The spatiotemporal characteristics of elementary audiovisual speech and music processing in musically untrained subjects

Stefan Elmer a,⁎, Martin Meyer b,⁎, Lutz Jäncke a,⁎

a Division Neuropsychology, Institute of Psychology, University of Zurich, Zurich, Switzerland
b Center for Integrative Human Physiology, Zurich, Switzerland

⁎ Research Unit "Plasticity and learning in the aging brain", Institute of Psychology, University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 7 October 2010
Received in revised form 16 June 2011
Accepted 11 September 2011
Available online 25 September 2011

Keywords:
Audiovisual interactions
Multimodal processing, Auditory-related cortex

ABSTRACT

Previously, the EEG technique has been used to investigate the spatiotemporal properties of audiovisual (AV) processing by taking advantage of the violation of the "additive model", which is considered to be a very conservative approach. In the present work, we used a less conservative and novel approach than the criterion of superadditivity for estimating AV interactions. Hence, we estimated AV interaction patterns by comparing the responses to AV stimuli with the averaged responses to the unimodal visual and auditory stimuli in musically untrained subjects and by presenting syllables and piano tones coupled with light flashes. Our results suggest that the two AV objects elicited consistent interaction patterns within the time course of unimodal processing in the time range between 80 and 250 ms post stimulus onset. The scalp topographies, as well as the source estimation approach we adopted, indicate that the first interaction pattern at around 100 ms was partially driven by auditory-related cortical regions. Additionally, we found evidence for a second interaction pattern at around 200 ms that was mainly associated with the responsiveness of extra-sensory brain regions. During this later processing stage, only the music condition was associated with putative responses that originated from auditory-related cortical fields. This study provides a novel approach to investigate the basic principles underlying elementary AV speech and music processing in subjects without formal musical education.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

An example of sensory interaction is provided by considering AV speech perception. In this context, it was previously shown that AV speech perception occurs automatically by integrating visual and auditory information, even though this sensory interaction can result in an illusory percept (McGurk and MacDonald, 1976). This example shows that unisensory functions are probably not conducted in isolation, but can be influenced by ongoing responses driven by other sensory modalities (Besle et al., 2009; Koelewijn et al., 2010; Murray and Spierer, 2009). The synthesis of inputs from different sensory modalities not only facilitates the detection, localisation, and recognition of a given event, but also supports cognition and communication (Teder-Salejarvi et al., 2002). Within the framework of the so-called “additive model”, sensory interactions are generally estimated by comparing the responses to AV stimuli with the algebraic sum of responses to the unisensory signals presented in isolation (Besle et al., 2009; Molholm et al., 2002, 2006; Sugihara et al., 2006; Teder-Salejarvi et al., 2002). Following the principles postulated by this model, in the hypothetical case of no interaction, the event-related-potentials (ERPs) of the AV condition should equal the summed unisensory responses. Otherwise, any violation of this equality (i.e., super-additivity or sub-additivity) points to an interaction between the unisensory responses. In this context, it should be mentioned that the criterion of superadditivity was previously postulated to be a “too conservative” approach (for examples consider Stanford and Stein, 2007). Consequently, some authors believe that superadditivity is not an adequate measure for characterizing multisensory interactions. For example, several single-cell recording studies adopted a “less conservative” method to estimate neurons with multisensory properties. In this context, it was advocated that a neuron has multisensory properties when its response to a bimodal stimulus is greater than the response elicited by a unimodal stimulus (for example consider Meredith and Stein, 1983, 1986). Since this “less conservative” approach is restricted to the investigation of AV interactions at the level of a single neuron, in the present work we propose an alternative approach for investigating multisensory interactions basing on voltage values collected over the scalp.
Even though some EEG studies provide evidence for the notion that bimodal AV signals interact in sensory-specific areas (Moholm et al., 2002), currently there is a plethora of data showing that AV interactions occur within the time course of sensory processing in brain regions that are not specifically devoted to assimilate sensory signals. In fact, there is mounting evidence showing that extra-sensory brain regions residing in occipital, parietal, frontal, and temporal cortices contribute to the synthesis of auditory and visual signals (De Santis et al., 2007; Giard, 2000; Inui et al., 2006; Massaro, 1999; Moholm et al., 2006; Spierer et al., 2007). Otherwise, recent neuroimaging and electrophysiological studies conducted with humans (Besle et al., 2009; Giard and Peronnet, 1999; Murray and Spierer, 2009; Teder-Salejarvi et al., 2002) and animals (Ghazanfar et al., 2005; Kayser et al., 2007), militate in favor of the view that the synthesis of auditory and visual information can occur in brain regions that are usually thought of as being responsive to sensory-specific stimuli. Generally, AV interactions that arise in sensory-specific brain regions can be explained by either 1) direct projections between primary or non-primary brain regions; or 2) by feedback projections from extra-sensory brain regions to sensory-specific cortical fields; or 3) by both mechanisms (1 and 2) operating together (Murray and Spierer, 2009).

