Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study

Riikka Möttönen a, b,*, Martin Schürmann b, Mikko Sams a

a Laboratory of Computational Engineering, Helsinki University of Technology, Espoo, Finland
b Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology, Espoo, Finland

Received 25 November 2003; received in revised form 26 February 2004; accepted 24 March 2004

Abstract

During social interaction speech is perceived simultaneously by audition and vision. We studied interactions in the processing of auditory (A) and visual (V) speech signals in the human brain by comparing neuromagnetic responses to phonetically congruent audiovisual (AV) syllables with the arithmetic sum of responses to A and V syllables. Differences between AV and A + V responses were found bilaterally in the auditory cortices 150–200 ms and in the right superior temporal sulcus (STS) 250–600 ms after stimulus onset, showing that both sensory-specific and multisensory regions of the human temporal cortices are involved in AV speech processing. Importantly, our results suggest that AV interaction in the auditory cortex precedes that in the multisensory STS region.

Keywords: Audiovisual speech; Auditory cortex; Magnetoencephalography; Multisensory processing; Superior temporal sulcus

During audiovisual (AV) speech perception, seeing a talker’s articulatory gestures improves recognition of the speech sounds especially in noisy conditions [17]. Functional magnetic resonance imaging (fMRI) studies have shown that activity in the auditory cortex and in the multisensory regions of the superior temporal sulcus (STS) is modified during AV speech perception in comparison with activity during auditory (A) or visual (V) speech perception [1,2,15,20].

We used magnetoencephalography (MEG) to study the time course of multisensory interactions in different cortical regions during AV speech perception. Earlier interactions in the multisensory than in the sensory-specific cortices would support the idea that modulated activity in the latter is caused by feedback from higher-level multisensory integration sites [1,2]. An opposite order of onsets would suggest that interaction begins in the lower-level processing structures.

Eight Finnish-speaking volunteers with normal hearing and vision (self-reported) participated in the study (two females, one left-handed, 21–31 years).

The stimuli were produced by a Finnish AV speech synthesizer ‘Talking Head’ (TH). Observing articulatory gestures of this TH improves recognition of synthetic A speech in noise [9]. The TH produced /pa/ syllables with stimulus onset asynchronies of 1640, 2100 and 2570 ms in random order in the A, V and AV conditions. In the A condition, synthetic 250 ms acoustic syllables were presented binaurally through plastic tubes and earpieces and the subjects saw the back of TH (height 17 cm, 80 cm from the subject’s eyes) on the screen. In the V condition, the face of the TH was shown on the screen so that the articulatory gestures (movement duration 600 ms) were visible. In the AV condition, acoustic syllables were presented simultaneously (onsets synchronized). Low-intensity continuous white noise was presented in all conditions (signal-to-noise ratio 7.5). The order of conditions was randomized across subjects.

The neuromagnetic responses were recorded in a magnetically shielded room with a 306-channel whole-scalp neuromagnetometer (Neuromag Vectorview, Helsinki, Finland). The positions of four marker coils, placed on the scalp, were determined in relation to three anatomical landmark points with a 3D-digitizer to allow alignment of the MEG and MRI coordinate systems. The coil locations were determined in the beginning of each condition. About

www.elsevier.com/locate/neulet


0304-3900/03/$ - see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.neulet.2004.03.076
120 responses (passband 0.01–172 Hz, digitized at 600 Hz) to stimuli were collected in each condition and averaged using a 1200 ms time-window that began 300 ms prior to the stimulus onset. Epochs containing large-amplitude artefacts on MEG or EOG channels were automatically rejected. Anatomical T1-weighted MR images of the subjects’ brains were obtained with a 1.5 T scanner (Siemens, Erlangen, Germany).

The responses evoked in the AV condition and the arithmetic sum of responses evoked in the A and V conditions were compared. The differences between AV and A + V responses were interpreted to reflect interaction between A and V modalities, as in numerous previous studies [1–5,7,10,20]. The individual signals were pre-processed by omitting noisy channels, applying a 300 ms pre-stimulus baseline and a 30 Hz low-pass filter.

