Audiovisual discrimination of CV syllables: a simultaneous fMRI-EEG study

Cyril Dubois\textsuperscript{1,2}, Rudolph Sock\textsuperscript{2}

\textsuperscript{1}Romanisches Seminar, University of Zürich, Zürich, Switzerland
\textsuperscript{2}Institut de Phonétique de Strasbourg - Equipe de Recherche Parole et Cognition, University of Strasbourg, U.R. Linguistique, Langues et Parole (LiLPa) 1339, Strasbourg, France
cyril.dubois@uzh.ch, sock@unistra.fr

1. Abstract

We carried out a simultaneous fMRI-EEG experiment based on discriminating syllabic minimal pairs involving three phonological contrasts characterized by different degrees of visual distinctiveness (vocalic labialization, consonantal place of articulation or voicing). Audiovisual CV syllable pairs were presented either with a static facial configuration or with a dynamic display of articulatory movements. In the sound-disturbed MRI environment, the significant improvement of syllabic discrimination achieved in the dynamic audiovisual modality, compared to the static audiovisual modality was associated with activation of the occipito-temporal cortex (MT + V5) bilaterally, and of the left premotor cortex. MT + V5 was activated in response to facial movements independently of their relation to speech, the latter was specifically activated by phonological discrimination. Significant ERP’s to syllabic discrimination were recorded around 150 and 250 ms. Our results provide arguments for the involvement of the speech motor cortex in phonological discrimination, and suggest a multimodal representation of speech units.

Index Terms: Visual distinctiveness, motor cortex, MT+ V5, multimodal representation.

2. Introduction

In a noisy environment, visual perception of articulatory movements improves natural speech intelligibility. Parallel to phonemic processing based on auditory signals; visemic processing constitutes a counterpart based on “visemes”; the distinctive visual units of speech. Indeed, depending on the speech-to-noise ratio, word length and verbal processing constraints, congruent facial display provides critical cues to healthy listeners in noisy environments [1, 2], and benefits hearing-impaired listeners [3]. Prior conceptions involving a clear-cut distinction between speech production and perception and their respective underlying neural substrates have been questioned both by neuropsychological investigations and functional neuroimaging studies. Over the last decades, among other theories of speech perception, several models referred also to speech production [4-6]. Thus, the Motor Theory of Speech Perception made three major assumptions: first, that speech was processed by a specific module; second, that perceiving speech meant perceiving vocal tract gestures; and third, that the motor system was recruited for speech perception. Recent neuro-imaging studies found that simply listening to speech stimuli activated, in the frontal cortex, areas largely overlapping those activated when subjects actually produced similar speech stimuli [7]. These results support the view that speech perception implies an auditory-to-articulatory mapping process put forward by the Motor Theory of Speech Perception [8]. The recent discovery of the mirror neuron system, wherein neurons respond both when an action is performed by oneself and seen to be performed by someone else [9, 10], supports knowledge about the central role of motor acts of speech production in speech perception. The Perception-for-Action-Control Theory (PACT), also gained support from the existence of an action-dependent matching system put forward by the discovery of the mirror neuron system [23]. This framework is based on the assumption that action shapes perception and vice-versa. As a result, “perceptuo-motor” units do not result from pure motor action, inasmuch as speech gestures also rely on perceptual - acoustic-auditory and visual - values, which in turn are useful to speech production. Despite some controversy about the early lateralisation of speech signal processing and its functional significance, current models of auditory speech processing agree that parallel streams are specialised in the analysis of different aspects of speech signals, following distinct processing pathways, dedicated to mapping sounds onto meaning and another involved in mapping speech sounds onto motor representations of articulation [11-13]. The brain regions, which constitute these networks, encompass the superior temporal gyrus (STG) and middle temporal gyrus (MTG) including middle and/or posterior parts of the superior temporal sulcus (STS) bilaterally, as well as the premotor cortex or Broca’s region.