In the present EEG study we adopted a "less conservative" criterion than the one postulated by the “additive-model” for investigating AV speech (syllables) and music (piano tones) processing in individuals without formal musical training. In particular, we addressed the question whether these two processing modes share similar spatiotemporal characteristics. This question is justified by the fact that even though speech and music are perceptually distinct, they share many acoustic commonalities. In fact, both signals convey acoustic information by means of timing, pitch, and timbre cues (Kraus and Chandrasekaran, 2010). Furthermore, in the domain of speech processing, one of the most cardinal cues pertinent to conveying phonetic information is the so-called voice-onset–time (VOT), which is a fast temporal cue that can be defined as the time between the release of a stop consonant and the onset of vocal folds vibrations (Elmer et al., 2011). For example, based on the discrimination of the VOT, a German listener is able to distinguish the consonant–vowel (CV) syllable da/from/ba/. These fast temporal cues are not only specific to speech processing, but they also occur during the perception of piano tones in the form of short, sharp, or impulsive tones (Schneider et al., 2005). Moreover, since the processing of word- and sight-reading is known to diverge (Schon et al., 2002; Sergent et al., 1992), in this study we used flashlights that were presented simultaneously with syllables and piano tones. This methodological approach is particularly fruitful because it enables us to compare AV speech and music processing in individuals, who have not attended formal musical training. In particular, we addressed the question whether speech and music stimuli interact with flashlights at about 100 and 200 ms post stimulus onset. Based on the strong evidence at hand, we believe that the temporal dynamics of elementary AV processing are relatively robust across a variety of stimuli characterized by different physical properties. Therefore, in the present study we explicitly addressed the question whether speech and music stimuli interact with flashlights at about 100 and 200 ms post stimulus onset by accessorily modulating the sensory-related neuronal sources which generate the unisensory responses.

2. Materials and methods

2.1. Subjects

Nineteen native Swiss-German volunteers (male, mean age 26.4 ± 2.6 SE) participated in the present study. None of the participants had formal musical training, and according to the Annett-Handedness-Questionnaire (AQH) (Annett, 1970) all subjects were consistently right-handed. Subjects gave written consent in accordance with procedures approved by the local ethics committee and were paid for participation. None of the participants had any history of neurological, psychiatric or audiological disorders.

2.2. Procedure

During the experiment, different presentation modes, namely, auditory (A), visual (V), and audiovisual (AV) signals were delivered in three separate runs (each of 6 min duration). The first run was comprised of 90 white and 90 red flashlights (V), the second run was made up of 90 syllables and 90 piano triads (A), and the third run consisted of 90 white flashes coupled with syllables and 90 red flashes coupled with piano triads (AV). In each run the different stimuli were presented consecutively in the context of a block design (first block = V; second block = A; third block = AV). This paradigm was previously used for investigating the spatiotemporal dynamics of auditory imagery.

2.3. Stimuli

2.3.1. Auditory stimuli

The auditory stimuli were registered as 16 bit stereo files and matched for intensity by using the MagIX Audio Studio software (MAGIX Audio Studio 03 deLuxe, Magix AG, Berlin, Germany). All auditory stimuli were delivered binaurally with a sound pressure level of about 50 Db (Digital Sound Level Meter 329, Voltcraft) using HiFi headphones (Sennheiser, HD 25–1, 70 Ω, Ireland). The verbal items consisted of three spoken consonant–vowel syllables (/Ka/ /Ta/ /Pa/), /VOT/Ka/= 59 ms; /VOT/Ta/= 49 ms; /VOT/Pa/= 39 ms) which were recorded by a professional male speaker. These stimuli have been used in previous experiments (Jancke et al., 2002; Jancke and Shah, 2004; Meyer et al., 2007; Elmer et al., 2011). The musical items were computer-generated artificial piano tones (MAGIX Audio Studio 03 deLuxe, Magix AG, Berlin, Germany) and consisted of one major (C4/E4/G4) and two minor triads (D4/F4/A4 and E4/G4/A4), American
notion). All auditory stimuli lasted 500 ms and were smoothed with a fall time of 10 ms, in order to avoid an abrupt decay. The Spectra and waveforms of the auditory stimuli used in this experiment are shown in a previous work of our group (Meyer et al., 2007).