The sources of auditory evoked fields peaking at a latency of about 100 ms (N100m) were modelled as current dipoles, using data of the A condition. The head was modelled as a spherically symmetric conductor; for each subject the sphere parameters were optimized from individual MR images. The equivalent current dipoles (ECDs) at a peak latency of the N100m responses were identified by a least-squares search using a fixed set of 48 gradiometers over the left or right temporal cortices. Thereafter the analysis was extended to cover the entire time period and a larger set of channels (100 gradiometers over left and right temporal cortices) was included in computing a time-varying two-dipole model for A, AV and A + V signals. The strengths of the ECDs were allowed to change as a function of time, while the dipole locations and orientations were kept fixed both within and across conditions. The grand average source waveforms served to define time windows within which the AV and A + V source strengths differed from each other. The statistical significance of the difference between mean source strengths within these selected time windows was tested by Wilcoxon matched pairs tests.

Fig. 1A shows the sources of N100m responses of one subject superimposed on his MRIs. In all subjects these dipoles were located at the auditory cortical areas in the Sylvian fissure (BA 41/42, BA 22). The mean (± standard error, n = 8) of dipole locations in Talairach coordinates was −54 (±2.4), −22 (±1.4), 10 (±1.6) in the left hemisphere and 46 (±1.8), −16 (±1.8), 10 (±1.5) in the right hemisphere. These current dipoles were used to explain the activity within the supratemporal auditory cortices. The source waveforms of the current dipoles for AV and A + V signals are shown in Fig. 1B for one subject. Fig. 1C shows averages of the source waveforms across all eight subjects. The grand average AV and A + V source strengths differ from each other at 150–200 ms in both hemispheres. Fig. 1D shows the mean source strengths at 150–200 ms. The difference between AV and A + V source strengths was significant in both hemispheres (n = 8, P < 0.05). Also, AV and A source strengths differed significantly in both hemispheres (n = 8, P < 0.05). The mean A and A + V source strengths did not differ statistically significantly at this latency in either hemisphere.

To estimate the sources of AV − (A + V) signals minimum current estimates (MCEs) were calculated [18]. MCE is an implementation of the minimum L1-norm estimate [6] and it explains the measured signals with a current distribution that has the smallest sum of current amplitudes. The procedure was as follows. (1) Estimates of individual signals were calculated for each time point. (2) The individual estimates were aligned on a standard brain [11]. The alignment applies first a 12-parameter affine transformation [19], followed by a refinement with an elastic non-linear transformation [13]. (3) Aligned estimates were averaged across subjects. (4) From the AV − (A + V) grand average estimates, the most active areas were selected as regions of interest (ROIs). A ROI was accepted for further analysis if the peak source activity exceeded 1 nAm and the activation duration exceeded 20 ms. (5) The activity

Fig. 1. AV interaction in the auditory cortices. (A) The current dipoles for N100m responses of subject 8 superimposed on his MRIs. (B) Source waveforms of the current dipoles of subject 8 (grey: AV, black: A + V). (C) Grand average source waveforms of current dipoles (grey: AV, black: A + V). (D) The mean source strengths for A + V, AV and A signals at 150–200 ms in the left and right hemispheres.
Fig. 2. AV interaction in the right STS. (A) Grand average MCE calculated from AV − (A + V) signals at 250–600 ms. The ROI was determined on the basis of the right temporal activation (inside the white ellipse). (B) The centre point of the right temporal ROI (the red square) superimposed on the standard MRI shows that it is located in the right STS. (The black squares indicate the mean source locations of N100m responses.) (C) Activation time courses for AV and A + V in right STS. (D) Strengths of estimated activity in the right STS at 250–600 ms for AV and A + V signals in all eight subjects (black diamonds). The mean of activation strengths across subjects is indicated as a red diamond. (E) Mean strength (+SE, n = 8) of right STS activation in A, V and AV conditions at 250–600 ms.

Within a selected ROI was calculated for AV and A + V grand average estimates as a function of time and time windows during which the strength of AV and A + V activities differed were defined. (6) The strengths of individual AV, A + V, A and V activities within the ROI during selected time windows were compared to each other statistically by Wilcoxon matched pairs tests. (7) For the determination of cortical interaction areas, the centre points of selected ROIs were superimposed on the MRI of the standard brain and transformed to Talairach coordinates.

