Hungarian Academy of Sciences, Hungary

^b Institute of Psychology, University of Szeged, Hungary

^c Pacific Development and Technology, LLC, USA

ARTICLE INFO

Article history:

Received 5 June 2009

Accepted 31 August 2009

Available online 4 September 2009

Keywords:

Audiovisual integration

Speech perception

Implicit memory

Deviance detection

Event-related brain potentials (ERP)

Mismatch negativity (MMN)

ABSTRACT

Detection of deviant speech syllables embedded in continuous noise was investigated in an oddball paradigm. Behavioral results showed improvement of detecting and identifying the syllables when congruent visual speech accompanied the utterances. A centrally maximal negative ERP difference wave peaking at approximately 290 ms post-stimulus was elicited by audiovisual but not by auditory- or visual-only task-irrelevant deviant syllables. Whereas the circumstances of the elicitation of this ERP response are similar to those of the mismatch negativity component (MMN and its visual counterpart, vMMN), its scalp distribution differs from that of both unimodal MMNs. Elicitation of an MMN-like ERP response (termed here as the audiovisual MMN: avMMN) suggests that detection of the audiovisual deviants involved integrated audiovisual memory representations. The pattern of behavioral and ERP results suggest that the formation of such cross-modal memory representation does not require voluntary operations and may even proceed for stimuli outside the focus of attention.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Perception of audiovisual (AV) speech is one of the natural situations in which strong cross-modal integration is assumed to take place in the human brain. Indeed, everyday experience as well as evidence from formal perceptual experiments agree that seeing the speaker's face improves the intelligibility of speech in noisy environments (e.g., Erber, 1975; Helfer and Freyman, 2005; Sams et al., 2005; Sommers et al., 2005; Sumbly and Pollack, 1954) and can disambiguate ambiguous speech utterances (Bertelson et al., 2003). One important question is whether the integration of auditory and visual information leading to the observed advantage in speech perception occurs early, at lower or late, at higher levels of processing. It has been suggested that visual information may fine-tune sound analysis, speeding it up by preparing for the most likely distinctive features of the acoustic signal and increasing its efficacy (e.g., van Wassenhove et al., 2005). Other studies argue that the effects of visual information are mediated by the articulatory preparation and motor system, possibly through mirror-neuron functions (e.g., Callan et al., 2003; Davis and Kim,

2004; Skipper et al., 2005, 2007) which may operate at the phonological level of perception (Mitterer, 2006). In the current study, we asked whether the integration of audio–visual information in speech perception may be subserved by more general integration processes. Specifically, we asked whether implicit sensory memory representations of AV speech stimuli combine information from both modalities. Implicit sensory memory representations are obligatorily formed for unimodal stimuli in all sensory modalities (Broadbent, 1958). However, little is known about whether or not sensory memory representations can integrate information for multiple modalities. We tested this question using implicit detection of changes in a sequence of a repeating congruent AV syllable.

A significant part of our perceptual abilities rely on the storage of sensory information. For example, in the auditory modality, segregation of sound sources by sequential rules assumes the existence of memory representations encoding the immediate history of the behavior of each tracked source (Bregman, 1990; Denham and Winkler, 2006; Winkler, 2007; for the role of implicit memory in the visual modality, see Czigler, 2007; Fernandez-Duque and Thornton, 2003; Merikle et al., 2001). It is assumed that sensory memory representations are created for all separable incoming stimuli. However, the information stored in them may become conscious only when selected for further processing. Such implicit memory representations encoding sensory information have been regarded as modality specific since Broadbent's (1958)

* Corresponding author at: Institute for Psychology, Hungarian Academy of Sciences, H-1394 Budapest, P.O. Box 398, Hungary.
Tel.: +36 1 3542296; fax: +36 1 3542416.

E-mail address: iwinkler@cogpsyphy.hu (I. Winkler).

introduction of the notion of sensory memory (cf. Coltheart, 1984; Neisser, 1967). Furthermore, sensory systems have been suggested to be modular (Fodor, 1983). However, recent findings showed that sensory cortical areas can be modulated by stimulation of a different modality (e.g., Fort and Giard, 2004). For example, Fu et al. (2003) found neurons responding to somatosensory stimulation in the caudomedial region of macaque auditory cortex. Also, cross-modal information, such as visual speech can affect implicit auditory memory traces (Besle et al., 2005) as was, for example shown by electrophysiological studies demonstrating that the McGurk illusion (McGurk and MacDonald, 1976) can lead to the pre-attentive detection of auditory sensory deviance (Colin et al., 2002, 2004; Saint-Amour et al., 2007; Sams et al., 1991). Thus it is possible that speech perception involves implicit memory traces integrating audiovisual information.

In recent years, functional neuroimaging studies (functional magnetic resonance imaging [fMRI], positron emission tomography [PET], high-density electroencephalography [EEG], and magnetoencephalography [MEG]) provided much information about the neural network underlying the processing of AV speech in the human brain. Several studies found increased, often supra-additive activation during congruent audiovisual (compared with unimodal) stimulation in the (left) superior temporal sulcus and gyrus (STS/G), the lateral part of Heschl's gyrus (Brodmann area 41), secondary auditory cortical areas (BA 42), and the region of the occipito-temporal junction, an area associated with processing visual motion (e.g., Callan et al., 2004; Calvert and Campbell, 2003; Kang et al., 2006; Miller and D'Esposito, 2005). Of these areas, STS was found to be especially sensitive to processing AV speech with degraded acoustic signals, suggesting that this multi-sensory area plays a crucial role in improving the identification of speech sounds under noisy conditions (for a review, see Assmann and Summerfield, 2004). STS activation was found to be sensitive to the synchrony of auditory and visual input (though synchrony has a ca. 250-ms acceptance window—see Dixon and Spitz, 1980; Navarra et al., 2005; van Wassenhove et al., 2007), whereas it showed no sensitivity to whether the auditory and visual input originated from the same physical location (~ventriloquism; Jones and Jarick, 2006; Macaluso et al., 2004). The evidence for STS being a critical player in AV speech integration is, however, not conclusive (Jones and Callan, 2003; Ojanen et al., 2005; Olson et al., 2002). Although imaging studies provided much information about the neural network involved in processing AV speech, fMRI and PET lack the temporal sensitivity needed to describe the timing and order of the AV speech related activations, and thus to engender functional models of processing AV speech.

Studies using EEG or MEG examined the time-course of audiovisual interactions for AV speech stimuli. Signs of early gating through attenuation of the mid-latency P50 component by congruent as opposed to incongruent AV speech have been found (Lebib et al., 2003; see, however, Reale et al., 2007). Besle et al. (2004) and van Wassenhove et al. (2005) showed suppression of the obligatory supratemporal auditory N1/P2 responses in the 100–190 ms latency range by congruent AV speech. Visual speech effects on the auditory N1 have also been observed by Jääskeläinen et al. (2004) and Pourtois et al. (2000), although the latter found facilitation instead of suppression. Neural activity evoked by congruent auditory and visual speech stimuli interacted in the 150–200 ms post-stimulus interval within auditory cortex and in the 250–600 ms interval in STS (Möttönen et al., 2004). Klucharev et al. (2003) suggested that the very early effects (found at 85 ms by these authors) are not speech-specific, whereas later effects (155–325 ms) possibly originate from heteromodal brain areas of phonetic nature. In summary, interaction between congruent auditory and visual information can be observed already shortly after the onset of the acoustic speech signal in auditory sensory

cortical areas, continuing for quite some time in various brain areas. However, there is a concern that the early effects, such as the P50 are not specific to speech processing (Klucharev et al., 2003; see, also Korzyukov et al., 2007). The effect of visual speech on sound processing is often of suppressing nature, an observation compatible with the results of some cross-modal neuroimaging studies (Bushara et al., 2003; Kang et al., 2006; Laurienti et al., 2002).

It is, however, not clear from the above results, how much of the effects can be attributed to task-related processes, such as participant strategies. One general problem of AV speech research lies in the difficulty of separating the effects of lip-reading (i.e., speech identification by visual cues alone) from possible improvements of the sensory stimulus representations. In search of auditory-specific effects of congruent visual information, it has been found that visual speech decreases the auditory detection threshold for congruent speech sounds by 2–3 dB (Grant, 2001; Grant and Seitz, 2000; Kim and Davis, 2003a). A lesser decrease in the auditory threshold can be observed with concurrent presentation of a non-speech visual stimulus (Bernstein et al., 2004) and Schwartz et al. (2004) showed a small but significant identification advantage of AV over auditory-only speech when different auditory speech stimuli were presented together with a common, but compatible visual speech stimulus (i.e., differences between the visual cues were eliminated).

Several studies compared the event-related potential (ERP) responses elicited by congruent and incongruent AV speech stimuli (Colin et al., 2002, 2004; Fingelkurts et al., 2003; Kaiser et al., 2005; Möttönen et al., 2002; Sams et al., 1991; Trejo et al., 2000). Most of these investigations were based on the McGurk illusion, in which, depending on the combination of the utterance and the simultaneously presented visual speech signal, information from the two modalities fuse into a percept that does not correspond to either unimodal stimulus or the visual stimulus dominates the auditory percept. Incongruent AV speech stimuli (compared with congruent ones) produced negative modulation of the ERP responses at about 300 ms from sound onset (Lebib et al., 2004). Many studies presented the congruent and incongruent AV stimuli in oddball sequences in which infrequent incongruent AV stimuli were presented amongst frequent congruent ones. Kaiser et al. (2005) found evoked gamma band activity peaking at ca. 160 ms from stimulus onset over posterior parietal cortex to infrequently changing the visual stimulus to a different syllable within a sequence of a frequently repeating congruent AV syllable. An ERP sign of auditory change, the mismatch negativity (MMN or its magnetic counterpart, termed MMNm or MMF; Näätänen et al., 1978; for recent reviews, see Garrido et al., 2009; Kujala et al., 2007) was elicited in the 140–300 ms latency range in auditory cortex by infrequently exchanging the visual component of a congruent AV speech stimulus to one which evoked the McGurk illusion (Colin et al., 2002, 2004; Saint-Amour et al., 2007; Sams et al., 1991). Saint-Amour et al. (2007) also found further brain activity elicited by infrequent deviants evoking the McGurk illusion. This ERP response was observed in the 350–400 ms post-stimulus interval; it originated both from temporal cortex and STG and showed strong left-hemispheric asymmetry. Although findings of auditory cortical effects of the McGurk illusion strongly support the notion that visual speech cues affect processing within the auditory system, by their very nature, illusions may not reveal the normal functioning of audiovisual integration of speech cues. It is possible that integrating incongruent auditory and visual speech information requires additional (or different) processing compared with the normal congruent case. For example, whereas cross-modal integration of AV speech is generally regarded to be automatic (see e.g., Soto-Faraco et al., 2004), Alsius et al. (2005) showed that when attention

is strongly focused on different stimuli, the McGurk illusion breaks down.