2.3.2. Visual stimuli

The visual stimuli consisted of red and white flashlights presented on a computer screen for 50 ms; the stimuli covered the whole surface of the computer screen (15 in.). The red, green, and blue components of the color were controlled by the Presentation software (Neurobehavioral Systems, USA, version 0.70, www.neurobs.com) and were numerically coded from 0 to 255. The data for each pixel was then used to generate the red, green, and blue values in this range. The flashlights consisted of the following values: White (255/255/255) and red (255/0/0). The white light of ~168.8 cd/m². The red flashlight had a density of light of ~36.4 cd/m² (Luminance Meter LS-110, Minolta).

2.4. Presentation mode and task

The auditory stimuli used in this experiment lasted 500 ms and were followed by an inter-stimulus-interval (ISI) of 1500 ms. If only a visual stimulus (50 ms) was presented, then the ISI lasted 1950 ms. During the AV condition, the visual and auditory signals were presented exactly at the same time. In order to ensure that the subjects attended the stimuli, an answer was required after each sentence was presented exactly at the same time. In order to ensure that the subjects attended the stimuli, an answer was required after each sensory stimulation (V/A/AV), that is, subjects were requested to respond by pressing a button with their right forefinger. This sort of task was used previously and it has been shown to ensure attention during the whole period of EEG recording (Molholm et al., 2006).

Since Fort and colleagues (Fort et al., 2002) could show that their AV stimuli, which were similar to the stimuli that we use in our study, are not mandatorily expected to facilitate perception, the subjects in this study were not explicitly instructed to respond as fast as possible. Thus, reaction time was not evaluated. Presentation software controlled the stimuli presentation (www.neurobs.com).

2.5. EEG recording and analysis

The subjects were placed in a chair at a distance of 110 cm from the monitor and supported their chin on a chin-rest, in order to avoid artifacts during the EEG measurement. The EEG (30 channels + 2 chromatic eye channels, subset of the 10/10 system) was recorded with a sampling rate of 500 Hz and a band pass filter from 0.1 to 100 Hz by using an EEG-amplifier (Brainproducts, Munich, Germany). We applied sintered silver/silver-chloride-electrodes (Ag/AgCl) and used the FCz position as online reference. By using Electrodeg conductant, the electrode impedance was reduced to ≤5kΩ. For all steps of digital EEG raw data processing, we used Brain Vision Analyser software (Version 1.04, Brainproducts, Munich, Germany). The data were high- and low-pass filtered at 1–45 Hz. The movement-related artifacts were removed manually and the data were re-referenced offline to an average reference derived from all channels. For computing ERPs, segments of 1100 ms, including a 100 ms pre-stimulus period, were created and a baseline correction relative to the −100 to 0 ms pre-stimulus time period was applied. The single segments were averaged for each subject according to stimulus type (A/V/AV) and condition (speech/music). Furthermore, we computed an average with the single segments related to the A and V stimuli (aAV) as well as computed multi-subject grand averages and difference waves (DWs = AV − aAV).

This novel approach we propose for investigating multisensory interaction derives from two previous observations: 1) the criterion of superadditivity was previously considered to be a “too conservative” approach, which may rely on a suboptimal measure for characterizing multisensory interactions (for example consider Meredith and Stein, 2007; 2) in the domain of single cells recording it was postulated that a neuron has multisensory properties when its response to a bimodal stimulus is greater than the response elicited by an unimodal stimulus (for example consider Meredith and Stein, 1983, 1986). Therefore, since 1) the criterion of superadditivity was proposed to be a “too conservative” approach and 2) it results difficult to convey the methodology used in the domain of single cells recording to whole-head EEG analyses, we propose a completely novel approach that takes into account these constraints. This novel approach consists in comparing the signals elicited by the bimodal objects (AV) with the averaged unisensory signals (aAV). This approach is justified by the similarity of the topographic scalp maps associated with the AV and aAV conditions (please consider Fig. 5). In fact, by considering the voltage distribution over the scalp, it becomes evident that the averaged A and V signals (aAV) have almost the same topographic properties as those associated with the AV condition. Exactly this observation militates in favor of the view that the approach we adopted in the present study may be promising for future studies.

