Neuroscience Letters 465 (2009) 204–209

Contents lists available at ScienceDirect

Neuroscience Letters

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

Perception based method for the investigation of audiovisual integration of speech

Zsófia Huhn[a,b,*], Gábor Szirtes[c,d], András Lőrincz[d], Valéria Csépe[a,c]

[a] Department of Psychophysiology, Research Institute for Psychology, Hungarian Academy of Sciences, Hungary
[b] Biophysics Department, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences, Hungary
[c] Department of Cognitive Psychology, Institute of Psychology, Eötvös Loránd University, Hungary
[d] Department of Information Systems, Eötvös Loránd University of Sciences, Hungary

ARTICLE INFO

Article history:
Received 22 April 2009
Received in revised form 26 August 2009
Accepted 29 August 2009

Keywords:
Audiovisual integration
Speech
Multisensory interaction
McGurk effect

ABSTRACT

Speech comprehension is significantly improved by visual input on the speaker's mouth movements. Audiovisual integration underlying this phenomenon is often studied in EEG experiments in which the event related brain potential (ERP) elicited by a bimodal stimulus is compared to the sum of ERPs triggered by auditory and visual signals of the same source. However, this method leads to spurious results in time ranges when ERP components common to all these stimulus types are present. A method that aims to filter out such common early anticipatory potentials is data high-pass filtering. In the present study, first, we demonstrated that subtle changes in filter cut-off frequency lead to remarkably different results on the interaction effect so that no reliable conclusion on the spatial distribution of the interaction could be drawn. Second, we suggested a different approach for the investigation of ERP correlates of audiovisual integration: bimodal syllables modified by light temporal asynchrony were presented to subjects and ERPs correlating with the fused and unfused perceptions were compared. We found that components corresponding to both auditory N1 and P2 waves were smaller in case of the fused perception, supporting the view that N1 and P2 generator activities are suppressed during multimodal speech perception. The N1 effect showed a clearly right hemisphere dominance while the effect around the P2 peak was most pronounced on centroparietal electrodes and dominated over the left hemisphere.

In face-to-face communication, acoustic information provided by speech sounds and visual information provided by lip movements of the speaker together form a subjective perception that supports language comprehension. The processing of spoken utterances is both accelerated [12,4] and improved [10,1] when the face of the speaker can be seen while listening to his or her voice. This audiovisual integration leads to the McGurk effect [13] as well, when some syllable-pairs undergo a perceptual fusion so that the perceived phoneme differs from *both* the auditory and the visual stimuli (e.g. visual /ga/ combined with auditory /ba/ is often heard as /da/).

The timing of integrative processes leading to these behavioural effects can be effectively studied by electrophysiological techniques. An often used method for the detection of crossmodal interaction in electro- and magnetoencephalographic (EEG and MEG) experiments is the '*sum of unimodal versus bimodal*' comparison. In this case event related potentials are measured in response to separate stimulation of either the auditory (A) or the visual (V) modalities, as well as to concurrent bimodal audiovisual (AV) stimulation. The sum of ERPs evoked by the unimodal stimuli is compared then to the ERP triggered by the bimodal input [12,16,4,18]. Difference between the A + V and AV ERP waves is interpreted as the indicator of integrative processes based on the following rationale: if the A and V modalities were to be processed independently after the AV stimulation, ERP(AV) would be equal to ERP(A + V) due to the law of linear superposition of electric fields. Hence, any difference between ERP(AV) and ERP(A + V) should reflect interactions between the inputs from the two modalities [2,5].

However, this additive model is valid only when there are no processes common to A, V and AV conditions: if there is a component X present in all three cases, the calculated interaction effect (ERP(A) + ERP(V) – ERP(AV)) would be equal to the real interaction plus X, since X was added twice for the unimodal cases but subtracted only once with the bimodal ERP [(X + X) – X = X]. Thus, interaction effect would be demonstrated even in the absence of any real interaction, or no interaction effect may be seen on electrodes where the signals measured from the interaction and the X processes has similar magnitude but opposite sign. Common activ-

* Corresponding author at: Biophysics Department, KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences, Konkoly Thege út 29-33, H-1121 Budapest, Hungary. Tel.: +36 392222222/1238.
E-mail address: zsofi@rmki.kfki.hu (Z. Huhn).