To deal with some of the above discussed problems of studying auditory-visual integration of speech cues, the current study tested implicit discrimination of congruent AV speech stimuli when the speech sounds were masked by concurrent noise, making the auditory discrimination very difficult. That is, visual speech cues were needed for successful discrimination of the different speech stimuli. In the main experiment, ERPs were recorded in an oddball paradigm (A, V, and AV conditions, separately) while subjects performed a visual target detection task, which was unrelated to the speech stimuli (both auditory and visual). Previous studies found ERP indices of detecting occasional deviants amongst a frequently repeating stimulus (standard), separately in the auditory and the visual modality: MMN and vMMN, respectively. Some studies also showed MMN-like effects for occasional audiovisual incongruence (Widmann et al., 2004). These ERP responses were shown to be based on implicit memory representations encoding sensory features of the repeating stimulus. Therefore, we searched for ERP signs of detecting the infrequent stimulus in the AV condition, because such ERP response could reflect the presence and quality of the underlying memory representations. If a discriminatory ERP response was found in the AV condition which could not be explained by combining the ERP response elicited by auditory-only (A condition) and visual-only (V condition) discrimination, this would suggest that (1) the brain formed an integrated audiovisual representation of the task-irrelevant AV speech, (2) which was used for implicit discrimination (deviance detection) and (3) this representation contained more distinctive information for the AV speech stimuli than the two unimodal representations combined. Alternatively, full additivity between the auditory and visual discriminative ERP responses would suggest that implicit discrimination of audiovisual speech stimuli uses auditory and visual information separately. Experiment 2 tested whether, due to the degraded auditory information delivered to them in the AV speech situations of Experiment 1, participants based their perceptual judgments on visual information. Finally, Experiment 3 tested whether in Experiment 1, visual information helped auditory discrimination only by providing information about the timing of the speech stimuli.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Thirteen healthy participants (3 male, age 20–25 years, mean 22.0) were recruited through a student part-time job agency. They received modest financial compensation for their participation. The study was approved by the Ethical Committee of the Institute for Psychology, Hungarian Academy of Sciences. Participants signed an informed consent form after the aims and procedures of the study were explained to them, before starting the experiment. One participant's data were rejected because of extensive electric artefacts in the EEG. Participants were pre-selected on the basis of a clinical audiometry with the criteria that the hearing threshold between 500 and 4000 Hz should not be higher than 25 dB, and the difference between the two ears not higher than 15 dB in the same frequency range. All participants had normal or corrected to normal vision.

2.1.2. Stimuli

Speech stimulation consisted of auditory (Condition "A"), visual (Condition "V"), and synchronized congruent audiovisual (Condition "AV") syllables ('ma' and 'ka'). The syllables were natural utterances recorded from the same actor speaking with minimal head movements in front of a homogeneous background (speech stimuli were originally recorded and edited for the study of Trejo et al., 2000). The audio-visual recordings were cut into 500-ms segments (15 frames at a 30 fps presentation rate). The syllables were delivered with a stimulus onset asynchrony (SOA; onset-to-onset interval) of 1000 ms. The duration of the auditory signal was 400 ms with the visual stimulus starting together (from a neutral expression) with the auditory one and lasting for 500 ms (this time included a short straight-face

period following the speech-related facial movements). The interval between the offset of the visual part of a stimulus and the onset of the next stimulus was filled with the last frame of the/ka/stimulus. Sound intensity was individually adjusted to 50 dB above sensation threshold (sensation level–SL). Sensation threshold was tested with a tone of 800 Hz frequency, whose sensation threshold was previously aligned (on a different set of subjects) with that of the speech sounds by establishing the intensity difference between the sensation threshold for the tone and the speech sounds. By using the tone for testing the sensation threshold, we avoided introducing the speech sounds to subjects prior to the main experiment. The visual recording had 320 × 240 pixel resolution and was shown on a 17" monitor screen with the participant sitting at a distance of 1.2 m (viewing angle 15.4 × 9.7°, that of the speaker's face 4.8 × 8.2°). The bottom 40-pixel-wide strip of the monitor was covered by black tape, because in this strip short flashes were embedded in the video pictures for synchronizing the ERP recordings. The flashes were detected by a photo-sensitive diode attached to the bottom left corner of the monitor and converted into electronic triggers for extracting ERP responses from the continuous EEG record (see below). These triggers were only used to mark the timing of the targets in the visual change detection task (see below). Because visual targets appeared at approximately the center of the screen, the visual target detection reaction times were corrected by adding 8 ms to the values measured from the trigger. Pilot results showed that the auditory speech stimuli elicited the N1 wave peaking at 112 ms from sound onset. This means that the sound onset detected by the auditory system was close to the physical onset of the sounds. Since the auditory and visual stimuli commenced synchronously, all ERPs were triggered from the onset of the sounds (or the expected sound onset in case of the V condition). Thus the measured ERP responses are aligned across the stimulus conditions and the relative component latencies are accurate, although the absolute latency values cannot be established with the same precision as can be measured for abruptly commencing stimuli.

Continuous masking noise of 75-dB (SL) intensity was delivered throughout the stimulus blocks. The noise was a cyclically repeating sound segment selected from a sound effects collection (Sound Effects, Vol. 6, Track 78: "Indian Dance") because its spectrum closely followed those of the syllables and thus strongly affected speech perception (see Assmann and Summerfield, 2004). The length of the repeating segment was 996.3 ms and exactly 271 noise cycles were presented during each 270-stimulus long stimulus block (because the SOA of the syllables was 1 s). Thus the phase difference between the test stimuli and the noise went through a full cycle during each stimulus block. The starting phase of the noise was randomized across stimulus blocks and subjects. Transients between successive noise segments have been smoothed by 5 ms linear rise and 5 ms fall amplitude windows.

2.1.3. Procedures

In the first part of the 5 h long experimental session, subjects performed a visual change detection task. During the stimulus blocks, a 0.1° wide and 0.3° tall red line segment was presented in the center of the viewing field, positioned on the nose of the speaker's image. Participants were instructed to press a response key whenever the orientation of the line changed from horizontal to vertical or back. Changes occurred randomly 13 times during each stimulus block with at least 1 s between consecutive changes.

The three stimulus conditions differed in the modality of the test stimuli: A, V, or AV. In the A condition, the display showed a still picture of the speaker's face with the superimposed target stimulus. The noise was also present in the V condition (but without the speech sounds). 90% 'ma' and 10% 'ka' syllables were presented in oddball sequences in a pseudo-randomized order forcing consecutive "ka" syllables to be separated by at least one "ma" syllable. For each stimulus condition, responses were also recorded for a control sequence in which the role of the two syllables was exchanged. Responses elicited by the frequent "ka" syllable in the control sequence were used to delineate the effects of stimulus deviation in the corresponding experimental condition, comparing the response elicited by 'ka' when it was the deviant with that elicited when it served as the standard.

Stimulus blocks consisted of 270 stimuli, including 27 deviants. Each condition received 1 control and 5 experimental stimulus blocks. The AV and A conditions were administered first to minimize the effects of learning to discriminate the visual stimuli on the ERP responses elicited in the AV condition. The AV and A stimulus blocks were delivered in a balanced order (A-AV-AV-A-A-AV-AV-A-A-AV for half and the opposite order for the other half of the participants). The A and the AV control stimulus blocks were delivered in random serial positions, one between the 2nd and the 5th and the other between 8th and 11th stimulus block. Short breaks separated the presentation of successive stimulus blocks with a longer break between the 6th and the 7th block. Following another longer break, the 5 experimental and 1 control stimulus block of the V condition were delivered. Again, the control block was presented randomly between the 2nd and the 5th serial positions.

In the second part of the experimental session, we tested the effects of congruent audio-visual information on detecting and identifying syllables in noise. Participants were instructed to detect deviant speech sounds in the same oddball stimulus blocks (deviance-detection conditions) as were presented in the first part of the experiment. In addition, in separate stimulus blocks, subjects performed a forced-choice reaction task (classification conditions) with 50% "ma" and 50% "ka" stimuli delivered in a random order (54 stimuli per stimulus block). During these stimulus blocks, no EEG was recorded. The instruction emphasized that both

detection and classification should be done on the basis of the speech sounds even if visual information were also available. Each task (deviance detection and classification) was performed with A, AV, and A-no-noise stimuli (the latter serving to establish the baseline performance), one stimulus block, each. For reducing the duration of the session, no active V condition was administered to subjects. The V condition had been previously tested in a pilot experiment (with a different group of subjects: 13 participants 9 female, age: 19–26 years, mean 21; 1 subject's data were discarded, because of technical problems in one condition; all subjects had normal or corrected to normal vision). In the AV deviance-detection condition, 7 incongruent audiovisual stimuli (auditory “ma” and visual “ka”) were randomly interspersed in the stimulus sequence. These were used as “catch” trials for participants who primarily relied on visual information in the active AV conditions (despite the instructions). However, the McGurk illusion evoked by this combination of stimuli (perception of “na”; Sekiyama, 1991) may have affected the responses. Therefore, in analyzing the data, a lenient criterion was used: only those participants' data was skipped, who produced more than 60% errors in the catch trials. Altogether, the data of 2 out of 12 subjects were rejected from the analysis of the behavioral responses. First, the deviance detection then the classification conditions for the A and AV stimuli were administered (the order balanced across subjects). These were followed by the same two conditions with A-no-noise stimuli.