The time segments used for statistical analyses were defined according to two consecutive global power (GFP) minima (derived from all EEG channels) of the grand average difference waves (DWs = AV − aAV) in the time range between 80 and 250 ms. In particular, the intervals between two adjacent GFP minima that overlapped with the time course of the visual P1/N2 and the auditory N1/P2 components were used as objective time windows for the ERP analyses (See Figs. 1 and 3). We selected the 80–250 ms time interval for analysis for two particular reasons. 1) The selection of this time interval relies on the results of previous studies conducted with similar stimulus material (Giard and Peronnet, 1999; Murray and Spierer, 2009; Teder-Salejarvi et al., 2002). 2) In the present study we were particularly interested in examining AV interactions that overlapped with the time course of sensory processing. Accordingly, we chose a time interval, which is known to typically overlap with the auditory N1/P2 and the visual P1/N2 complex.

The GFP defined time windows (TW) were selected separately for the verbal and musical conditions and used for all the statistical analyses reported in this work. The GFP-based analyses yielded the following TWs: TW 1 speech = 86–164 ms/TW 2 speech = 164–252 ms; TW 1 music = 86–154 ms/TW 2 music = 154–238 ms. Since we were principally motivated to explore AV interaction effects in visual- and auditory-related areas, we have decided to evaluate voltage values at electrodes Oz and Cz. By considering the topographic maps implemented in Fig. 5, it becomes visible that electrodes Oz and Cz reflect maximal voltage values during visual and auditory stimulation. A similar pattern was visible by considering the difference waves (DW) which may reflect AV interaction patterns. Furthermore, when one considers the topographic maps implemented in Fig. 5, it becomes evident that the topographies show a somewhat symmetrical distribution of voltage values over the scalp and that the electrode positions Pz and Fz are often associated with a reversal of polarity. Thus, we chose to evaluate voltage values at electrodes Oz, Pz, Cz and Fz. The ERP analyses are based on the evaluation of mean amplitudes in the GFP-defined TWs.

2.6. LORETA source estimation

In order to estimate the putative neuronal sources underlying AV interactions, we applied an established inverse linear solution approach (LORETA, low-resolution electromagnetic tomography; www.unizh.ch/keyinst/NewLORETA/LORETA01.htm) (Pascual-Marqui et al., 1994, 1999). This approach is particularly helpful because, unlike conventional dipole fitting, it does not require a-priori assumptions about the number and the localization of the dipoles. In fact, LORETA calculates the three dimensional distribution of electrically active neuronal generators in the brain as a current density value (A/m²) that is based on the recorded electric scalp potentials. LORETA provides a solution for the
This figure shows the grand average waveforms associated with the speech condition at electrodes Fz, Cz, Pz and Oz. The y-axis depicts voltage values in μV, the x-axis time in milliseconds. The time segments used for statistical analyses were defined according to two consecutive global field power (GFP) minima (derived from all EEG channels) of the grand average difference waves (DWs = AV − aAV) in the time range between 80 and 250 ms. "AV": (red), “aAV”: (blue), “A”: (black), “V”: (magenta), "DWs": (green).

This figure shows the mean amplitudes of the “AV” and “aAV” stimulation modes associated with the speech condition during the two time windows (TW 1 86–164 ms; TW 2 164–252 ms). The mean amplitudes are shown separately for the following electrodes: Oz, Pz, Cz, and Fz. The left two bars of each plot refer to TW 1 and the two right bars refer to TW 2. The dark bars depict the “AV” stimulation mode; the light bars depict “aAV”. Standard error bars are visible. ** p < .01; *** p < .001, Bonferroni corrected.
inverse problem by assuming that the smoothest of all possible activity distributions is the most plausible one for explaining the data. The characteristic feature of this particular inverse solution approach is the low spatial resolution, which conserves the location of maximal activity, but with a certain degree of dispersion (Mulert et al., 2004). Here we determined the current density distribution for epochs of brain electrical activity on a dense grid of 2394 voxels at 7 mm spatial resolution. The localisation error of LORETA’s source identification may vary between 7 (Pascual-Marqui et al., 1994) and 14 mm (Phillips et al., 2002). LORETA refers to a three-shell spherical model registered to the Talairach human brain atlas. The source estimations are provided as x, y, z coordinates and are situated relative to the intercommissural line (AC–PC line) in horizontal (x), the anterior/posterior (y), and vertical (z) directions. The solution space is confined to the gray matter portion of the human cortex, which rules out the option that subcortical tissue and white matter contribute to the solution.