Fig. 2A presents the grand average activity in the right hemisphere estimated for AV − (A + V) signals at 250–600 ms. The ROI was defined on the basis of the activated area in the right temporal lobe (indicated by an ellipse). The centre point of the ROI was located in the ventral bank of the right STS (BA 21, Talairach coordinates: 55, −28, −13) 12 mm posterior and 23 mm inferior to the average source location of the right N100m response (Fig. 2B). The strength of the grand average activity for AV and A + V signals within the right STS is presented in Fig. 2C as a function of time. The activity for A + V was stronger than that for AV at about 250–600 ms. Fig. 2D shows the mean strength of activity for AV and A + V signals in all eight subjects during this period. In all subjects the strength of activity for AV was weaker than that for A + V (n = 8, P < 0.01). Fig. 2E shows the mean strength of activity in the right STS for A and V stimuli, when presented alone, and that for AV stimuli. Differences in strength of activity elicited by AV vs. either unimodal stimuli were not statistically significant.

The present results confirm those of previous studies by demonstrating modified activity in the supratemporal auditory cortices during AV speech perception [1,9,13]. The timing of the auditory-cortex AV interaction (150–200 ms) is in agreement with the previous MEG studies showing that a visual change in an AV speech stimulus sequence elicits a ‘mismatch response’ in the auditory cortex within 200 ms after stimulus onset [8,12]. In the event-related potential (ERP) study of Klucharev et al. [5] the earliest AV interactions in the auditory cortex were not sensitive to phonetic congruency vs. incongruency of AV vowels suggesting that they are not speech-specific. It remains open whether the auditory-cortex interactions in the present study are specific to speech. Furthermore, the present results do not show in which subregions of the auditory cortex AV interaction occurs.

There is evidence that V input to the auditory association cortex is mediated by feedback rather than feedforward connections in monkeys [14]. This input likely originates from the multisensory cortical regions. Furthermore, it has been proposed that A and V speech inputs would initially interact in the multisensory cortices (like STS) and that the activity in the sensory-specific cortices would be modulated by feedback connections from these integration sites [1,2]. The latter proposal is not supported by the present results which suggest that AV interaction in the right STS is clearly later (250–600 ms) than that in the auditory cortices (150–200 ms). This time course rather suggests that A and V speech inputs would initially interact in low-level cortical (or subcortical) structures. This interpretation is further supported by the (non-speech) ERP studies showing very early (onset within 50 ms after stimulus onset) AV interactions in the sensory-specific cortices [4,7]. However, our findings do not suggest via which pathways the V input reaches the auditory cortex.

The present findings indicate that the right STS region plays a role at a late stage of AV speech processing. Previous fMRI studies have also demonstrated multisensory interaction in the STS region during AV speech perception [2,15,20]. Calvert et al. [2] found enhanced activation (AV > A + V) in the posterior parts of the left STS during observation of congruent meaningful AV speech and suppressed activation (AV < A + V) during observation of incongruent AV speech. The lack of significant AV interaction in the left STS in the current study could be due to different stimulus material. Wright et al. [20] found both...
enhanced and suppressed activations in bilateral STS regions during observation of congruent AV words. In their MEG study, Raij et al. [10] found late suppressed activations in bilateral STS regions to AV letters.

Two interpretations are conceivable to the suppressive (AV < A + V) interaction effect in the right STS. (1) Simultaneous A and V speech signals could inhibit each other within a population of STS neurons. In animals, activity of multisensory neurons in the superior colliculi is strongly enhanced when A and V inputs are spatially congruent and suppressed when they are incongruent (for a review, see Ref. [16]). FMRI studies of AV temporal [3] and semantic congruence [2] have suggested that this principle would hold also in the human superior colliculi and STS. In contrast, our results suggest that multisensory suppression in the right STS neurons might play a role in integration of temporally, spatially and phonetically congruent speech signals. (2) Alternatively, as MEG measures activity of a large set of neurons, the suppressed MEG response to AV stimulation could also be caused by competition of unisensory inputs in a population of multisensory but not integrative STS neurons.

In conclusion, the time course of multisensory interactions found in the present study illuminates the processing sequence of AV speech stimuli by suggesting that interactions in the auditory cortices precede those in the multisensory regions of the temporal cortex.

Acknowledgements

We thank Professor Riitta Hari and Dr Iiro Jääskeläinen for the comments on the manuscript and Mika Seppälä for help in data analysis. The MRIs were obtained at the Department of Radiology, Helsinki University Central Hospital. This study has been supported by the Academy of Finland and the Sigrid Juselius Foundation.

References