2.1.4. EEG recording

EEG was recorded with Ag/AgCl electrodes from 16 scalp locations (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, T5, T6, O1, Oz, and O2, according to the international 10–20 system and the left and right mastoids, LM and RM, respectively) against a common reference electrode attached to the tip of the nose. Signals were amplified (Neuroscan Synamps amplifier), on-line filtered in the 0–40 Hz frequency range, and digitized with a sampling rate of 250 Hz. Horizontal EOG was recorded between two electrodes placed lateral to the outer canthi of the two eyes; vertical EOG was recorded between an electrode placed above and one below the right eye. Signals were off-line filtered with a 2–16 Hz pass-band. Epochs of 700 ms duration including a 100 ms pre-stimulus interval were extracted for each stimulus. Only control and deviant stimuli separated by at least 1 s from a target visual change and the participants' response to the change were included in the analysis. The first three epochs of each stimulus block, and epochs exceeding a voltage range of 100 μ V on any EEG or EOG channel were also discarded from the analysis.

2.1.5. Data analysis

2.1.5.1. EEG data. ERP components were identified from the group-averaged waveforms, separately for deviant and control stimuli, as well as for the deviant-minus-control differences. Amplitudes were measured for each participant and ERP component as the average voltage in a 24-ms long window centered on the group-averaged peak. The mean voltage in the 100-ms pre-stimulus interval was used as the baseline for amplitude measurements (see Table 2). Amplitude measurements were submitted to 2-factor (stimulus modality \times electrode), repeated measures ANOVAs. Greenhouse–Geisser correction was used as appropriate. Correction factors for the degrees of freedom (ϵ) and effect sizes (η^2) are reported. All significant effects are discussed.

2.1.5.2. Behavioral data. For the deviance detection (10–90%) and classification (50–50%) conditions, responses were accepted within 50–1000 ms post-stimulus time windows. Responses to non-targets were classified as false alarms using the same time window. To allow calculation of d' values in cases when the hit rate or the false alarm rate was 0 or 1, respectively, the measured value was replaced by $1/2N$ or $1-1/2N$, respectively (where N stands for the number of targets for hit rate or the number of non-targets for false alarm rate; see Macmillan and Creelman, 1991). Although the number of targets was the same (27) in the two active task conditions (deviance detection and classification), there were 9 times more non-targets in the deviance detection than in the classification condition, allowing a much wider range for the d' measure. To make the results from the two conditions compatible, 27 non-target deviance detection trials were randomly sampled, separately for each individual. Due to the low number of correct responses occurring mainly in the A condition, reaction times could not be analyzed. d' values were evaluated by a stimulus condition (A, AV, A-no-noise) \times task (classification vs. deviant detection) repeated measures ANOVA. For comparing performance in the V stimulus condition, which was conducted on a different group of participants, with that in the other three stimulus conditions, three ANOVAs were performed with task as a within-subject factor (classification vs. deviant detection) and stimulus condition as a between-group factor (V vs. A-no-noise; V vs. A; and V vs. AV respectively).

Reaction times and hit rates for the visual change-detection task performed during the EEG part of the experiment (50–1000 ms post-change time window) were analyzed with one-way ANOVAs of stimulus condition (A vs. V vs. AV).

2.2. Results

2.2.1. Discrimination of the speech stimuli

Fig. 1 shows the results of the deviance detection and classification tasks. Subjects showed poor performance both in detecting and in classifying the

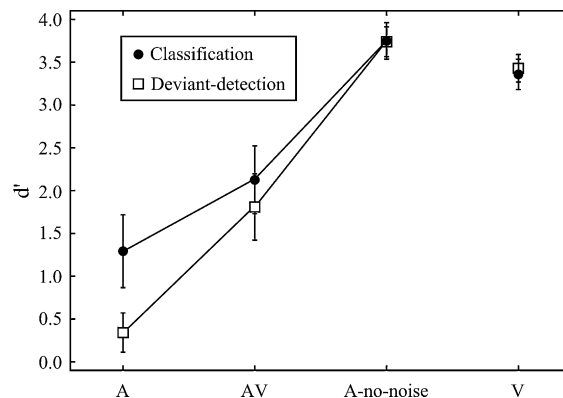


Fig. 1. Experiment 1. Group-averaged ($N = 10$) d' values measured in the deviant detection (empty squares) and classification (filled circles) task conditions for the A, AV, A-no-noise conditions and, separately, for the V stimulus condition. Performance for the V stimuli was measured in a different group of participants than that for the other stimuli. Standard error of mean values are marked by vertical lines at each measurement point.

infrequent deviants without visual speech cues. Performance with the combined AV speech stimuli was significantly better, but still less so than with the speech sounds alone without noise (ANOVA of the d' values: $F(2,18) = 45.51$, $\epsilon = 0.94$, $p < 0.001$, $\eta^2 = 0.84$, with Tukey HSD post hoc tests showing significant difference between each pair of levels by at least $p < 0.01$). Classification performance was significantly better than deviance-detection performance: $F(1,9) = 5.26$, $p < 0.05$, $\eta^2 = 0.37$ and there was also a tendency for an interaction between the two tasks and the three stimulus conditions ($F(2,18) = 3.49$, $\epsilon = 0.76$, $p = 0.07$, $\eta^2 = 0.28$), which stemmed from the significantly better (Tukey HSD, $p < 0.01$) classification than deviance-detection performance in the A condition. Performance in the V condition (different group of subjects) was only slightly (not significantly) lower than that in the A-no-noise condition and significantly better than in the A (noise) condition (main effect of stimulus modality [V vs. A]: $F(1,20) = 77.44$, $p < 0.001$, $\eta^2 = 0.79$; main effect of task: $F(1,20) = 4.57$, $p < 0.05$, $\eta^2 = 0.19$; and interaction: $F(1,20) = 6.15$, $p < 0.05$, $\eta^2 = 0.23$; post hoc Tukey HSD tests showed that, except for the V deviant detection and V classification scores, each pair of conditions differed significantly by at least $p < 0.05$). Performance in the V condition was also significantly higher than in AV (main effect of stimulus modality [V vs. AV]: $F(1,20) = 14.60$, $p < 0.01$, $\eta^2 = 0.42$).

2.2.2. Visual target detection

ANOVA tests on hit rates (HR) and reaction times (RT) (see Table 1) showed that target detection was more difficult in the presence of lip movements compared with that over the background of the picture of the still face (HR: $F(2,22) = 3.90$, $p < 0.05$, $\eta^2 = 0.26$ with Tukey HSD tests showing a $p < 0.05$ significant difference between the A and V conditions; RT: $F(2,22) = 9.61$, $p < 0.001$, $\eta^2 = 0.47$, with the A condition differing by $p < 0.01$ from both V and AV conditions). However, no difference was found as a function of the presence or absence of the acoustic component of the speech stimulus (i.e., no difference between the V and AV conditions).

2.2.3. ERP results

The repeating (control) syllable embedded in noise without a corresponding visual stimulus elicited ERP components characterized by rather low-amplitudes. A significant ($t(11) = -2.59$, $p < 0.05$) central (Cz) waveform peaking at 310 ms was elicited. This peak also appears in a control condition in which the syllables were delivered without noise (see Fig. 6 of Experiment 3). The low-amplitude ERPs obtained in the A condition suggest that the speech sounds could not be easily distinguished from the background noise. In the V and AV conditions, prominent

Table 1

Experiments 1 and 3: group-averaged ($N = 12$ and $N = 10$) hit rates and reaction times in the visual change detection task (performed during the EEG recordings) for the A (with noise in Experiment 1, no-noise in Experiment 3), V (Experiment 1 only), and AV stimulus conditions.

Experiment	Stimulus condition	Reaction time in ms (SE)	Hit rate in % (SE)
1	A (with noise)	441 (21)	96 (0.5)
	V	515 (18)	92 (1.5)
	AV	513 (23)	94 (0.7)
3	A (no-noise)	526 (38)	93 (1.2)
	AV	458 (27)	96 (0.8)

Comparison between the AV and the sum of the corresponding A and V responses to frequent syllables

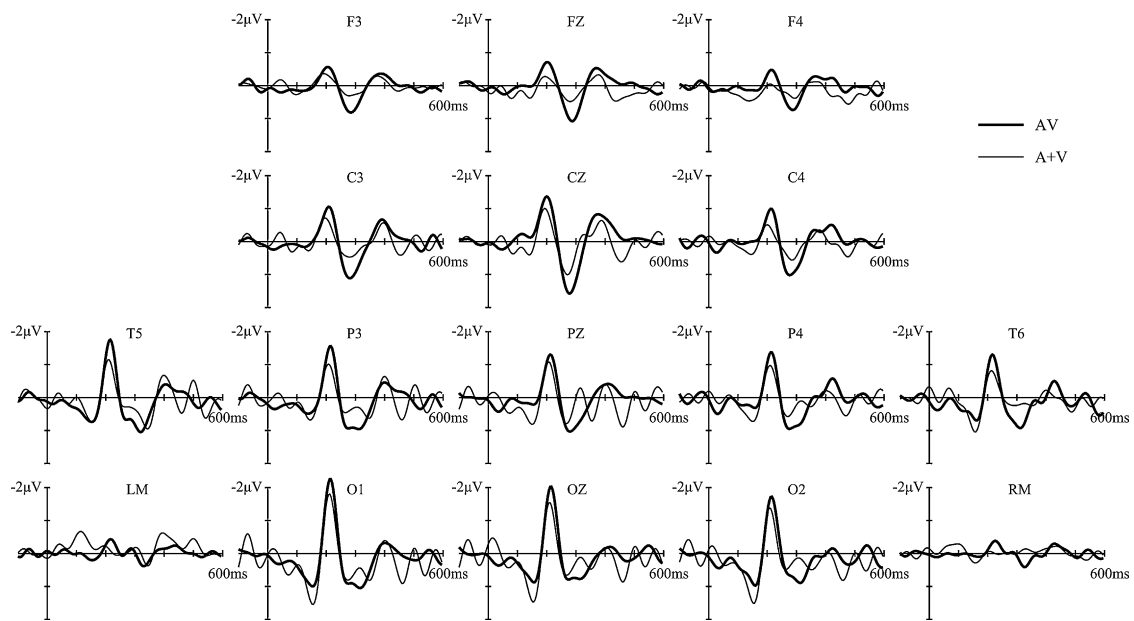


Fig. 2. Experiment 1. Group-averaged ($N = 12$) ERP responses elicited by frequent (control) AV (thick solid line) 'ka' syllables compared with the sum of the corresponding responses elicited by A and V stimuli (marked A + V, thin solid line) at all recording electrodes.