When calculating the neuronal generators that contribute to AV interaction patterns (DWs), we estimated mean LORETA current density values within the GFP-defined TWs. We used a transformation matrix with high regularization (1e-3*([first eigenvalue]), in order to increase signal to noise ratio. For all sources estimated by LORETA and implemented in Table 1, we adopted a threshold of 0.0004 prop.A/m². In the present work, we consistently reported the first four maxima above threshold for each stimulation mode and condition. Even though we are fully aware that the LORETA approach we used is not a statistical procedure, we want to emphasize that this method has often been shown to reasonably estimate the current density maxima of AEP components (Meyer et al., 2006; Pascual-Marqui et al., 1994, 1999).

3. Results

3.1. ERP waves

The auditory presentation of both syllables and piano triads elicited a classical N1/P2 complex at electrode Cz. As apparent from Figs. 1 and 3 (electrode Cz), the elicited ERP responses to speech and music stimuli show waveforms that are similar in latency (syllables: N1, 120 ms; P2, 208 ms; piano triads: N1, 122 ms; P2, 194 ms). Otherwise, the piano triads elicited stronger responses than the syllables at electrode Cz (syllables: N1: −1.66 μV; P2: 2.21 μV; piano triads: N1: −3.12 μV; P2: 3.83 μV). The more complex spectral profile of the piano triads probably accounts for the enhanced amplitudes. Both the red and the white flashlights elicited a classical visual P1/N1 (white: P1: 7.33 μV, 118 ms; N1: −0.48 μV, 158 ms; red: P1: 4.43 μV, 102 ms; N1: 3.2 μV, 132 ms) and P2/N2 (white: P2: 2.06 μV, 194 ms; N2: 0.74 μV, 252 ms; red: P2: 4.46 μV, 148 ms; N2: 2.49 μV, 168 ms) complex between 100 and 250 ms after stimulus presentation at electrode Oz. The differences in amplitude of the visual ERPs occurred since the red flash stimulated additive areas involved in color and feedback processes. Alternatively, differences in luminance may account for this observation.

In the present work AV interactions were estimated by using a "less conservative" procedure than the one postulated by the more traditional "additive model". In particular, AV interactions were estimated by comparing the responses to AV stimuli with the averaged responses to the unimodal visual and auditory stimuli. These difference waves (DW = AV – aAV) showed P1/N2 like responses at electrode Oz (speech condition: P1: 2.15 μV, 128 ms; N2: −0.76 μV,
Audiovisual interaction patterns

Audiovisual (AV) interaction patterns were evaluated by comparing the mean amplitudes elicited by the AV signals with those of the averaged A and V responses (aAV) in the GFP-defined time windows (TWs). For each of the four midline electrodes (i.e., Oz/Pz/Cz/Fz), separate 2-way ANOVAs with the following independent variables were computed: two TWs and two stimulus types (STs, i.e., “AV” and “aAV”). Since we only evaluated voltage values at four a priori-defined electrodes, all statistical analyses (ANOVA) were not corrected for multiple comparisons (the Bonferroni-corrected threshold for four ANOVAs is alpha = .0125).

3.2.1. Speech condition

The four ANOVA analyses, which are related to the speech condition, yielded significant TW effects, as well as significant TW × ST interactions at three out of four electrodes, namely, at Oz (TW F(1,18) = 22.57, p < .0001; TW × ST F(1,18) = 11.32, p < .003), Cz (TW F(1,18) = 196.18, p < .0001; TW × ST F(1,18) = 25.90, p < .0001), and Fz (TW F(1,18) = 6.06, p = .024; TW × ST F(1,18) = 13.20, p = .002).

In order to further investigate the significant TW × ST interactions revealed at electrodes Oz, Cz, and Fz, post-hoc comparisons between the two STs (i.e., “AV” vs. “aAV”) were carried out for each TW and electrode by using t-tests for paired samples (Bonferroni corrected). These post-hoc analyses revealed interaction patterns at occipital (Oz 86–164 ms t(18) = 4.060, p < .001) and central (Cz 86–164 ms t(18) = −3.456, p < .003) electrodes during TW 1 (86–164 ms). We also noted AV interaction effects at frontal scalp sides (Fz 164–252 ms t(18) = 3.026, p = .007) during TW 2 (164–252 ms). The ERP curves and the significant interaction patterns are visible in Figs. 1 and 2.

