Neurocognition of speech-specific audiovisual perception

Mikko Sams, Ville Ojanen, Jyrki Tuomainen and Vasily Klucharev

Laboratory of Computational Engineering
Helsinki University of Technology
Mikko.Sams@hut.fi

Abstract

We have studied neurocognitive mechanisms of audiovisual perception of non-meaningful speech. We demonstrate different event-related brain potential (ERP) differences to non-speech and speech audiovisual stimuli. Our imaging studies show that viewing and listening to speech activate overlapping areas in the frontal cortex, which is also activated during speech production. We also demonstrate that subjects have to expect the auditory stimuli to be speech to integrate auditory and visual speech.

1. Introduction

Perception is often multisensory. For example, during conversation speech is perceived both by audition and vision. Viewing speech improves speech recognition especially in a noisy environment. Combining phonetically conflicting acoustic and visual speech stimuli may produce even new percepts (“McGurk effect”). When, e.g., the auditory /ba/ syllable is dubbed onto the visual articulation of /ga/, subjects typically hear a syllable /da/ [1]. Combining information from more than one modality is useful also in perceiving non-speech stimuli. For example, detection of bisensory stimuli is faster than that of its unisensory components. Integration of auditory and visual information is primarily based on temporal and spatial coincidence of the stimuli. Speech sounds and corresponding articulatory gestures are approximately simultaneous and arise from the same source. In addition, specific articulatory gestures and acoustic patterns provide also phonetic information. Therefore, both general and speech-specific multisensory mechanisms are important in audiovisual perception of speech.

Different brain structures have been suggested to be involved in different aspects of audiovisual integration. Detection of temporal synchrony activates insula, superior colliculus and claustrum. Intraparietal sulcus is sensitive to spatial discordance in the auditory and visual stimuli. Some features of audiovisual speech are integrated in the superior temporal sulcus (STS). Furthermore, during audiovisual speech perception activity is enhanced in sensory-specific cortices (for a review see Ref. [2].

Recent non-speech ERP studies on integration [3] have demonstrated very early multisensory interactions in the sensory-specific brain regions. With speech stimuli, interaction effects are usually later. Sams et al. [4] found that infrequent conflicting “McGurk stimuli” presented among frequent matching stimuli elicits a “mismatch response” at ~180 ms after stimulus onset in the auditory cortices. This responses is typically elicited automatically by infrequent deviant sounds. Mismatch response evoked by the McGurk stimuli suggests that illusory auditory changes are treated in the auditory cortices like real acoustic changes and that visual information from visible articulation has access to the auditory cortex within 200. MEG studies have also reported later integration effects in the STS [5].

Simultaneous acoustic and visual stimuli coming from the same location are naturally bound to audiovisual percepts. More abstract, phonetic and linguistic features need to be integrated in speech perception. In the following, we describe recent research from our laboratory in which we specifically studied neurocognitive mechanisms of audiovisual perception of non-meaningful phonetic stimuli.

2. ERP indicators of phonetic and non-phonetic audiovisual interactions during speech perception

We recently studied how phonetic and non-speech interactions are reflected in ERPs [6]. Stimuli were the Finnish vowels /a/, /o/, /i/ and /y/. Unisensory auditory stimuli were presented together with a static face on the screen. Unimodal visual stimuli were articulating faces without sounds. Congruent stimuli were audiovisual /a/ + /a/ and /i/, /o/ and /y/. Incongruent stimuli were acoustic /a/ + visual /y/, /y/ + visual /a/, acoustic /a/ + visual /o/, /i/, acoustic /i/ + visual /o/. These incongruent pairs did not cause a McGurk effect. In audiovisual tokens the first articulator movements started 95 ms prior to onset of the sound.

When subjects identified the stimuli according to what they heard, mean RTs to the incongruent audiovisual stimuli, auditory stimuli and congruent audiovisual stimuli were 681 ms, 611 ms and 571 ms.

The left side of Fig. 1 shows ERPs at three EEG channels to the congruent and incongruent stimuli, as well as the sum of ERPs to auditory and visual stimuli (A+V). ERPs to A+V and audiovisual stimuli differed significantly at three consecutive latencies. In each case, ERP to A+V appears larger than ERPs to audiovisual stimuli. The right side of Fig. 1 depicts ERPs to the congruent and incongruent stimuli. Also these responses differed at three latencies.

Our results demonstrate both non-phonetic and phonetic multisensory interactions during audiovisual speech perception. The sum of the ERPs to unisensory stimuli was different from the ERPs to both congruent and incongruent stimuli at three latencies. No differences between congruent and incongruent stimuli were found at these latencies. Therefore differences likely reflect interactions in general, non-phonetic, audiovisual processing. Timing and scalp distribution of the non-phonetic interactions support the involvement of sensory-specific cortices. These effects were stronger in the right hemisphere. ERPs to incongruent and congruent stimuli showed the earliest difference at ~150 ms. We suggest that these differences reflect phonetic-level
interactions, probably occurring in the heteromodal brain regions. Our findings support a view that both sensory-specific and heteromodal cortices are involved in the AV integration of speech. Furthermore, sensory-specific and heteromodal cortical regions are involved in the audiovisual interactions at separate latencies and are sensitive to different features of speech stimuli.