0304-3940/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.neulet.2009.08.077

ities causing such problems in the interpretation may emerge at virtually any latency. 200 ms after the stimulus onset, semantic processes, target processing (e.g. N2b/P3 waves in ERP/MEG recordings), response selection or motor processes emerge [11], therefore the analysis is usually restricted to the first 200–300 ms latency window. In turn, in this early time range, anticipatory ramp-like deflections (e.g. contingent negative variation, CNV) starting before the stimulus onset might give rise to spurious interaction effects [17]. In order to overcome this problem, data are often high-pass filtered with a relatively high cut-off frequency which filters out slow anticipatory potentials. Accordingly, Teder-Salejarvi et al. [17] showed that early spurious interaction effects around 40 ms could be removed by using a high-pass filter of 2 Hz cut-off frequency. However, we hypothesize that this filtering might also modify the components related to auditory, visual or interaction processes, and leads to unknown changes in the observed interaction effect. Therefore in the first part of the paper we investigate the sensitivity of the sum of unimodal vs. bimodal comparison method against the filtering conditions: we examine the effect of using high-pass filters with different cut-off frequencies (0.1, 1, 1.5 Hz) on the timing and scalp distribution of the interaction effects. In the second part of this study, we use a different paradigm for the investigation of audiovisual integration, the '*fused versus unfused perception*' method. This method has already been applied in an fMRI study by Miller and D'Esposito [14] but was not used in EEG experiments yet.

Twenty-three native Hungarian speaker volunteers (21 females, 2 males, mean age: 21.2 years, range: 19–35 years) participated in the study, for which they gave a written informed consent in accordance with the Declaration of Helsinki. All but one was right-handed and all were free of any neurological disease or audiological problems and had normal or corrected-to-normal vision.

The audiovisual syllable [ba:] was produced by a female Hungarian native speaker (Fig. 1A) and the scene was recorded by using a DV camera at a video sampling rate of 25 fps and a remote microphone at audio sampling rate of 44.1 kHz. For this natural speech stimulus, the first detectable lip movement occurred at 360 ms preceding the sound onset (Fig. 1A, Bc). Audiovisual stimuli of 10 different temporal offsets were produced so that the asynchronies between the onsets of the auditory and visual stimuli (hereafter stimulus onset asynchrony, SOA) were 360, 320, . . ., 0 ms, the natural timing of the audiovisual stimulus corresponding to a SOA of 0 ms. A behavioural pilot was conducted in order to determine the temporal offset threshold for perceptual fusion. Subjects were presented with 10 repetitions of the audiovisual stimuli for each SOA, delivered in a random order. The speech sounds were presented via headphones and the corresponding video was displayed on a monitor located at a distance of 1 m from the subject. Participants were instructed to indicate whether they perceived the stimuli as fused in a single event (fused perception) or as temporally segmented (unfused perception), respectively. The SOA where the number of 'fused' and 'unfused' answers were similar, i.e. the number of fused perceptions divided by the number of all trials was nearest to 0.5 (Fig. 1C), was chosen for presentation (AV-Am stimulus, see below) in the EEG experiment.