3.2.2. Music condition

The four ANOVA analyses, which were related to the music condition, yielded significant TW effects, as well as significant TW × ST interactions at three out of four electrodes, namely, at Oz (TW F(1,18) = 5.24, p = .034; TW × ST F(1,18) = 7.09, p = .016), Cz (TW F(1,18) = 97.10, p < .0001; TW × ST F(1,18) = 43.04, p < .0001), and Fz...
Furthermore, the ANOVA analysis performed at electrode Oz yielded a significant effect of ST (ST F(1,18)=11.60, p=.003).

In order to further elucidate the significant interactions and main effects, post-hoc comparisons between the two ST (i.e., “AV” vs. “aAV”) were carried out for each electrode and TW by using t-tests for paired samples (Bonferroni corrected). These post-hoc analyses revealed significant differences within TW 1 at electrode Oz, Cz and Fz (Oz 86–154 ms t(18)=3.698, p = .002; Cz 86–154 ms t(18)= −3.930, p = .001; Fz 86–154 ms t(18)= −3.293, p = .004) and within TW 2 at electrode Cz and Fz (Cz 154–238 ms t(18)= 4.333, p = .0001; Fz 154–238 ms t(18)= 3.962, p = .001). The ERP curves and the significant interaction patterns are visible in Figs. 3 and 4.

3.3. Audiovisual speech vs. music processing

In order to compare AV interactions between the speech and music conditions, we statistically evaluated the mean amplitudes of the interaction patterns (DWS). With this purpose in mind, we computed two 2-way ANOVAs (4 electrodes × 2 conditions), one for each TW (TW 1 and TW 2). The outcome of these two analyses revealed a main effect for electrode (TW 1 F(1,18)=25.51, p = .0001; TW 2 F(1,18)=13.38, p = .001). Neither the electrode×condition interaction, nor the main effect of condition reached significance.

3.4. Topographic scalp maps

Analyzing topographic features is beneficial because it generally enables to gain more information about the putative neuronal sources underlying AV interaction patterns. Fig. 5 depicts the topographic maps associated with the following stimulation modes: White flashlights (Vsyllables), red flashlights (Vpiano), syllables (Asyllables), piano triads (Apiano), audiovisual syllables (AVsyllables), audiovisual piano (AVpiano), average audio and visual syllables (aAVsyllables), average audio and visual piano (aAVpiano), DWS speech condition (DWsyllables), and DWS music condition (DWpiano). We consider it important to note the similarity in topography between the unimodal auditory activity within the time course of the N1 wave (TW 1), as generated by syllables and piano triads presented alone (Asyllables and Apiano), and the estimated interaction patterns (DWsyllables and DWpiano). The similar polarities in both cases (central negativity and positive potentials around the mastoid region) may indicate that AV interactions modulated auditory-related cortical regions. Furthermore, Fig. 5 denotes some
similarities between the unimodal auditory activity elicited by the presentation of the piano triads (Apiano) and the DWs associated with the music condition during TW 2. This consistent pattern, which is characterized by a central positivity and negative potentials around the mastoid region, leads us to suggest that a modulation of auditory-related cortical regions occurred only during the music condition.

3.5. LORETA source estimation

In order to gain more information on the putative neuronal sources underlying AV interaction patterns, we estimated the current density maxima related to the different stimulation modes (V/A/DW/AV/aAV) and conditions (speech/music). Table 1 depicts the mean current density maxima estimated by LORETA within TW 1 and TW 2.

For the first time window (TW 1), current density maps of red and white flashlights indicate that the visual-related brain regions (BA 17 and 18), as well as the middle temporal (BA 39) and parietal (BA 40) areas, contribute to the perception of elementary visual stimuli. As regards TW 2, both visual stimuli elicited activity in the following areas: the visual-related regions (BA 17 and 37), the middle temporal gyrus (BA 21 and 39), and the posterior cingulum (BA 31). During TW 1 and TW 2, both the syllables and the piano triads were associated with activity in the auditory cortex (BA 22 and 42), the inferior parietal lobe (BA 40), the posterior cingulate (BA 31), and in the medial frontal gyrus (BA 6).