Table 2

Experiment 1: group-averaged ($N = 12$) ERP amplitudes for frequent (control) 'ka' syllables in the AV condition compared with the sum of the corresponding amplitudes in the A and V control conditions (A + V) from two measurement intervals at Fz, Cz, Pz, and Oz.

	Amplitude in μV (SE)			
	188–212 ms		264–288 ms	
	AV	A+V	AV	A+V
Fz	-0.67 (0.18)	-0.25 (0.36)	0.95 (0.29)	0.46 (0.26)
Cz	-1.29 (0.29)	-0.90 (0.33)	1.51 (0.37)	0.93 (0.34)
Pz	-1.00 (0.31)	-0.90 (0.29)	0.98 (0.32)	0.67 (0.39)
Oz	-1.39 (0.36)	-1.16 (0.38)	0.76 (0.48)	0.59 (0.41)

visual ERP components can be discerned showing the largest amplitude over occipital areas. Fig. 2 shows that the control AV speech stimuli elicited a higher-amplitude negative wave in the 188–212 ms interval compared with the sum of the A- and V-condition responses (termed A + V) ($F(1,11) = 6.13$, $p < 0.05$, $\eta^2 = 0.36$; repeated measures ANOVA with factors of stimulus modality [AV vs. A + V] and electrode [F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, O1, Oz, O2]; the electrode main effect was also significant: $F(11,121) = 16.84$, $\epsilon = 0.15$, $p < 0.001$, $\eta^2 = 0.62$; Table 2, left). There was also a statistical tendency for a higher-amplitude positive response in the 264–288 ms interval for AV stimuli as compared with A + V sum ($F(1,11) = 3.93$; $p = 0.07$, $\eta^2 = 0.26$; Table 2, right).

Fig. 3 shows the deviant-minus-control difference waveforms for the three stimulus conditions. No significant difference between the deviant and control responses was observed for the A stimuli. For the V stimuli, the occipital deviant-stimulus response was more positive than the control response in the 308–332 ms latency range (peak: 320 ms; the ANOVA [deviant vs. control \times O1 vs. Oz vs. O2]

Deviant minus-control difference waveforms in the A, V, and AV conditions

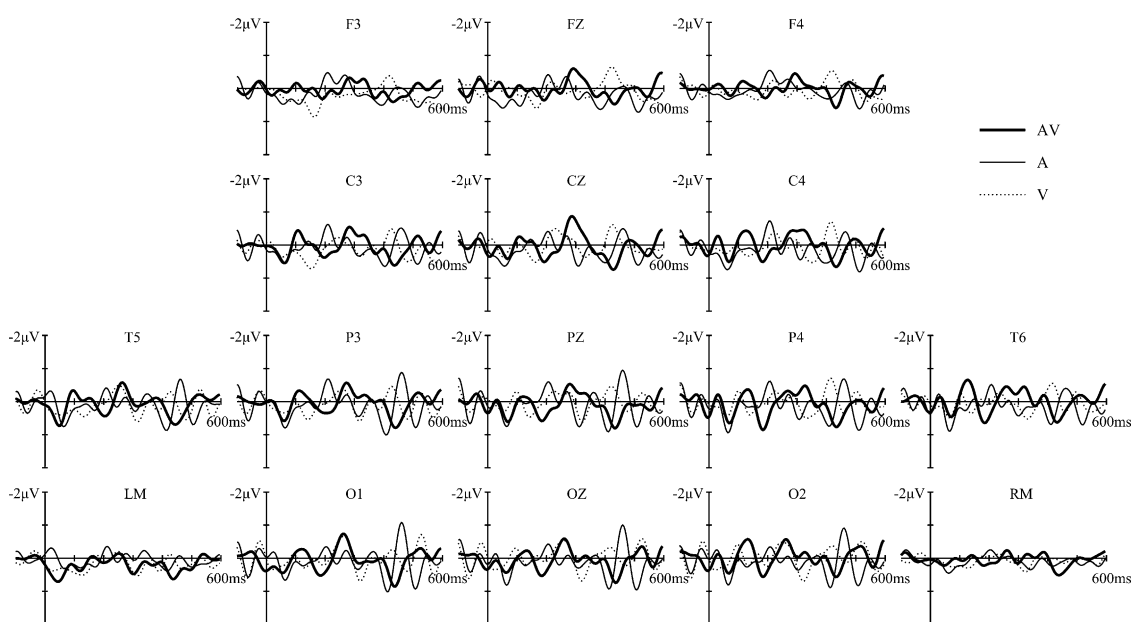


Fig. 3. Experiment 1. Group-averaged ($N = 12$) deviant-minus-control difference waveforms for A (thin solid line), V (dotted line), and AV (thick solid line) 'ka' syllables at all recording electrodes.

Table 3
Experiment 1: group-averaged ($N=12$) mean amplitudes (24-ms windows) for the A, V, AV, and A+V difference waveforms at Fz, Cz, Pz, and Oz at the 260, 288, 320, and 428 ms peaks.

Latency (ms)	Electrode	Amplitude in μV (SE)			
		A	V	A+V	AV
260	Fz	-0.26 (0.21)	0.12 (0.30)	-0.14 (0.32)	-0.02 (0.18)
	Cz	-0.14 (0.21)	-0.26 (0.36)	-0.40 (0.48)	-0.26 (0.25)
	Pz	-0.17 (0.32)	-0.22 (0.39)	-0.38 (0.52)	-0.24 (0.27)
	Oz	0.04 (0.39)	-0.51 (0.37)	-0.47 (0.55)	-0.54 (0.23)
288	Fz	0.01 (0.27)	0.06 (0.25)	0.06 (0.24)	-0.54 (0.30)
	Cz	0.16 (0.23)	-0.15 (0.33)	0.01 (0.43)	-0.81 (0.30)
	Pz	0.13 (0.27)	-0.12 (0.30)	0.01 (0.48)	-0.46 (0.28)
	Oz	0.07 (0.25)	-0.09 (0.28)	-0.02 (0.46)	-0.13 (0.28)
320	Fz	0.24 (0.20)	0.16 (0.20)	0.40 (0.25)	-0.36 (0.31)
	Cz	0.47 (0.22)	0.26 (0.29)	0.73 (0.40)	-0.42 (0.36)
	Pz	0.61 (0.26)	0.54 (0.32)	1.15 (0.45)	-0.25 (0.35)
	Oz	0.23 (0.21)	0.62 (0.27)	0.86 (0.36)	0.01 (0.32)
428	Fz	0.11 (0.22)	-0.59 (0.29)	-0.48 (0.33)	0.43 (0.28)
	Cz	0.14 (0.24)	-0.55 (0.36)	-0.40 (0.38)	0.69 (0.339)
	Pz	0.37 (0.32)	-0.40 (0.30)	-0.02 (0.39)	0.71 (0.31)
	Oz	0.46 (0.41)	-0.25 (0.27)	0.22 (0.48)	0.61 (0.25)

showed a stimulus type $F(1,11) = 5.62, p < 0.05, \eta^2 = 0.34$ and an electrode main effect $F(2,22) = 7.00; p < 0.01; \epsilon = 0.82, \eta^2 = 0.39$; see Table 3). For the AV stimuli, deviants elicited a significant negative ERP difference in the 248–272 ms interval over occipital electrodes (peak: 260 ms; the ANOVA [deviant vs. control \times O1 vs. Oz vs. O2] showed only a stimulus type main effect $F(1,11) = 6.18; p < 0.05; \eta^2 = 0.36$; see Table 3). Also a significant fronto-centrally negative AV difference wave was elicited in the 276–300 ms interval (peak: 288 ms; the ANOVA [deviant vs. control \times C3 vs. Cz vs. C4] showed a stimulus type $F(1,11) = 4.83; p = 0.05; \eta^2 = 0.30$, an electrode main effect $F(2,22) = 4.89; p < 0.05; \epsilon = 0.90, \eta^2 = 0.31$, and an interaction between these two factors $F(2,22) = 5.08; p < 0.05; \epsilon = 0.70, \eta^2 = 0.32$; see Table 3). The interaction was due to significantly higher Cz than C3 or C4 amplitudes for deviants ($p < 0.01$ at least, Tukey HSD post hoc tests), but not for controls. Finally, a late centrally maximal positive difference was observed for AV stimuli in the 416–440 ms latency range (peak: 428 ms; the ANOVA [deviant vs. control \times C3 vs. Cz vs. C4] showed only a stimulus type main effect $F(1,11) = 4.74; p = 0.05; \eta^2 = 0.30$; see Table 3).