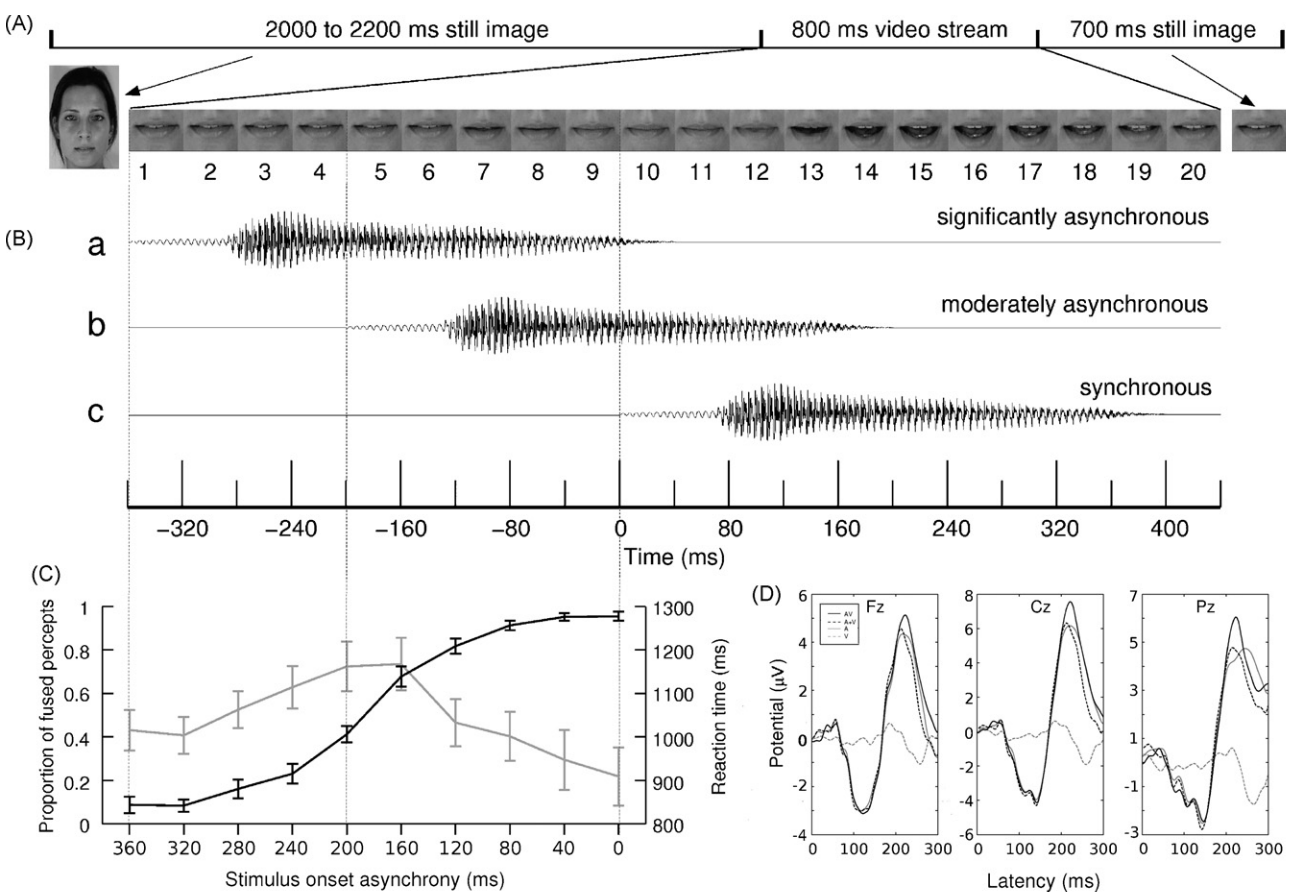

Fig. 1. Time courses of audiovisual trials. (A) Visual stimulus. (B) Auditory stimulus onsets in case of the three synchrony levels. (C) Timing of the moderately asynchronous audiovisual stimuli (AV-Am) was determined based on the behavioural pilot: AV stimuli with 10 different SOAs were presented, and the SOA for which the proportion of the number of fused perceptions and the number of all trials (black curve) was closest to 0.5 was chosen (160 or 200 ms). The reaction time (grey curve) measured from the onset of auditory stimulus is the highest for these SOAs. (D) ERPs evoked by AV-As stimuli (black continuous line) differ significantly ($p < 0.05$) from the sum of ERPs (black dashed line) triggered by the unimodal auditory (grey continuous line) and visual stimuli (grey dashed line) only after 200 ms latency, where the bimodal ERP has higher amplitudes then the sum of unimodal ERPs.

In the main experiment using EEG recording, the syllables were presented in auditory-only (A), visual-only (V) and audiovisual (AV) conditions. In the A condition, the 400-ms-long auditory syllable was presented with the visual image of the still face while in the V condition no sound was played during the video stream. In the AV conditions, each trial began with the presentation of a still image of a face for a random interval of [2000, 2100, 2200] ms (Fig. 1A). Then, the 800 ms long movie of the articulating face was played and the last frame of the video was displayed for 700 ms. AV stimuli with three different SOAs, 0, 160 or 200 ms (depending on the subject's pilot results on perceptual fusion threshold) and 360 ms, were presented that corresponded to synchronous (AV-S), moderately asynchronous (AV-Am) and significantly asynchronous (AV-As) stimuli, respectively (Fig. 1B). Based on the pilot results, the expected ratio of fused perceptions was 1, 0.5 and 0 for the AV-S, AV-Am and AV-As stimuli, respectively. The AV-As stimulus type was presented in order to counterbalance the amount of fused and unfused perceptions therefore ERPs elicited by this stimulus are not analysed in detail. However, sample ERPs are provided in Fig. 1D. Subjects were asked to press a button with right or left hand after each AV stimulus to indicate fused or unfused perception. The left and right positions of the response-buttons were counterbalanced across the subjects. A, V, AV-S and AV-As stimuli were presented 90 times each while AV-Am was repeated 180 times. Stimuli were pseudo-randomized and the 540 stimuli in total were divided into three blocks. Five short pauses were kept in each block and there were longer breaks between the blocks.