In order to gain further understanding of the respective contribution of visemic and phonemic information to speech perception, we compared the effect of static and dynamic visual cues in a syllable pair discrimination task. To complement event-related BOLD-contrast fMRI data with information about the time-course of dynamic AV speech processing, an electroencephalogram (EEG) was recorded simultaneously using interleaved EEG and fMRI measurements [14]. Several event-related potential (ERP) studies have investigated the neurophysiological correlates of AV integration in speech processing. They were mainly based on passive listening, as in [15] who studied the mismatch negativity (MMN) component of the auditory ERP resulting from the automatic detection of incongruent AV speech units. Other studies involving identification tasks [16] or recognition tasks [17], showed decreased N1 latency and amplitude when processing congruent bimodal AV speech stimuli vs unimodal auditory or visual stimuli. In view of our unusual paradigm, we did not expect to replicate the results of previous electrophysiological studies, but one may predict that dynamic visual cues should, to a degree, facilitate phonological discrimination and influence the latency...
and amplitude of the ERPs evoked by the discrimination of minimal phonetic features in such noisy background.

With respect to fMRI, we predicted increased activation in components constituting the network dedicated to auditory speech discrimination and, potentially involved in “visemic processing”. The PAC and STG, even the premotor cortex might be involved in this processing, as might the Sylvian parieto-temporal (Spt) area and the mid-posterior STS, considered respectively as sensory-motor and AV integration sites [18, 19]. In addition, when subjects were provided with dynamic visual cues to perform phonological discrimination tasks, we expected bilateral occipito-temporal activation (area MT + V5) related to visual movement processing.

3. Method

3.1. Subjects

The subjects were 30 adult French native speakers (mean age = 22.6; range = 18-27, 15 females), who reported no neurological or psychiatric deficit, with normal hearing and vision (or corrected to normal), and had not experienced any verbal learning disability. All subjects were recruited among students of the University of Strasbourg, and were right-handed according to the Edinburgh handedness inventory [20]. The volunteers first acquainted themselves with the MRI scanner environment by lying down within the scanner for a few minutes, and were then made aware of the purpose and conditions of the experiment, namely that they would be wearing a set of electrodes (in a magnetic environment). All participants gave their informed consent prior to their inclusion in the study, and were paid for their participation in the investigation. The local ethical committee approved the experimental procedure.

3.2. Stimuli

In the dynamic AV presentation modality, the stimuli consisted in a set of eight French syllables and eight non-phonological stimuli. The eight consonant-vowel (CV) syllables [pi] [py] [bi] [by] [ti] [ty] [di] [dy] were produced by a female native speaker of French and recorded in video files lasting for 360 ms each. Each video showed the bottom part of the speaker’s moving face (frame rate 25 images/s, audio sample rate 44.1 KHz in 16 bits) starting and ending with a closed mouth position. The non-phonological stimuli consisted in videos of natural syllables [fa] [fo] [va] [vo] played backward. Additionally, in order to obliterate any phonological cues, their audio part was altered using Audacity Software (http://audacity.sourceforge.net) by applying a low-frequency oscillation effect (5 Hz, start phase at 180°, deepness 75%, resonance factor 6, and periodicity 85%).

In the static AV modality, the same set of stimuli was used, but the moving face was replaced with a stilled face, randomly selected out of the video frames of the stimuli. In each pair, the two audio stimuli were presented with the same still.

3.3. Experimental procedures

The task consisted in forced-choice discrimination between pairs of different and similar AV stimuli. The trials lasting for 1120 ms consisted of two 360-ms AV stimuli displayed on a screen, with a 400-ms inter-stimulus interval (plain blue screen) between them. The subjects perceived the AV stimuli through prismatic goggles and headphones with high noise attenuation (AVOTEC SS-3100 silent scan). They did not wear earplugs in order to preserve the auditory perception. After each trial they had to decide whether the two stimuli were similar or different by clicking on one of a two-button mouse.

In each dynamic or static modality, the participants were presented with a total of 160 trials with pairs of different stimuli (40 of each category; one set of non-phonological and three of phonological pairs), besides 40 empty trials which served as null events. The trials were pseudo-randomly distributed using Optseq (http://surfer.nmr.mgh.harvard.edu/optseq) which automatically optimises the order and timing of events for event-related fMRI experiments [21], on the basis of the duration and sampling time of the hemodynamic responses, the number and types of stimuli and null events. In each modality of presentation, the task was performed within 24 minutes (360 4-second trials), the complete acquisition lasted for 48 minutes.