Fig. 4 compares the deviant-minus-control difference waveforms measured for AV stimuli with the sum of the corresponding waveforms for A and V stimuli (A+V). The earlier occipital negative difference (248–272 ms) found for the AV stimuli corresponds to a similar (although not significant) difference observed for the V

stimuli and thus there was no significant difference between the A+V and AV amplitudes (ANOVA [AV vs. A+V \times O1 vs. Oz vs. O2]; for the stimulus modality factor, $F(1,11) = 0.0048$; see Table 3). However, the fronto-central negative difference (296–320 ms) and the late central positive difference (424–448 ms) found for the AV stimuli had no similar-sized counterpart in either in the A or the V stimulus responses. Therefore, the AV difference waveforms significantly differed from the sum of the A and V waveforms in these two latency ranges (ANOVA [AV vs. A+V \times C3 vs. Cz vs. C4]: significant stimulus modality main effect $F(1,11) = 4.91; p < 0.05; \eta^2 = 0.31$ for the negative and $F(1,11) = 12.97; p < 0.01; \eta^2 = 0.54$ for the positive difference peak, respectively). A subsequent topographical comparison (ANOVA [AV vs. A+V \times Fz vs. Cz vs. Pz vs. Oz]) of the normalized (McCarthy and Wood, 1985) amplitudes in the 296–320 ms showed significant interaction between the two factors ($F(3,33) = 3.53, \epsilon = 0.65, p < 0.05, \eta^2 = 0.24$) as well as an electrode main effect ($F(3,33) = 3.96, \epsilon = 0.65, p < 0.05, \eta^2 = 0.26$).

2.3. Discussion

The very low d' values in the auditory-only condition demonstrated that the noise had a strong masking effect on the

Comparison between the AV and the sum of the corresponding A and V difference waveforms

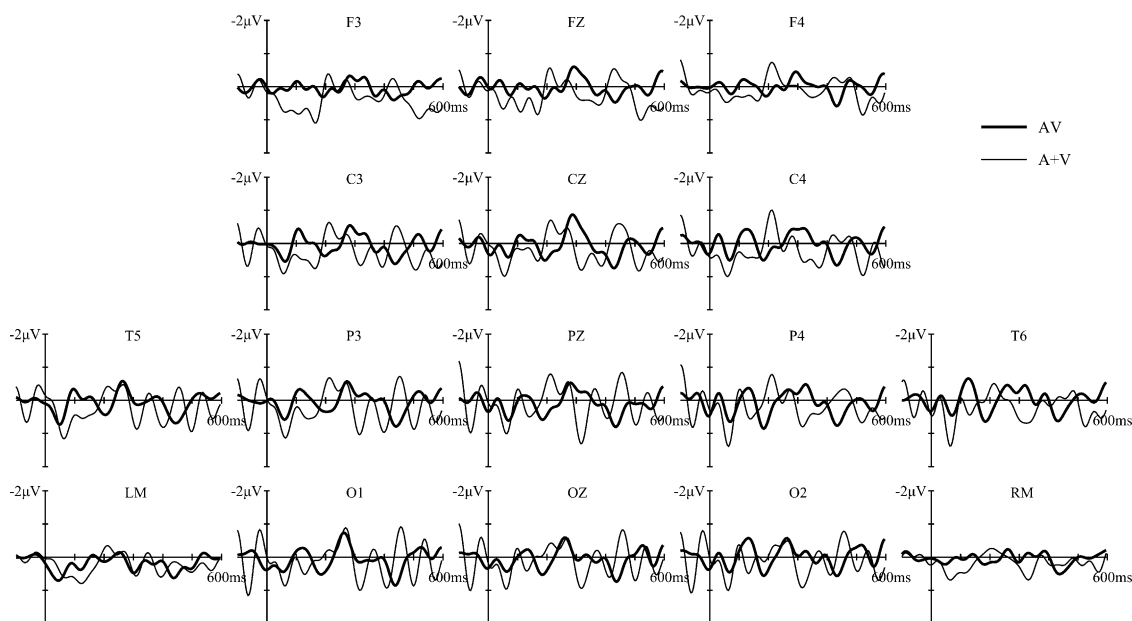


Fig. 4. Experiment 1. Group-averaged ($N = 12$) AV deviant-minus-control difference waveforms (thick solid line) compared with the sum of the differences obtained for A and V 'ka' syllables (marked A+V, thin solid line) at all recording electrodes.

speech sounds. Performance in detecting rare deviant syllables amongst a repeating frequent one as well as classifying the same two syllables presented equiprobably within the sequences was significantly higher for AV than for auditory-only speech in noise. These results confirm the conclusion of previous studies that visual speech cues improve the intelligibility of speech in noisy environments. However, good performance in the visual-only sequences (different group) brings up the possibility that participants may have relied on visual speech cues in the AV condition. Therefore, an alternative account of the results suggests that the auditory stimuli only acted as noise for visual perception. This view is supported by the current results showing that, although measured in different groups, discrimination and identification performance based on visual cues alone were better than the measures obtained in the corresponding AV tasks (see Fig. 1). That is, on this alternative, despite our instructions, participants may have primarily based their responses on visual information (when available) and this task was easier in the absence than in the presence of auditory stimulation. Thus this alternative explanation questions our claim that visual information improved the quality of auditory representations, suggesting instead that, when given the chance, participants based their perceptual judgments on visual information. Experiment 2 (see later) was designed to directly test this alternative explanation.

In accordance with the very low performance for auditory-only speech stimuli embedded in noise (A condition), no significant discriminative ERP response was elicited in the corresponding EEG experiment. Furthermore, frequent auditory-only syllables elicited rather low-amplitude ERP responses, suggesting that these speech sounds could not be easily distinguished from the background noise. The absence of a significant N1 response specifically suggests that no sharp syllable onset was detected. Furthermore, attention directed to a visual task may have further reduced the early ERP responses, since primary auditory cortical circuits are sensitive to the modality of selective attention (Fritz et al., 2003) and specifically, the N1 response is significantly modulated by attention (Hillyard and Picton, 1979). On the other hand, the significant later response, which also characterizes the response to the syllables in the absence of noise, demonstrates that the syllables were processed to some degree. AV standards elicited a supra-additive negativity peaking at ca. 200 ms from stimulus onset, which coincides with the visual N1 wave. This result may be analogous to the increased auditory N1 observed by Pourtois et al. (2000), since no significant auditory N1 was elicited in the current A condition. Deviant V syllables elicited a late occipital difference response, and possibly also an earlier negative difference, which, however, did not reach significance (except for a tendency at O1, peak: 260 ms; $t(11) = 1.81$, $p < 0.1$). Deviant AV syllables also elicited this negative difference followed by a centrally distributed negative and a subsequent positive difference wave. The early occipital negative difference wave probably corresponds to vMMN, whereas the late positive difference to P3(a).

Most importantly, the centrally maximal negative (296–320 ms) and the late central positive difference (424–448 ms) found for AV speech deviants had no counterpart in the A and V stimulus conditions. Therefore, these ERP components could not be modeled by summing the corresponding difference waves and neither their latency nor their scalp distribution matched those of the unimodal responses. Thus, at least the first of these components may represent a genuine audiovisual interaction in processing deviant speech stimuli. In turn, the later P3-like response probably reflects further processing following the detection of a deviant (for reviews, see Escera et al., 2000; Friedman et al., 2001; Polich and Criado, 2006). The latency range of the centrally maximal negativity is compatible with that of the interactions between auditory and visual speech found in some

previous studies (Kaiser et al., 2005; Klucharev et al., 2003; Möttönen et al., 2004; van Wassenhove et al., 2005) and the component would be classified as speech-specific according to Klucharev et al. (2003), whereas modulations of earlier ERP components may reflect shared temporal or spatial features of the acoustic and visual stimulation, such as the supra-additivity observed for the visual N1 to AV standards. In contrast to the ERP components showing subadditivity between unimodal and multi-modal presentations (see e.g., Besle et al., 2004; van Wassenhove et al., 2005), the current component has no closely corresponding unimodal equivalent. We suggest that this component may be a mismatch negativity based on multi-modal information. That is, this component reflects the violation of a regularity, the representation of which is based on both auditory and visual information. Accordingly, this component could be termed audiovisual MMN (avMMN). Supporting the assumption that the observed negative component is analogous to the unimodal MMN components is the elicitation of a following positive difference waveform, which could be identified as a P3(a) component. The P3(a) often follows the auditory and visual MMN (Näätänen, 1990; Czigler, 2007) and it is thought to reflect further processing of deviant stimulus events (Escera et al., 2000; Friedman et al., 2001). P3(a) was elicited in the V and AV conditions of the current experiment, and it peaked later in the AV than in the V stimulus condition, just as the avMMN observed in the AV condition peaked later than the assumed vMMN in the V condition. Finally, avMMN was preceded by a somewhat increased positivity in the AV standard-stimulus responses, a phenomenon that has often been observed for the auditory MMN and has been suggested to reflect the formation/reinforcement of implicit memory representation of the standard (repetition positivity, RP; Baldeweg, 2006; Haenschel et al., 2005).

Our results thus suggest for the first time that the perceptual advantages observed for congruent AV compared to auditory-only speech stem from the formation of an enhanced representation for speech sounds. This is because MMN-type responses are based on the formation of a memory representation of the regular stimulation (for recent reviews of the memory pre-requisites of auditory and visual MMN, see Winkler, 2007 and Czigler, 2007, respectively). The assumption of an improved speech representation is compatible with the improvement of speech identification by visual speech cues (e.g., Grant and Seitz, 2000; Schwartz et al., 2004) as well as with the interpretation of MMN results obtained when frequent congruent AV speech stimuli were contrasted with infrequent incongruent ones (evoking the McGurk effect) by changing only the visual stimulus component (Colin et al., 2002, 2004; Fingelkurts et al., 2003; Kaiser et al., 2005; Möttönen et al., 2002; Saint-Amour et al., 2007; Sams et al., 1991; Trejo et al., 2000). The late emergence of the avMMN (later than either the auditory or the visual MMN would be expected to appear) however contrasts results showing early (50–200 ms) visual effects on auditory processing (Lebib et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005). This may suggest that the formation of a multi-modal representation takes longer than that of the unimodal ones. Alternatively, it is possible that discriminability of the two syllables was still low even in the AV condition, an effect known to delay the MMN response (e.g., Näätänen, 1990; Schröger and Winkler, 1995). Favoring the second alternative are the findings showing that the current vMMN preceded the avMMN only by ca. 50 ms and no auditory-only MMN was elicited at all. Furthermore, deviance detection and identification performance level for the current AV syllables in noise was significantly lower than that for the auditory stimuli without noise.