EEG recordings were made using a BrainAmp amplifier (sampling rate: 1000 Hz, online bandpass filter: 0.1–100 Hz), BrainVision Recorder program and EasyCap electrode caps with 31 standard electrode sites. An electrode was put on the outer canthus of the right eye in order to register horizontal eye movements. Electrode impedances were kept below 5 kΩ and voltages were referenced to the nose. EEG data were digitally filtered using a zero phase-shift low-pass filter with cut-off frequency at 45 Hz. For the sum of unimodal vs. bimodal analysis, data were also high-pass filtered offline with cut-off frequencies 1 and 1.5 Hz. Epochs were extracted with the following time ranges: from 100 ms before to 500 or 900 ms after stimulus in A and V conditions, respectively and from 100 ms before visual stimulus to 500 ms after auditory stimulus onset in all the AV conditions. Epochs were baseline corrected on an interval of 100 ms preceding the visual (V and AV conditions) or the auditory stimulus (A condition). Trials with signal amplitude exceeding 80 μV at any electrode at any time point in the epoch were rejected and the remaining trials were averaged. The mean number of averaged trials was 83.3, 77.8 and 82.6 for the A, V and AV-S conditions, respectively. In the AV-Am condition, the SOA of the presented AV stimulus was 200 ms for 15 subjects and 160 ms for 8 subjects. In the fused vs. unfused analysis, only those 10 subjects from the 200 ms SOA group were included, where the number of trials in each perceptual category exceeded 25. The mean number of perceptually fused trials in this group was 70.8 whereas the number of trials perceived as unfused was 74.3 on average.