During TW 1, the current density maps of the DW related to the speech and music conditions revealed maxima residing in the auditory cortex (BA 22 and 42). Whereas during the same time window both conditions were associated with activity in extra-sensory brain regions (anterior and posterior cingulate, middle temporal and frontal gyrius), only the syllables condition elicited additional responses in visual-related areas (BA 18). During TW 2, LORETA uncovered AV interaction effects in auditory-related brain regions for the music condition only (BA 22). This is in line with the information that is provided by the topographic maps, which are visible in Fig. 5. In addition, the interaction between syllables and flashlights elicited responses in extra-sensory brain regions, namely, in medial frontal regions (BA 6 and 11), the cingulate gyrus (BA 24), and the right insula (BA 13).

4. Discussion

To date, the spatiotemporal characteristics underlying elementary AV speech and music processing are still unclear. In fact, the question whether speech and music interact with simple visual stimuli like flashlights by accessory modulating the sensory-related neuronal sources which generate the unisensory responses, has yet not been addressed. With this purpose in mind, we performed an EEG study and investigated subjects, who had no musical experience, while they processed visual, auditory, and AV signals. The topographic characteristics, as well as the source estimation approach that we adopted, both indicate that a common interaction pattern, which occurred at around 100 ms, was associated (at least in part) with the modulation of auditory-related cortical regions. During the second time window, at about 200 ms, the AV responses partially diverged between the speech and music condition. Whereas the AV interaction patterns associated with the music condition evoked at least some responses in the auditory-related cortex, the responses related to the speech condition were mainly supported by extra-sensory brain regions. In turn, we will discuss these results in more detail by placing particular emphasis on the functional contribution of auditory-related cortical regions to AV speech and music processing.

4.1. AV interaction patterns at around 100 ms

A first consistent interaction pattern became manifest during both the speech and the music conditions at around 100 ms over occipital (Oz) and central (Cz) scalp sides. The observation that these consistent interaction patterns occurred within the time course of unisensory visual and auditory processing, leads us to suggest that at least one part of the neural mechanisms, which contribute to the auditory N1 or to the visual P1/N1 responses, were modulated by the perception of the bimodal objects presented. The similarity between the topographic characteristics (i.e., polarity and topographic voltage distribution) elicited by the unimodal auditory N1 responses and the AV interactions at around the same latency, is compatible with the view that neural sources originate from the supratemporal plane (Vaughan and Ritter, 1970). Along this line of argumentation, the topographic maps depicted in Fig. 5 show that the unisensory auditory responses and the interaction patterns were actually characterized by positive potentials around the mastoid region with polarity reversal at fronto-central scalp sites. These particular topographic characteristics are typically associated with neuronal responses originating from supratemporal regions and may suggest that the fusion of auditory and visual features modulated the activity of the auditory N1 generators. To date, the auditory N1 responses are known to result from the activity of at least two dipolar sources that are situated in the auditory cortex. While the first is tangentially oriented and primarily responsible for the fronto-mastoid polarity reversal (Vaughan and Ritter, 1970), the second is radially oriented and probably originates in the superior temporal gyrus (Giard et al., 1994).

In order to gain more information about the putative spatial location of the AV interaction patterns that were elicited by the presentation of speech and music stimuli coupled with flashlights, we adopted the LORETA source estimation approach. As depicted in Table 1, LORETA consistently estimated current density maxima in auditory-related regions (BA 22 and BA 42). Furthermore, the fusion of syllables and flashlights was associated with putative sources in visual-related areas (BA 18). Since this estimation is compatible with the information arising from the topographic maps, our results provide clear evidence for the notion that the conjunction of syllables and piano tones with simple visual features modulated the activity of auditory-related regions. Otherwise, it should be considered that electrical brain imaging cannot be interpreted in an fMRI like manner. Nevertheless, it should be mentioned that this approach has been successfully used in a number of EEG studies pertaining to visual and auditory processing and has been shown to reliably estimate the respective sources (Esslen et al., 2004; Gottselig et al., 2004; Laufer and Pratt, 2005; Meyer et al., 2006; Mulert et al., 2002, 2004; Sinai and Pratt, 2003; Vitacca et al., 2002).