There are, however, some alternative explanations to be considered. Is it possible that the centrally negative difference found only for AV stimuli reflected the results of attentive

discrimination? That is, although participants were not instructed to identify or discriminate the syllables, they still attempted to do so. This assumption is supported by the result showing that the presence of visual speech cues affected performance in the primary visual detection task compared with the still face background presented in the auditory-only stimulus blocks. This finding suggests that facial movements either masked the target change to some degree or drew capacity away from performing the change detection task. However, no difference in performing the visual detection task was found between the V and AV sequences. Therefore, the ERP components observed only for AV stimuli could not have been caused by attentive visual discrimination, which could also be successfully performed in the V stimulus condition. Furthermore, participants were not informed that speech sounds were embedded in the continuous noise or that more than one type of spoken syllables would be presented within the sequences and the active task conditions were conducted always after the EEG recording. Thus it is highly unlikely that participants performed an active sound discrimination task, which was quite difficult even in the AV stimulus condition (see behavioral results) while they also performed the visual change detection task as they were instructed.

Another alternative explanation of the current results with respect to the formation of audiovisual representations suggests that the performance improvement as well as the ERP components found for AV speech were not specifically related to the visual speech cues, rather the synchronous periodic visual stimulation enhanced sound processing by informing the auditory system about the start and progress of the speech sounds. That is, the visual speech stimuli simply helped participants in timing the sound analysis processes. Previous behavioral studies showed no improvement of speech intelligibility by non-speech visual timing cues (Kim and Davis, 2003b; Schwartz et al., 2004; Summerfield, 1979). Because these results do not apply to the current ERP experiment in which the speech stimuli were task-irrelevant, in a follow-up experiment, we tested whether the current ERP effects could be explained by the visual stimuli providing timing information to the auditory system. Furthermore, we also wished to determine the timing of auditory MMN for the speech stimuli used in Experiment 1 in the absence of noise. These questions were addressed in Experiment 3.

3. Experiment 2

3.1. Methods

3.1.1. Subjects

Eleven participants (4 male, age 19–22 years, mean 20.6), none of whom participated in Experiment 1, were recruited through a student part-time job agency. They received modest financial compensation for their participation. The study has been approved by the Ethical Committee of the Institute for Psychology, Hungarian Academy of Sciences. Participants signed an informed consent form after the aims and procedures of the study were explained to them, before starting the experiment. Participants were pre-selected on the basis of a clinical audiometry with the same criteria as were used for Experiment 1. All participants had normal or corrected to normal vision. One participant's data was dropped from the analyses, because of above-criterion false alarms on catch trials.

3.1.2. Procedure and analysis

The stimulation in Experiment 2 was identical to the deviance-detection condition of Experiment 1 with only the V and AV conditions repeated. Just as in Experiment 1, the participants were asked to detect infrequent 'ka' syllables amongst frequent 'ma' syllables. The AV condition was delivered two times with different instructions: detect the infrequent syllables based on (1) visual information (AV-visual) and (2) auditory information (AV-auditory). The AV sequences included the same catch trials (auditory "ma" and visual "ka") as the corresponding sequences of Experiment 1. The catch trial criterion (same as in Experiment 1) was, however, only used in the AV-auditory condition, because in the AV-visual condition, the catch trials were correct targets. All other aspects of the data analysis were identical to the corresponding procedures in Experiment 1. One-way dependent-measures ANOVA (Condition [V vs. AV-visual vs. AV-auditory]) was

used to compare the d' (discrimination sensitivity) measures across the different conditions.

3.2. Results and discussion

Almost identical, high d' values were obtained in the V and AV-visual conditions (3.45 ± 0.49 and 3.48 ± 0.65 for the V and AV-visual conditions, respectively) with a much lower value for the AV-auditory condition (2.42 ± 0.70). This observation was supported with the significant result of the ANOVA ($F(2,18) = 9.58$; $p < 0.002$; $\eta^2 = 0.52$) and the subsequent post hoc comparisons (Tukey HSD tests showed $p < 0.01$ difference between the AV-auditory and the other two conditions, but no difference between the V and the AV-visual conditions). These results match those obtained in Experiment 1 and the pilot V condition. They suggest that when asked to base their judgment on auditory information, participants indeed attempt to do so. Should they have used a visual strategy in the AV-auditory condition, their performance would have equaled that of the AV-visual condition (which explicitly asked them to use the visual information). Furthermore, the lack of a significant difference between the V and the AV-visual performance results show that the presence of auditory stimuli (noise and syllables) did not significantly affect visual detection (in fact, the AV-visual results are numerically slightly higher than those in the V condition). In summary, the performance increase from the A (noise) to the AV (noise) condition in Experiment 1 can be safely attributed to the effect of visual information on auditory deviance detection and, as a consequence, to the underlying memory representation.

4. Experiment 3

4.1. Methods

4.1.1. Subjects

Ten participants (4 male, age 19–23 years, mean 20.9), none of whom participated in Experiment 1 or 2, were recruited through a student part-time job agency. They received modest financial compensation for their participation. The study has been approved by the Ethical Committee of the Institute for Psychology, Hungarian Academy of Sciences. Participants signed an informed consent form after the aims and procedures of the study were explained to them, before starting the experiment. Participants were pre-selected on the basis of a clinical audiometry with the same criteria as were used for Experiment 1. All participants had normal or corrected to normal vision.

4.1.2. Procedure, recording, and analysis

The stimulation in the first part of Experiment 3 was identical to that in the AV condition of Experiment 1, except that instead of the speaker's face, a circle with a 3.4° diameter was presented on the screen. During the initial 15 frames on which the speech-related facial movements appeared in the original video recording, the circle expanded to 4.8° in diameter then shrunk frame-by-frame back to 3.4° during the following 14 frames. The visual stimulation was the same for all trials (standards and deviants alike), since its role was to mark the timing of the speech sounds without specifying the identity of the syllables. Six experimental and one control stimulus block were delivered, the latter in the third serial position. In the control stimulus block, the role of the two syllables was exchanged (same as in Experiment 1). All other parameters, including those of the visual change detection task were identical to the corresponding parameters of the AV condition of Experiment 1. In the second part of Experiment 3, the A condition from Experiment 1 was repeated without the background noise. The screen showed the same still face and the task was identical to that employed in Experiment 1. Six experimental and one control stimulus block were presented, the latter in the third sequential position. The two parts of the experimental session was separated by a longer break, during which the participant could leave the experimental chamber. Otherwise, short breaks were inserted between successive stimulus blocks unless the participant asked for a longer rest period.

Data acquisition and analysis procedures were identical to those described for Experiment 1.

4.2. Results and discussion

4.2.1. Behavioral results

Subjects showed somewhat better performance and shorter response times to visual target changes with the simpler but animated as compared with the more

ERP responses to frequent and infrequent auditory syllables with synchronous visual non-speech animation

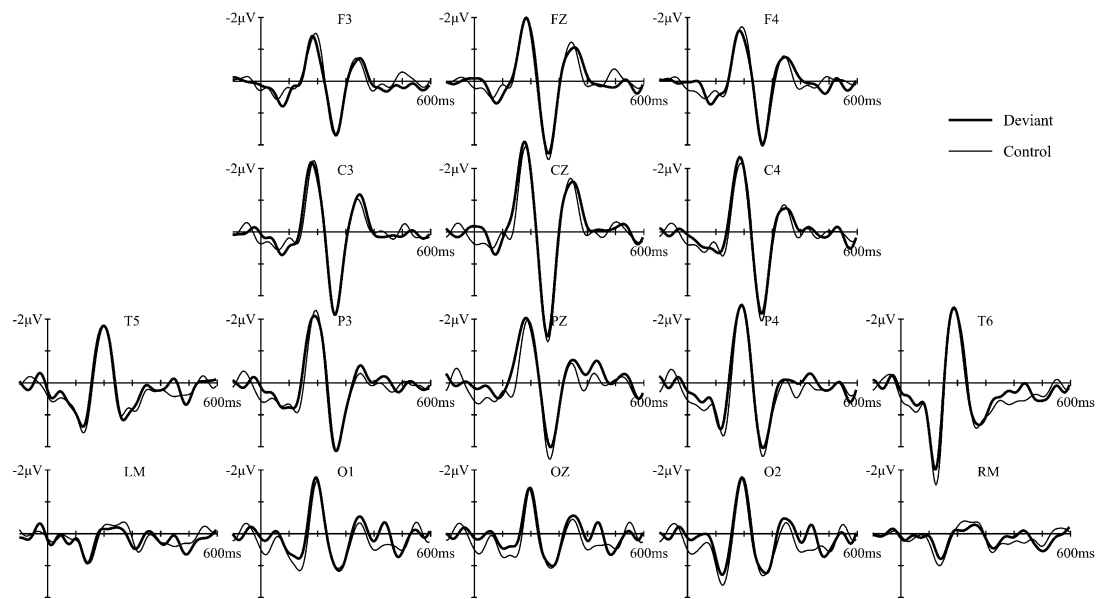


Fig. 5. Experiment 3. Group-averaged ($N = 10$) ERP responses elicited by frequent (control, thin solid line) and deviant (thick solid line) syllables in the AV (animated circle) condition at all recording electrodes.

complex but constant visual display (Table 1; $t(9) = 3.56, p < 0.01$ for reaction times and $t(9) = 2.15, p = 0.06$ for hit rates).

4.2.2. ERP results

Fig. 5 shows that although both standard and deviant AV (animated circle) stimuli elicited clear ERP responses, no differential ERP response was elicited by deviants. This result shows that the differential ERP response found in Experiment 1 cannot be explained by visual stimuli providing timing cues for the auditory system. It should be noted that because the visual stimulation did not differ between the standard and deviant stimuli, no discriminative ERP component based on integrated audiovisual memory could be expected in this condition. The motivation for this condition was restricted to clarify whether or not the apparently AV-specific

component found in the main experiment did indeed require congruent audiovisual information.