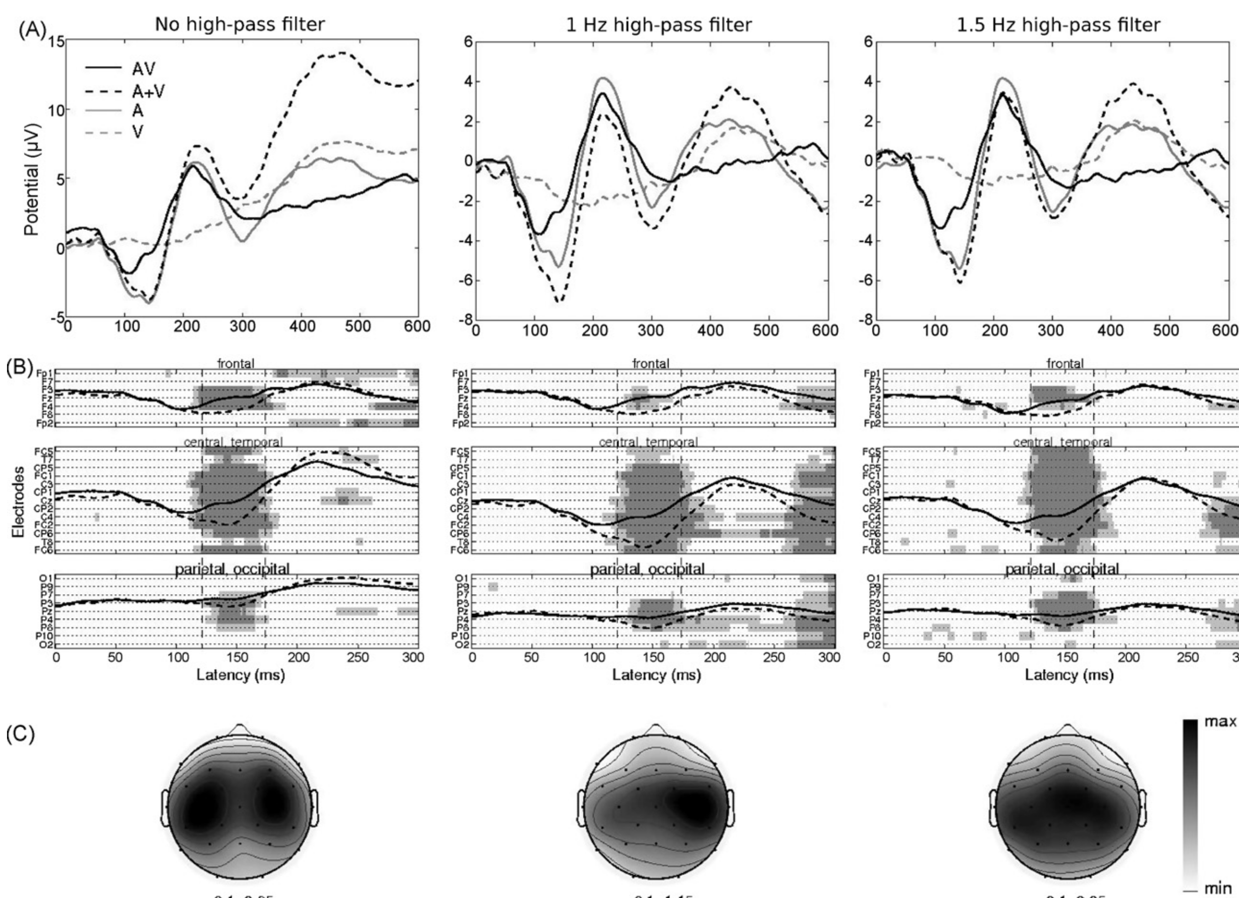

Fig. 2. Bimodal vs. sum of unimodal ERPs analysis. (A) Comparisons shown for ERPs evoked by the audiovisual stimulus (black continuous line) and the sum of unimodal ERPs (black dashed line) triggered by the auditory-only (grey continuous line) and the visual-only stimulus (grey dashed line) at electrode Cz, for the three different filtering conditions. High-pass filtering removed the slow positive shift of the ERPs. (B) Significant differences between the AV and A + V traces (light grey shading: $p < 0.05$, dark grey shading: $p < 0.01$) at different electrodes (grouped according to scalp locations). Significant differences can be seen on most electrodes near the N1 peak, in the 120–170 ms latency range (vertical dashed lines). (C) Scalp distribution of averaged effect sizes in this range depended strongly on the filtering condition. Ranges of effect size values are indicated under the scalp maps.