![Figure 1: Left side shows grand-averaged ERPs to congruent and incongruent stimuli, the sum (A+V) of the responses to the unisensory stimuli as well as difference waves (ERPs to A+V minus ERP to incongruent stimulus). Right side of the figure shows the ERPs to congruent and incongruent stimuli and corresponding difference waves (ERP to incongruent minus ERP to congruent stimulus).](image)

3. Broca’s area and processing of audiovisual speech

We further studied the differential processing of conflicting and matching audiovisual phonetic stimuli in an fMRI study, using the same stimuli as in the ERP experiment [7].

Figure 2 shows the contrast ‘incongruent audiovisual stimuli > congruent audiovisual stimuli’. Three distinct clusters of activation were detected: 1) Broca’s area in the left inferior frontal gyrus (BA 44/45), extending to the middle frontal gyrus (BA 9), 2) the left prefrontal cortex of the middle frontal gyrus (BA 10) and 3) the superior parietal lobule (BA 7). No statistically significant voxels were detected in the contrast ‘congruent audiovisual stimuli > incongruent audiovisual stimuli’.

Both congruent and incongruent audiovisual stimuli increased the activity in Broca’s area in comparison to the baseline. Significant activations in the prefrontal cortex and the superior parietal lobule were due to deactivation during the congruent audiovisual stimulation and activation during the incongruent audiovisual stimulation. These responses were not significant in comparison to the baseline but reached significance when contrasted with each other.

![Figure 2: Across-subjects z-statistic maps overlaid on an anatomical template.](image)

Broca’s area is involved in speech production but it participates also in various auditory speech processing tasks (e.g. [8]). Importantly, Broca’s area is also activated by visual speech (e.g. [9]). This area was activated more robustly by the phonetically incongruent than congruent audiovisual stimuli. The incongruent stimuli were not perceptually integrated, but produced two separate and simultaneous phonetic percepts. We suggest that the unisensory components of the incongruent stimuli activated distinct neural populations in Broca’s area. Phonetically congruent auditory and visual inputs probably converge on the same neurons. Therefore, the stronger activation by the incongruent stimuli might be due to processing of two instead of a single phonetic input.

We suggest that activations in the prefrontal cortex and the superior parietal lobule are related to some other aspects on information processing. These areas have been shown to be activated in tasks inducing response conflict and in attention tasks. Left-lateralized prefrontal activation has been shown when attention has to be divided between visual and auditory stimuli[10].

4. Brain areas activated by viewing and listening to speech

In a study described above we found that the speech motor areas are also activated by audiovisual phonetic stimuli. To study this issue further, we mapped the neural network activated by the unimodal visual and auditory vowels. We used the same vowel stimuli as in our previous fMRI experiment, but they were presented only unimodally.

Figure 3 shows brain areas that were activated by viewing and listening to speech. Viewing speech activated visual areas extending to the motion specific area V5/MT bilaterally (BA 17, 18 and 19), inferior parietal areas in the left postcentral gyrus (BA 43), left frontal area from the insula to the left middle and inferior frontal gyri, including Broca’s area (BA 44 and 45) and extending superiorly along the precentral sulcus to the lateral premotor cortex (BA 6) and middle frontal gyrus (BA 8 and 9). The lateral premotor cortex and middle frontal gyrus were activated also on the right side (BA 6 and 9). Activity can be seen also in the right middle and superior temporal gyrus/sulcus (BA 21, 22).
Figure 3: Across-subjects z-statistic maps overlaid on an anatomical template. Brain areas that were activated during viewing speech are indicated in red and those activated by hearing speech in blue (compared to the baseline condition). Areas activated by both viewing and listening to speech are shown in yellow.

Auditory speech activated the auditory processing areas bilaterally (BA 41, 42, 21, 22), the anterior cingulate gyrus (BA 32) extending to the medial frontal gyrus (BA 8 and 9), left frontal areas in the inferior and middle frontal gyri (BA 44, 45, 46, 47), extending superiorly to the precentral gyrus (BA 9), the anterior area of the left superior frontal gyrus (BA 10) and the insula, caudate nucleus and putamen bilaterally.

Areas which were activated by both viewing and listening to speech included the left frontal areas from insula, inferior and middle frontal gyri (BA 44, 45, 46, 47) to the precentral gyrus (BA 9), the inferior parietal areas (BA 43/40) and the posterior STG/STS in the right hemisphere.

The major finding was that the left insula, Broca’s area, the left lateral premotor cortex, and the inferior parietal area as well as the right STS/STG constitute a neural network common to visual and auditory speech processing. The results support the idea that the brain areas related to speech production provide a common representational space for auditory and visual speech and that the phonetic features of auditory and visual speech could be combined in this area.