The findings arising from this study are novel in that they consistently indicate a modulation of auditory-related cortical regions during elementary AV speech and music processing. In addition, these results are in line with two previous electrophysiological studies (Fort et al., 2002; Giard and Peronnet, 1999) that investigated AV interactions in subjects who were instructed to recognize two objects defined by either visual features alone (i.e., disks), auditory features alone (i.e., pure tones), or by the conjunction of auditory and visual stimulus attributes. Along this vein, Giard and Perronet, as well as Fort and colleagues, revealed non-linear AV interactions between 90 and 140 ms post stimulus onset that corresponded in latency, polarity, and topography to the N1 component of auditory ERPs. Comparable results were also found in synaesthete subjects (Paulesu et al., 1995), monkeys (Watanabe and Iwai, 1991), and rats (Barth et al., 1995).

Furthermore, we revealed that the simultaneous presentation of piano tones and flashlights was associated with interaction patterns at electrode Fz. These additional interactions (See Figs. 3 and 4), in association with the estimated sources (See Table 1), indicate that additional regions residing proximate to the cingulate gyrus and in the frontal cortex may contribute to the synthesis of auditory and visual signals.
4.2. AV interaction patterns at around 200 ms

As a second main result, we revealed robust interaction patterns for both the AV music and speech objects at electrode Cz that occurred between 150 and 250 ms post stimulus onset (See Figs. 2 and 4). Even though the spatial resolution of ERPs does not allow us, with total confidence, to associate the effects measured over the scalp with the underlying electromagnetic dipoles, the LORETA source estimation approach suggests that putative current density maxima, which are located in extra-sensory regions, can be regarded as the neuronal signature for these significant interaction patterns (See Table 1). In fact, while the music condition was related to current-density maxima presumably originating from the cingulate, in the speech condition AV interaction patterns were rather associated with neuronal sources residing in the right insular cortex and in the medial frontal gyrus. These putative anatomical locations are consistent with previous reports, which showed the privileged role of extra-sensory regions in AV processing. To date, at least two electrophysiological studies (Giard and Peronnet, 1999; Molholm et al., 2002) and one neuroimaging (Calvert et al., 2001) investigation found similar evidence for a convergence of AV processing in frontal or cingular regions. Several foci of multimodal convergence in the frontal cortex have also been revealed in synaesthete individuals (Pauls et al., 1995; Walsh, 1996) and monkeys (Benevento et al., 1977; Ito, 1982). Furthermore, the insula is advocated as being an extra-sensory associative region that supports the integration of auditory information with other sensory functions (Bamiou, 2003; Bushara et al., 2001; Lewis et al., 2000; Calvert et al., 2001).

Notably, our results provide further evidence for the notion that only the music condition elicited sensory interaction patterns at electrode Cz during TW 2. In this context, the topographic maps depicted in Fig. 5 (i.e., DSyllables and DPiano) suggest that the neuronal sources underlying AV speech and music interaction patterns may qualitatively diverge. In fact, while the AV interaction patterns associated with the music condition (DPiano) showed a scalp distribution that roughly corresponds to that of pure auditory processing and, therefore, may have originated in the supra-temporal plane (See Table 1) (Vaughan and Ritter, 1970), the topographic maps associated with the speech condition (DSyllables) indicate a specific constellation of neuronal activation patterns that was not present during unisensory processing. These different topographic characteristics suggest that the neuronal sources associated with relatively late interaction patterns diverge as a function of the AV objects presented.

5. Limitations

In this study we adopted a block design. However, the incorporation of a completely randomized event-related design would have better controlled for attention variability over time and across subjects. Nevertheless, there are several excellent reasons that favor the applicability of a block design for investigating AV interactions. In fact, since the entire experiment lasted only 18 min and the task applied was not cognitively demanding, it isn’t plausible to assume that any attention effects may have influenced the data. Furthermore, by computing DWs between AV and aAV possible attention effects would have been washed out. Therefore, possible enhanced attention effects during the first two runs are widely eliminated in the DWs. Accordingly, the interaction patterns that we revealed in the present study may have been underestimated, as opposed to overestimated. Otherwise, even though it is improbable, we can not completely exclude that specific attention affects may have influenced the data in some directions.

In the present work we propose a novel approach for investigating AV interaction patterns. Even though several authors are the opinion that the criterion of superadditivity is a “too conservative” approach and that multisensory interactions can also arise independently from non-linearity (Stanford and Stein, 2007; Meredith and Stein, 1983, 1986), further studies are necessary for validating the approach we propose in this study.

Acknowledgments

The authors acknowledge ongoing research support from the Swiss National Science Foundation (Grants 320030-120661 to MM and 32003B-121927 to LJ) and from the “Fonds zur Förderung des akademischen Nachwuchses (FAN) des Zürcher Universitätsvereins (ZUNIV).

References