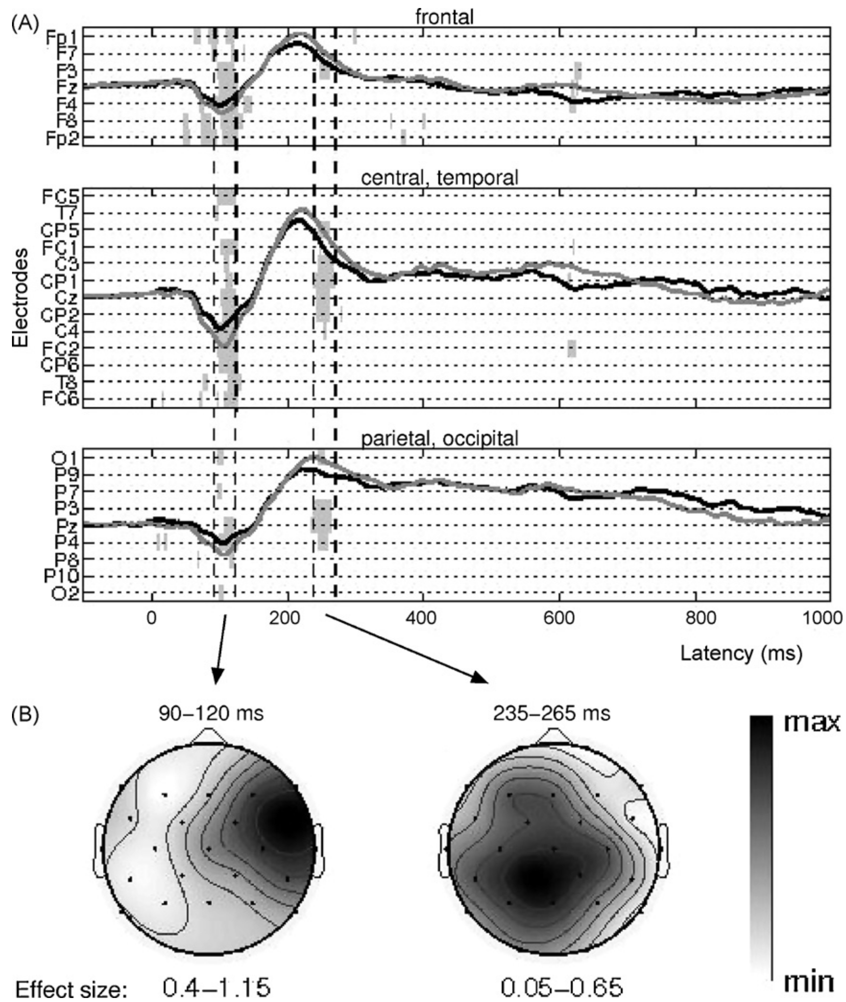

Fig. 3. Fused vs. unfused perception analysis. (A) Significant differences ($p < 0.01$, grey shading) between ERPs corresponding to fused (black line) and unfused perceptions (grey line) appear at many of the electrodes in two time ranges: near the N1 peak (90–120 ms latency, first vertical dashed line pair) and in the descending part of P2 wave (235–265 ms latency, second vertical dashed line pair). (B) Scalp distribution of average effect size in the former range is highly asymmetric and shows right hemisphere dominance, while in the latter range it is less asymmetric and is more pronounced on the left hemiscalp.

Significant differences between (1) bimodal AV-S ERP and the sum of unimodal A and V ERPs as well as (2) ERPs belonging to fused and unfused perception were assessed by using randomization tests [6], where the test statistic was the Student's t-value for matched samples. The compared traces were randomly permuted separately for each subject and histograms of test statistics were constructed based on 1000 random permutations, for each time point and electrode. The p value was calculated as the proportion of test statistics t that resulted in a larger absolute value than the absolute value of the test statistic observed for the original, non-permuted data. In time points where statistically significant effect was observed on most of the electrodes (120–170 ms latency for the bimodal vs. sum of unimodal, 90–120 and 235–265 ms latency for the fused vs. unfused perception method), Cohen's d value [8] was calculated to quantify the size of the observed effect. The effect size is defined as the difference between two means divided by a standard deviation for the data: $d = (\overline{x_1} - \overline{x_2})/s$, where s is the pooled standard deviation: $s = \sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2)}$, with $\overline{x_k}$ and $\overline{s_k}$ as the mean and standard deviation for group k, for $k = 1, 2$. The hemisphere dominance was determined by comparing the effect size values of corresponding right and left hemisphere electrodes, by Studest's t-test for matches samples. The scalp distribution was determined to be asymmetric when p value was smaller than 0.05.

Fig. 2A presents the ERPs elicited by unimodal A, V, and bimodal AV-S stimuli from the onset of the auditory signal up to 600 ms latency, at electrode Cz. The first significant negative and positive deflections of the auditory ERP are the N1 and P2 waves, peaking around 140 and 210 ms. The relatively late maximum of the N1 wave is likely caused by the long (~80 ms) voice-onset time of the auditory consonant-vowel stimulus (Fig. 1B). These two auditory event related components are also unequivocally seen on the audiovisual ERP, however the peak latency of N1 is significantly shorter, it is around 100 ms.

Crossmodal interactions were first analysed by the AV vs. A + V comparison of the synchronous AV data and the effect of filtering using this method was studied by comparing the results in three conditions: using no offline high-pass filter, high-pass filtering with a cut-off frequency of 1 Hz and 1.5 Hz (Fig. 2). The slow positive shift (putative anticipatory potential) seen on all A, V and AV curves without high-pass filtering could be removed by the application of the high-pass filter. In all three filtering conditions, significant differences between the AV and A + V traces near the N1 peak (120–170 ms) have been found at many electrodes (Fig. 2B): the N1 wave was smaller in response to the bimodal stimulation than the sum of unimodal ERPs. The scalp distribution of the difference wave, however, strongly depended on the filtering condition: the effect size, expressing the actual difference between AV and A + V

ERP values, showed a symmetric distribution ($p = 0.0503$) with temporal maxima when no offline high-pass filter was used, whereas high-pass filtering the data with 1 Hz cut-off frequency resulted in a strong right hemisphere dominance ($p = 3.6e{-}04$) and a filter of 1.5 Hz cut-off frequency resulted in a slight left hemisphere dominance ($p = 0.044$) of the effect size with central maximum (Fig. 2C). The P2 wave was even more substantially effected by the filtering parameter: no offline high-pass filtering resulted in smaller AV than A + V values, while 1 Hz filtering resulted in larger ones (Fig. 2A). These effects were significant ($p < 0.05$) at electrodes Fp1, F7, Fp2, Fc5, T7, Cz, CP2, Pz and C4, Cp6, P4, P8, O2, respectively (Fig. 2B). On the other hand, no significant differences on any of the electrodes were found with the 1.5 Hz filter near the P2 peak (Fig. 2B).