The three sets of phonological pairs were designed to examine the influence of visual speech cues relevant to three classes of phonetic features, categorized as corresponding to three different visemes. The first contrast between the spread [i] vs. rounded [y] vowels, was based on the feature of labiality. The second contrast was relevant to the place of articulation, i.e. labial [p b] vs. coronal consonants [t d] and the last, voicing-based contrast, opposed voiced [b d] vs. voiceless consonants [pt]. We will refer to these contrasts as “labiality”, “place of articulation” and “voicing”. The three contrasts were chosen in order to compare different levels of visual distinctiveness (labiality = or > place of articulation > voicing). Accuracy and response times (RT) measured from the onset of the second syllable in each trial were recorded.

3.4. MRI/EEG acquisition

MR images were acquired using a Bruker 2T 8200 scanner (Bruker Medical GmbH, Ettlingen, Germany), equipped with an SK330 insert gradient coil (30 mT/m, 150 µs rise time) and a radio-frequency head coil. The repetition time was set to 4 sec, encompassing a 1.8-s gradient artefact-free window for interleaved EEG recording. After discarding the first three volumes used to reach signal equilibrium in each modality of presentation, a total of 360 volumes (320 volumes with presentation of pairs of similar or different stimuli and 40 volumes without stimuli) were acquired.

EEG signals were acquired continuously during fMRI acquisition using a magnetic resonance-compatible system (EMR32: Schwarzer, Munich, Germany), equipped with a digital signal processor board, which also received three synchronized trigger inputs: from the stimuli, the MRI volumes’ onset dating, and a separately amplified electrocardiogram ECG channel (Physiogard, Bruker, SARL, Wissenbourg, France). A set of 23 Ag/AgCl electrodes, with iron-free copper leads fixed individually using EEG paste (Elefix, Nikon Khoden) served to record EEG data and eye movements. EEG signals were recorded from 19 electrodes positioned according to the international 10-20 system (F7, F3 Fz, F4, F8, FC3, FC4, T3, C3, Cz, C4, T4, CP5, CP6, T5, P3, Pz, P4, T6) with nasion reference, and the three remaining electrodes monitored horizontal and vertical eye movements, for subsequent off-line validation (i.e. rejection of artefact-ridden recordings).

Due to technical constraints, ERPs were exclusively computed for the second stimulus of each pair.
4. Results

4.1. fMRI results

In the dynamic AV presentation mode, correct discrimination of phonological as well as non phonological pairs activated a neural network involving bilaterally the STG, MTG (BA 22/41and 21) and the occipito-temporal region (BA 19 and 37) in the right hemisphere. Additional activation in the left premotor cortex (BA 6) was observed as subjects were engaged in the discrimination of phonological pairs contrasting with respect to voicing or place of articulation. In the static AV presentation mode, significant activation involved only the STG and MTG bilaterally (see figure 1).

![Figure 1: Activation areas to correct discrimination of the phonological contrasts, in the dynamic (top) and static (bottom) AV presentation. Activations were significant at p FEW-corrected < 0.005, with an extent threshold of 25 voxels.](image)

4.2. ERP results

In the dynamic AV modality, correct discrimination of syllable pairs elicited significant positive waves within the 125-175 ms window, for the three phonological contrasts, at Fz and Pz electrodes, respectively. Significant negative waves were recorded within the 225-275 ms window at Fz and Cz electrodes. In the static AV modality, correct discrimination of syllable pairs elicited a significant positive wave at the Fz electrode in the 125-175 ms window only for the voicing-based contrast. Within the 225-275 ms window, significant negative waves were recorded at different electrode sites for the different phonological contrasts, as follows: at Fz and Cz electrodes for voicing; at Cz for the place of articulation and at CP5 for labiality. In the 125-175 ms window the difference in amplitude as a function of the contrast, i.e. labiality, which elicited waves of higher amplitude at the Fz and Pz electrodes in the dynamic vs static AV modality. As to the 225-275 ms window, the difference in negative ERP amplitudes as a function of the modality of stimulus presentation was significant for all three contrasts, which, in the dynamic AV mode, elicited at the Fz electrode waveforms of higher amplitudes as follows: (labiality: $T = 4.67 \ p < 10^{-4}$; voicing: $T = 2.95 \ p < 0.009$; and place of articulation: $T = 3.37 \ p < 0.008$). The latter additionally elicited a negative waveform of increased amplitude at Pz (see figure 2).