5. Is audiovisual speech special?

We investigated whether subjects’ expectations about the nature of the auditory stimulus has an effect on audiovisual speech perception. Sine-wave replicas of Finnish nonwords /omso/ and /onso/ were presented to the subjects either alone or dubbed onto a visual display of a congruent or incongruent articulating face. In non-speech mode, the subjects were trained to classify the SWS stimuli in two arbitrary categories and were not told about their speech-like nature. In speech mode, the same subjects were trained to perceive the same SWS stimuli as speech. We studied whether subjects integrated the acoustic and visual signals in a similar way in these two modes. We hypothesized that if audiovisual speech perception is special, then integration would only occur when the subjects perceived the SWS stimuli as speech. For comparison, natural speech stimuli were also employed. The subjects reported how they heard the different auditory-only and congruent or incongruent audiovisual stimuli.

The results (Fig. 4) revealed that the subjects could accurately classify the auditory-only SWS stimuli in two categories both when they were perceived as non-speech and speech. Classification was as accurate as for auditory-only natural speech. In a visual-only control task without any auditory stimulus, subjects correctly identified 97% of the articulations.

Matching audiovisual SWS and natural speech stimuli were also classified very well. Conflicting audiovisual SWS stimuli in non-speech mode showed a slight decrease in correct identifications to 80% from 90% for the auditory-only stimuli. However, when the same SWS stimuli were perceived as speech, the effect of conflicting visual speech was much stronger, the correct identifications decreasing to 25% from 92% for the auditory-only stimuli, indicating strong audio-visual integration. For natural speech, the correct identifications decreased to 4% from 93%. The proportion of correct identifications of both natural speech (4%) and SWS stimuli in speech mode (25%) were significantly lower than those of SWS stimuli in non-speech mode (80%). Correct identifications of the natural speech and SWS stimuli in speech mode differed significantly from each other, but this difference was due to two subjects who, we suggest, did not enter the speech mode and therefore correctly identified 80% and 95% of the auditory component of the incongruent SWS stimuli.

These results demonstrate that acoustic and visual speech were integrated strongly only when the perceiver interpreted the acoustic stimuli as speech. If the SWS stimuli had been processed always in the same way, the influence of visual speech should have been the same in both speech and non-speech modes. We suggest that when SWS stimuli were perceived as non-speech, the acoustic and visual tokens did not form a natural multisensory object, and were processed more independently. When the SWS stimuli were perceived as speech, the acoustic and visual signals combined naturally to form a coherent phonetic percept. We interpret our present findings to be strong evidence for the existence of a specific speech mode of perception.

6. Discussion

We have shown evidence of different brain activations related to binding global and speech-specific features of audiovisual speech stimuli. Non-speech interactions were suggested to occur in sensory-specific brain areas. Our fMRI studies showed a strong participation of Broca’s area and frontal premotor speech areas in viewing and listening to speech, as well as to binding of phonetic features of audiovisual speech.

The involvement of Broca’s and motor speech areas in speech perception supports the motor theory of speech perception, which suggests that the objects of speech perception are the talker’s articulatory gestures [11]. We suggest that these areas contain motor representations of articulatory gestures into which both acoustic and visual
speech inputs are mapped during audiovisual speech observation.

The activation of Broca’s area and motor speech regions by auditory and visual speech has been argued to be related to the functioning of “mirror neurons”, which are activated both during execution actions and observation of similar actions performed by other individuals [12]. In humans, there is evidence of mirror neurons also in the primary motor cortex [13].

The mirror neurons have been suggested to play an important role in action recognition and speech communication [14]. Recent transcranial magnetic stimulation studies have found evidence of the activation of the speech production system during speech observation. Motor evoked potentials from articulatory muscles are enhanced during observation of visual and auditory speech when the left (but not right) motor cortex is stimulated [15]. Interestingly, in monkeys some of the mirror neurons are audiovisual, meaning that they are activated by both seen and heard actions [16].

Our psychophysical study showed that the subject’s expectation of the nature of the auditory stimuli is crucial for the audiovisual integration of speech. When subjects thought that the acoustic stimuli are not speech, viewing articulation did not alter perception. When they had learned that the acoustic stimuli were speech, they experienced a strong McGurk effect, providing evidence of integration. We suggest that our subjects attended to a different set of features depending on whether the SWS stimuli were perceived as speech or non-speech. In speech mode, they attended to those features which form a phonetic object. When the same stimuli were perceived as non-speech, attention was focused on some other features that could be used to discriminate the stimuli. Speech mode would then guide attention to speech-specific features both in auditory and visual stimuli, which are integrated only when they provide coherent information about a phonetic object.

7. Conclusions
Perception of speech is special in a sense that both viewing and listening to speech activates those brain areas which are also used in speech production. In addition, integration of auditory and visual speech requires that the subjects expect the stimuli to speech.

8. Acknowledgements
This research was funded by the Academy of Finland.

9. References