In conclusion, slight changes in the cut-off frequency value substantially influenced the results of the sum of unimodal method vs. bimodal method, therefore few reliable conclusion on the interaction effect pattern could be drawn. Audiovisual interactions were further studied by the fused vs. unfused perception method, by comparing the ERPs belonging to fused and unfused perception in response to the moderately asynchronous AV stimulation. Fig. 3 depicts significant differences ($p < 0.01$) on each electrode, from −100 to 1000 ms after auditory stimulus onset. Remarkable differences between the two ERPs were found in two distinct time ranges, around the first large negative wave (N1, 90–120 ms latency) and after the maximum of the latter large positive deflection (P2, 235–265 ms latency). The minima of p values in the given time ranges varied for the different electrodes in the ranges [2.5e−6, 0.04] and [2.8e−5, 0.64] in case of the N1 and the P2 effects, respectively, which suggests a larger N1 than P2 effect. In both ranges, the average potentials in absolute value were smaller for fused than for unfused percepts. Fig. 3B shows the mean of the effect size values in the two time ranges: near the N1 peak, the effect size was largest on the right temporal electrodes (right hemisphere dominance, $p = 3.2e{-}04$), whereas in the second time range, after P2 peak, the effect size had largest values on the central/parietal electrodes and on average was larger on the left than the right hemiscalp ($p = 2.6e{-}05$).

Although audiovisual integration related to speech perception has been widely studied by both fMRI [7,15,14] and EEG/MEG [12,4,18,16] experiments, still no clear picture about the mechanism of integrative processes has been emerged yet. One approach in EEG/MEG experiments is to find reliable changes of the ERPs elicited by bimodal stimulation as opposed to the sum of the ERP correlates of the unimodal stimulations. By this method a reduction in N1 and P2 waves in response to bimodal stimulation was observed, however the scalp distributions of these effects are controversial. Besle et al. [4] have shown that the N1 amplitude reduction is most pronounced at left hemisphere electrodes, while another study reported right hemisphere dominance [12], and in a third one the distribution of N1 effect was found to be rather symmetric and had a central maximum [16]. The P2 wave usually has its peak after 200 ms poststimulus which falls in the time range when common late processes are probably already present, making the sum of unimodal vs. bimodal method invalid. However, when the method was applied in the late time ranges too, P2 reduction was found to be largest at central sites [12,16].

In our study, using the same method, the N1 peak was found to be reduced in response to AV stimulation compared to the sum of ERPs triggered by the unimodal A and V stimuli. However, the scalp distribution of this difference largely depended on the filtering. It has shown right, central or left hemisphere dominance when data were high-pass filtered with 0.1, 1, or 1.5 Hz cut-off frequency, respectively. The observed effect of bimodal processing on the P2 wave was even more strongly affected by the filtering method. The bimodal trace could show smaller, larger, or equal amplitude than the sum of unimodal ERPs, depending on the high-pass filter used. As the three different filters gave different result on N1 and P2 effects, no conclusion about the real scalp distributions could be drawn.

In the present study we proposed a different method for investigating the electrophysiological correlates of audiovisual integration: AV interactions associated with fused perception were compared to those correlated to unfused perception. The advantage of our method is that it may reveal information about time ranges when non-specific processes are also present. Accordingly, early anticipatory potentials did not need to be removed and interactions more than 200 ms after auditory stimulus onset could also be studied.