![Figure 2: ERPs to correct discrimination of the three phonological contrasts, in the dynamic (grey line) and static (black line) AV presentation. The time window is from -100 ms to 600 ms, and the vertical line indicates the onset of the second item of contrasting pair.](image)

5. Discussion

In accordance with current functional anatomic models of speech processing, in the present study discrimination of phonetic features consistently activated bilaterally the STG and MTG, including the mid-posterior STS, in dynamic as well as static AV presentation modes. However, when the stimuli were presented dynamically, the discrimination task activated additional brain regions, i.e. the right occipito-temporal cortex (BA 19/37) recruited for all pairs of stimuli, and the left Premotor cortex (BA 6) for pairs contrasting with respect to voicing or place of articulation. While the premotor cortex, a probable part of the mirror neuron system, was specifically activated by dynamic speech articulation, activation of the MT+ V5 area did not depend on whether facial movements were associated or not with speech. The fact that activation in the premotor cortex does not rely on the visual distinctiveness of speech movement, and may be observed as a result of speech listening in absence of any visual cue [7], suggests that its involvement could be related to motor speech representations rather than visemic processing per se. In accordance with current functional anatomic models of speech processing, in the present study discrimination of phonetic features consistently activated bilaterally the STG and MTG, including the mid-posterior STS, in dynamic as well as static AV presentation modes. However, when the stimuli were presented dynamically, the discrimination task activated additional brain regions, i.e. the right occipito-temporal cortex (BA 19/37) recruited for all pairs of stimuli, and the left Premotor cortex (BA 6) for pairs contrasting in respect of voicing or place of articulation.

As shown by electrophysiological data, the cues related to articulatory speech configuration provided by dynamic AV presentation induced significant changes in evoked potentials around 150 ms and 250 ms after the onset of the second syllable of each pair. In order to discuss the ERP data of this study, one should bear in mind the particular context of signal acquisition. ERP analyses focused on the identification of significant waveforms allowing comparisons across static and dynamic AV modalities. Two significant waveforms peaking in the 125-175 ms and 225-275 ms temporal windows were identified. A positive wave over Fz, Cz, Pz in the 125 - 175 ms window was recorded in response to the correct discrimination of syllables contrasting in voicing, in the static AV modality, and in response to the correct discrimination of all three phonological contrasts, in the dynamic AV modality. The influence of visemic processing on syllabic discrimination in a noisy environment occurred as early as 150 ms following the acoustic stimulus.
onset. As to the significant negative wave recorded over the vertex sites in the 225-275 ms temporal window, elicited in response to correct discrimination of the three phonological contrasts regardless of the modality of stimulus presentation, it was consistent with the N250. N250 might be a suitable candidate to study auditory perception in subjects with specific language-impairment [22]. Our study did not evidence faster auditory speech processing. Indeed, in both conditions the syllable discrimination task involved AV stimuli, and improved behavioural performances, reported for the vocalic labiality features, only consisted in an increased accuracy rate without significantly reduced response time.

6. Conclusion

In line with an early study [22], our results suggest that the more the auditory speech component is degraded, the more relevant congruent visual cues become for syllabic discrimination. As shown by electrophysiological data, the cues related to articulatory speech configuration provided by dynamic AV presentation induced significant changes in evoked potentials as early as 150 ms and around 250 ms after the onset of the second syllable of each pair. Our fMRI data showed that the visual movement area (MT + V5) and the premotor cortex were involved in visemic processing. This view would be in accordance with theoretical frameworks advocating multimodal speech representations based on perceptuo-motor features, as postulated by the Perception-for-Action-Control Theory [23].

7. Acknowledgements

We thank the CIC of the Hôpitaux Universitaire de Strasbourg for examining the healthy volunteers. The work was supported of the French Ministry of Research (Grant ACI TTT, Mesures et Données, 2003 – 2006; ANR (DOCVACIM) », 2008-2012) as well as Maison Interuniversitaire des Sciences de l’Homme d’Alsace (MISHA).

8. References