Fig. 6 shows that in the absence of noise, infrequent spoken “ka” syllables elicited a large fronto-central MMN response peaking at 164 ms from sound onset. MMN appeared with reversed polarity at the mastoid leads. The ANOVA [deviant vs. control \times C3 vs. Cz vs. C4] showed a stimulus main effect: $F(1,9) = 8.02, p < 0.05, \eta^2 = 0.47$; an electrode main effect: $F(2,18) = 4.67, \epsilon = 0.85, p < 0.05, \eta^2 = 0.34$; and an interaction: $F(2,18) = 4.24, \epsilon = 0.79, p < 0.05, \eta^2 = 0.32$; with post hoc Tukey HSD tests showing that deviant and standard amplitudes significantly differed at all electrodes and that the deviant amplitude at C4 was significantly lower than at Cz. The [deviant vs. control \times Lm vs. Rm] ANOVA showed a stimulus main effect: $F(1,9) = 7.87, p < 0.05, \eta^2 = 0.47$. In full accordance with the expectations based on previous MMN results, in the absence of the masking noise, auditory deviant

ERP responses to frequent and infrequent auditory only syllables with no background noise

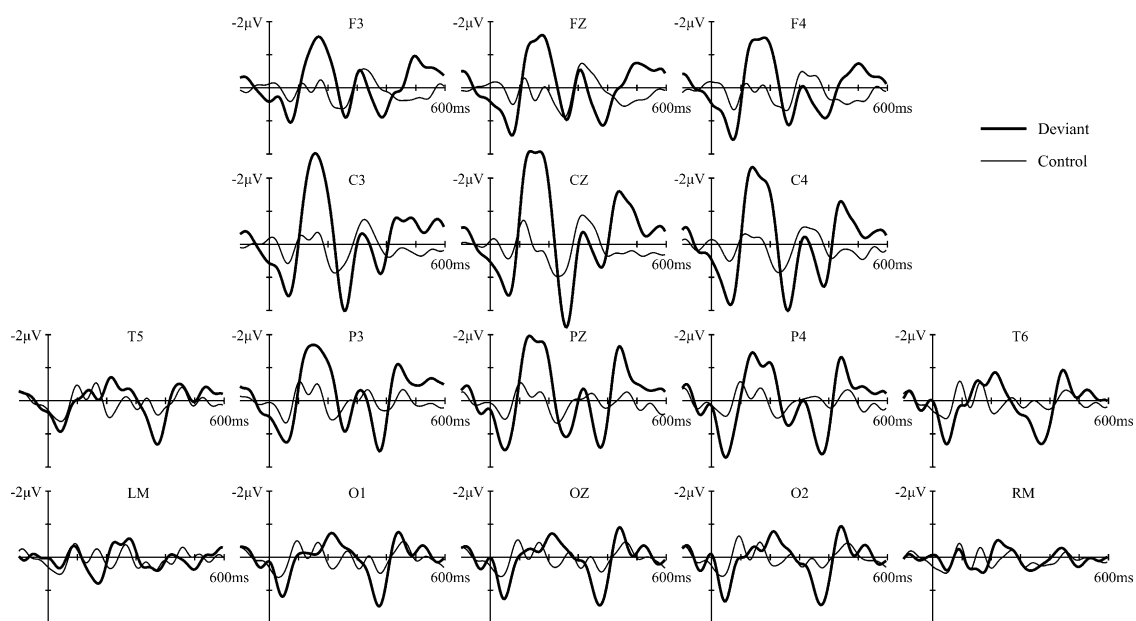


Fig. 6. Experiment 3. Group-averaged ($N = 10$) ERP responses elicited by frequent (control, thin solid line) and deviant (thick solid line) syllables in the A (no noise) condition at all recording electrodes.

syllables elicited the auditory MMN with a typical scalp distribution and peak latency, which is much earlier than that of the avMMN observed in Experiment 1. However, as was argued previously, in the absence of noise, the two syllables were easy to discriminate, which can account for the latency difference. Previous studies showed that background sounds can increase the MMN peak latency (Martin et al., 1999; McArthur et al., 2003; Muller-Gass et al., 2001) and the reaction time of detecting deviants (e.g., Novitski et al., 2006).

5. General discussion

The current results revealed the existence of an ERP response elicited by infrequent congruent audiovisual syllables delivered amongst frequent repetitions of a different congruent audiovisual syllable while participants performed a visual detection task unrelated to the speech stimuli. The component was not elicited by the frequent syllables. Elicitation of a component under these conditions makes the observed component similar to the well-known MMN response. Therefore, we tentatively termed this component avMMN (audiovisual MMN). The latency as well as the scalp distribution of this avMMN differs from those of both corresponding unimodal MMN responses as well as from the sum of these responses. It has also been established that when the visual components of the congruent AV syllables were exchanged for a non-speech stimulus that preserved the timing information but did not differentiate between standards and deviants, no ERP component similar to avMMN was elicited. This result does not, however, mean that the avMMN is specific to speech stimuli. In Experiment 3, visual stimulation was the same for standards and deviants (its role being to facilitate the timing of auditory analysis processes). Therefore, the results of Experiment 3 provide evidence only for the claim that the elicitation of avMMN requires concurrent discriminative auditory and visual information, not that this component would be specific to speech. Finally, behavioral and ERP results together suggested that although the role of attention in avMMN elicitation cannot be ruled out, attention to visual speech cues is not a sufficient pre-requisite of avMMN elicitation. In summary, avMMN appears to be similar to the previously observed MMN responses, but it is based on integrated auditory and visual information.

The elicitation of the avMMN and a possibly supra-additive repetition positivity (the latter by the standard stimuli) suggest that task-irrelevant auditory and visual (speech) information is encoded in an integrated memory representation in the brain and it is then used for detecting deviations in the sensory input. The latter conclusion is further supported by the fully compatible behavioral results measured in the active conditions of the current study. Deviance detection is an essential function of the human sensory systems, because deviance, defined as a violation of some detected regularity, represents new information for the organism (as opposed to change, which can be regular; for a discussion of this issue, see Winkler, 2003). New information is potentially relevant for survival and goal-directed behavior as well as for updating the internal models describing the current environment (Winkler, 2007; Winkler et al., 1996). The finding that avMMN was elicited for task-irrelevant stimuli suggests that the formation of such representations requires no or only small attentional capacities and that the resulting memory representation may be of implicit nature. Deviance detection based on implicit memory representations is advantageous for the organism, because detecting vital information as well as maintaining a valid model of the environment are important functions for survival irrespective of the current goals of the organism. For example, identifying and tracking animals in the environment on the basis of partial visual and auditory cues without focusing on each of them can leave more attentional capacity for negotiating a difficult terrain.

Although early automatic inter-modal integration of speech cues would fit with the goals of deviance detection, the current

results cannot decide this question. The latency of the observed avMMN response was longer than those of the corresponding unimodal MMNs. This may be a sign of a hierarchical organization, in which integration across modalities occurs following modality-specific processing. However, as was argued before, the relatively long latency of the avMMN response may also reflect poor discriminability of the frequent and infrequent syllables caused by the background noise. Further studies using the current method will target the generality vs. speech-specificity issue and the question of early (and possibly automatic) formation of integrated audiovisual memory representations. However, the finding of an ERP indexing an implicit audio-visual deviance detection process allows one to study audiovisual representations of speech without confounds stemming from the strategies employed by experimental participants.

Acknowledgements

This work was supported by the Hungarian National Research Fund (OTKA T048383). We thank Ms. Zsuzsanna D'Albini, Kinga Gyimesi, and Gabriella Pályfi for collecting the data.

References

- Alsius, A., Navarra, J., Campbell, R., Soto-Faraco, S., 2005. Audiovisual integration of speech falters under high attention demands. *Current Biology* 15 (9), 839–843.
- Assmann, P.F., Summerfield, A.Q., 2004. The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (Eds.), *Speech Processing in the Auditory System*. Springer Handbook of Auditory Research, vol. 14. Springer Verlag, New York, pp. 231–308.
- Baldeweg, T., 2006. Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences* 10 (3), 93–94.
- Bernstein, L.E., Auer Jr., E.T., Takayanagi, S., 2004. Auditory speech detection in noise enhanced by lipreading. *Speech Communication* 44 (1–4), 5–18.
- Bertelson, P., Vroomen, J., de Gelder, B., 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science* 14 (6), 592–597.
- 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 (8), 2225–2234.
- Besle, J., Fort, A., Giard, M.-H., 2005. Is the auditory sensory memory sensitive to visual information? *Experimental Brain Research* 166 (3–4), 337–344.
- Bregman, A.S., 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA.
- Broadbent, D.E., 1958. *Perception and Communication*. Pergamon Press, New York.
- Bushara, K.O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., Hallett, M., 2003. Neural correlates of cross-modal binding. *Nature Neuroscience* 6 (2), 190–195.
- 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 (17), 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 (5), 805–816.
- Calvert, G.A., Campbell, R., 2003. Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of Cognitive Neuroscience* 15 (1), 57–70.
- 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.
- 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.
- Coltheart, M., 1984. Sensory memory—a tutorial review. In: Bouman, H., Bouwhuis, D.G. (Eds.), *Attention & Performance X: Control of Language Processes*. Erlbaum, Hillsdale, NJ, pp. 259–285.
- Czigler, I., 2007. Visual mismatch negativity: violation of non-attended environmental regulations. *Journal of Psychophysiology* 21 (3–4), 224–230.
- Davis, C., Kim, J., 2004. Audio-visual interactions with intact clearly audible speech. *Quarterly Journal of Experimental Psychology A* 57 (6), 1103–1121.
- Denham, S.L., Winkler, I., 2006. The role of predictive models in the formation of auditory streams. *Journal of Neurophysiology—Paris* 100 (1–3), 154–170.
- Dixon, N.F., Spitz, L., 1980. The detection of auditory visual desynchrony. *Perception* 9 (6), 719–721.
- Erber, N.P., 1975. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders* 40 (4), 481–492.
- Escera, C., Alho, K., Schröger, E., Winkler, I., 2000. Involuntary attention and distractibility as evaluated with event related brain potentials. *Audiology and Neuro-Otology* 5 (3–4), 151–166.