This method revealed two distinct time ranges important for audiovisual integration: around the auditory N1 peak (90–120 ms after auditory stimulus onset) and after P2 peak, in the ascending phase of P2 component (235–265 ms poststimulus). In both intervals, the ERP amplitude was smaller in case of fused than unfused perception. This is in agreement with results that N1 and P2 amplitude in response to bimodal stimulation is smaller than the sum of unimodal ERPs [12,4,18,16], which is often interpreted so that suppressive interaction takes place during audiovisual integration, leading to depressed N1 and P2 generator activities. This assumption was recently confirmed by human intracranial recordings from the auditory cortex, where audiovisual interaction were shown to consist of both a suppression of the visual response to lipreading and a decrease of the auditory response to the speech sound [3]. The scalp distribution of the N1 suppression effect determined by our method was found to be highly asymmetric, it is most pronounced on the right temporal areas, whereas P2 decrease is largest on central electrodes with a weak dominance in left hemisphere. In the fMRI study with similar paradigm [14] the right STS and the right medial superior frontal gyrus showed decreased activity with perceptual fusion, which might correspond to the N1 decrease in our experiment. The left hemisphere dominance of the latter effect is supported by the results of Doesburg et al. [9] who showed a left hemisphere bias in gamma-band synchrony increase from 170 to 250 ms after bimodal stimulus onset. However, time–frequency analyses of the present results and further, combined EEG–fMRI, experiments will be needed to determine the frequency ranges and brain areas that play important roles in the formation of the observed N1 and P2 effects.

Acknowledgement

This study was supported by the Internal Grant provided by the Research Institute for Psychology.

References

[1] P. Arnold, F. Hill, Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact, British Journal of Psychology 92 (2001) 339–355.
[2] D.S. Barth, N. Goldberg, B. Brett, S. Di, The spatiotemporal organization of auditory, visual and auditory–visual evoked-potentials in rat cortex, Brain Research 678 (1995) 177–190.
[3] J. Besle, C. Fischer, A. Bidet-Caulet, F. Lecaignard, O. Bertrand, M.H. Giard, Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans, Journal of Neuroscience 28 (2008) 14301–14310.
[4] J. Besle, A. Fort, C. Delpuech, M.H. Giard, Bimodal speech: early suppressive visual effects in human auditory cortex, European Journal of Neuroscience 20 (2004) 2225–2234.
[5] J. Besle, A. Fort, M.H. Giard, Interest and validity of the additive model in electrophysiological studies of multisensory interactions, Cognitive Processing 5 (2004) 189–192.
[6] R.C. Blair, W. Karniski, An alternative method for significance testing of waveform difference potentials, Psychophysiology 30 (1993) 518–524.
[7] G.A. Calvert, R. Campbell, M.J. Brammer, Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex, Current Biology 10 (2000) 649–657.

[8] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, 1988.
[9] S.M. Doesburg, L.L. Emberson, A. Rahi, D. Cameron, L.M. Ward, Asynchrony from synchrony: long-range gamma-band neural synchrony accompanies perception of audiovisual speech asynchrony, Experimental Brain Research 185 (2008) 11–20.
[10] K.W. Grant, P.F. Seitz, The use of visible speech cues for improving auditory detection of spoken sentences, Journal of the Acoustical Society of America 108 (2000) 1197–1208.
[11] S.A. Hillyard, W.A. Teder-Salejarvi, T.F. Munte, Temporal dynamics of early perceptual processing, Current Opinion in Neurobiology 8 (1998) 202–210.
[12] V. Klucharev, R. Mottonen, M. Sams, Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception, Cognitive Brain Research 18 (2003) 65–75.
[13] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (5588) (1976) 746–748.
[14] L.M. Miller, M. D'Esposito, Perceptual fusion and stimulus coincidence in the cross-modal integration of speech, Journal of Neuroscience 25 (25) (2005) 5884–5893.
[15] I.R. Olson, J.C. Gatenby, J.C. Gore, A comparison of bound and unbound audio-visual information processing in the human cerebral cortex, Cognitive Brain Research 14 (1) (2002) 129–138.
[16] J.J. Stekelenburg, J. Vroomen, Neural correlates of multisensory integration of ecologically valid audiovisual events, Journal of Cognitive Neuroscience 19 (2007) 1964–1973.
[17] W.A. Teder-Salejarvi, J.J. McDonald, F. Di Russo, S.A. Hillyard, An analysis of audio–visual crossmodal integration by means of event-related potential (ERP) recordings, Cognitive Brain Research 14 (2002) 106–114.
[18] V. van Wassenhove, K.W. Grant, D. Poeppel, Visual speech speeds up the neural processing of auditory speech, Proceedings of the National Academy of Sciences of United States of America 102 (4) (2005) 1181–1186.