- Fernandez-Duque, D., Thornton, I.M., 2003. Explicit mechanisms do not account for implicit localization and identification of change: an empirical reply to Mitroff et al. (2002). *Journal of Experimental Psychology: Human Perception and Performance* 29 (5), 846–858.
- Fingelkurts, A.A., Fingelkurts, A.A., Krause, C.M., Möttönen, R., Sams, M., 2003. Cortical operational synchrony during audio-visual speech integration. *Brain and Language* 85 (2), 297–312.
- Fodor, J.A., 1983. *Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA.
- Fort, A., Giard, M.-H., 2004. Multiple electrophysiological mechanisms of audio-visual integration in human perception. In: Calvert, G., Spence, C., Stein, B. (Eds.), *The Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, pp. 503–514.
- Friedman, D., Cycowicz, Y.M., Gaeta, H., 2001. The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience and Biobehavioral Reviews* 25 (4), 355–373.
- Fritz, J., Shamma, S., Elhilali, M., Klein, D., 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience* 6 (11), 1216–1223.
- Fu, K.M., Johnston, T.A., Shah, A.S., Arnold, L., Smiley, J., Hackett, T.A., Garraghty, P.E., Schroeder, C.E., 2003. Auditory cortical neurons respond to somatosensory stimulation. *The Journal of Neuroscience* 23 (20), 7510–7515.
- Garrido, M., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009. The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology* 120 (3), 453–463.
- Grant, K.W., 2001. The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America* 109 (5 Pt 1), 2272–2275.
- Grant, K.W., Seitz, P., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America* 108 (3 Pt 1), 1197–1208.
- Haenschel, C., Vernon, D.J., Dwivedi, P., Gruzeliier, J.H., Baldeweg, T., 2005. Event-related brain potential correlates of human auditory sensory memory-trace formation. *The Journal of Neuroscience* 25 (45), 10494–10501.
- Helfer, K.S., Freyman, R.L., 2005. The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America* 117 (2), 842–849.
- Hillyard, S.A., Picton, T.W., 1979. Event-related potentials and selective information processing in man. In: Desmedt, J.E. (Ed.), *Progress in Clinical Neurophysiology*, vol. 6. Cognitive Components in Cerebral Event-related Potentials and Selective Attention. Karger, Basel, pp. 1–52.
- Jääskeläinen, I.P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., Tarnanen, I., Sams, M., 2004. Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport* 15 (18), 2741–2744.
- Jones, J.A., Callan, D.E., 2003. Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport* 14 (8), 1129–1133.
- Jones, J.A., Jarick, M., 2006. Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research* 174 (3), 588–594.
- Kaiser, J., Hertrich, I., Ackermann, H., Mathiak, K., Lutzenberger, W., 2005. Hearing lips: gamma-band activity during audiovisual speech perception. *Cerebral Cortex* 15 (5), 646–653.
- Kang, E., Lee, D.S., Kang, H., Hwang, C.H., Oh, S.H., Kim, C.S., Chung, J.K., Lee, M.C., 2006. The neural correlates of cross-modal interaction in speech perception during a semantic decision task on sentences: a PET study. *Neuroimage* 32 (1), 423–431.
- Kim, J., Davis, C., 2003a. Hearing foreign voices: does knowing what is said affect masked visual speech detection? *Perception* 32 (1), 111–120.
- Kim, J., Davis, C., 2003b. Testing the cuing hypothesis for the AV speech detection. In: *Proceedings of AVSP'2003*, St Jorioz, France, pp. 9–12.
- Klucharev, V., Möttönen, R., Sams, M., 2003. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research* 18 (1), 65–75.
- Korzyukov, O., Pflieger, M.E., Wagner, M., Bowyer, S.M., Rosburg, T., Sundaresan, K., Elger, C.E., Boutros, N.N., 2007. Generators of the intracranial P50 response in auditory sensory gating. *Neuroimage* 35 (2), 814–826.
- Kujala, T., Tervaniemi, M., Schröger, E., 2007. The mismatch negativity in cognitive and clinical neuroscience: theoretical and methodological considerations. *Biological Psychology* 74 (1), 1–19.
- Laurienti, P.J., Burdette, J.H., Wallace, M.T., Yen, Y.-F., Field, A.S., Stein, B.E., 2002. Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience* 14 (3), 420–429.
- 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 (3), 185–188.
- Lebib, R., Papo, D., Douiri, A., de Bode, S., Gillon Dowens, M., Baudonniere, P.-M., 2004. Modulations of 'late' event-related brain potentials in humans by dynamic audiovisual speech stimuli. *Neuroscience Letters* 372 (1–2), 74–79.
- Macaluso, E., George, N., Dolan, R., Spence, C., Driver, J., 2004. Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21 (2), 725–732.
- Macmillan, N.A., Creelman, C.D., 1991. *Detection Theory: A User's Guide*. Cambridge University Press, Cambridge.
- Martin, B.A., Kurtzberg, D., Stapells, D.R., 1999. The effects of decreased audibility produced by high-pass noise masking on N1 and the mismatch negativity to speech sounds/ba/and/da/. *Journal of Speech Language & Hearing Research* 42 (2), 271–286.
- McArthur, G.M., Bishop, D.V., Proudfoot, M., 2003. Do video sounds interfere with auditory event-related potentials? *Behavior Research Methods, Instruments, & Computers* 35 (2), 329–333.
- 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.
- McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. *Nature* 264 (5588), 746–748.
- Merikle, P.M., Smilek, D., Eastwood, J.D., 2001. Perception without awareness: perspectives from cognitive psychology. *Cognition* 79 (1–2), 115–134.
- Miller, L.M., D'Esposito, M., 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience* 25 (25), 5884–5893.
- Mitterer, H., 2006. On the causes of compensation for coarticulation: evidence for phonological mediation. *Perception & Psychophysics* 68 (7), 1227–1240.
- Möttönen, R., Krause, C.M., Tiippana, K., Sams, M., 2002. Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research* 13 (3), 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 (2), 112–115.
- Muller-Gass, A., Marcoux, A., Logan, J., Campbell, K.B., 2001. The intensity of masking noise affects the mismatch negativity to speech sounds in human subjects. *Neuroscience Letters* 299 (3), 197–200.
- Näätänen, R., 1990. The role of attention in auditory information-processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences* 13 (2), 201–288.
- Näätänen, R., Gaillard, A.W.K., Mäntylä, S., 1978. Early selective attention effect on evoked potential reinterpreted. *Acta Psychologica* 42 (4), 313–329.
- Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., Spence, C., 2005. Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research* 25 (2), 499–507.
- Neisser, U., 1967. *Cognitive Psychology*. Appleton Century Crofts, New York.
- Novitski, N., Maess, B., Tervaniemi, M., 2006. Frequency specific impairment of automatic pitch change detection by fMRI acoustic noise: an MEG study. *Journal of Neuroscience Methods* 155 (1), 149–159.
- Ojanen, V., Möttönen, R., Pekola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T., Sams, M., 2005. Processing of audiovisual speech in Broca's area. *Neuroimage* 25 (2), 333–338.
- Olson, I.R., Gatensby, J.C., Gore, J.C., 2002. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research* 14 (1), 129–138.
- Polich, J., Criado, J.R., 2006. Neuropsychology and neuropharmacology of the P3a and P3b. *International Journal of Psychophysiology* 60 (2), 172–185.
- Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., Crommelinck, M., 2000. The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport* 11 (6), 1329–1333.
- Reale, R.A., Calvert, G.A., Thesen, T., Jenison, R.L., Kawasaki, H., Oya, H., Howard, M.A., Brugge, J.F., 2007. Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience* 145 (1), 162–184.
- 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.
- Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., Simola, J., 1991. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters* 127 (1), 141–145.
- Sams, M., Möttönen, R., Sihvonen, T., 2005. Seeing and hearing others and oneself talk. *Cognitive Brain Research* 23 (2–3), 429–435.
- Schröger, E., Winkler, I., 1995. Presentation rate and magnitude of stimulus deviance effects on pre-attentive change detection. *Neuroscience Letters* 193 (3), 185–188.
- 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–B78.
- Sekiyama, K., 1991. McGurk effect and incompatibility: a cross-language study on auditory-visual speech perception. *Studies and Essays in Behavioral Sciences and Philosophy*, vol. 14. Kanazawa University, pp. 29–62.
- Skipper, J.I., Nusbaum, H.C., Small, S.L., 2005. Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25 (1), 76–89.
- Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L., 2007. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex* 17 (10), 2387–2399.
- Sommers, M.S., Tye-Murray, N., Spehar, B., 2005. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear & Hearing* 26 (3), 263–275.
- Soto-Faraco, S., Navarra, J., Alsius, A., 2004. Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 92 (3), B13–B23.
- Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26 (2), 212–215.
- Summerfield, Q., 1979. Use of visual information for phonetic perception. *Phonetica* 36 (4–5), 314–331.
- Trejo, L.J., Johnson, T., Hyatt, A., 2000. Visual-auditory interactions and mismatch negativity: a study of the McGurk effect. In: *Second International Congress on Mismatch Negativity and its Clinical Applications*, Barcelona, Spain, June 15–18.

- 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.
- van Wassenhove, V., Grant, K.W., Poeppel, D., 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45 (3), 598–607.
- Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A., Schröger, E., 2004. From symbols to sounds: visual symbolic information activates sound representations. *Psychophysiology* 41 (5), 709–715.
- Winkler, I., 2003. Change detection in complex auditory environment: beyond the oddball paradigm. In: Polich, J. (Ed.), *Detection of Change: Event-Related Potential and fMRI Findings*. Kluwer Academic Publishers, Boston, pp. 61–81.
- Winkler, I., 2007. Interpreting the mismatch negativity (MMN). *Journal of Psychophysiology* 21 (3–4), 147–163.
- Winkler, I., Karmos, G., Näätänen, R., 1996. Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. *Brain Research* 742 (1–2), 239–252.