Neuroscience Letters 452 (2009) 219–223

Contents lists available at ScienceDirect

Neuroscience Letters

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

Two cortical mechanisms support the integration of visual and auditory speech: A hypothesis and preliminary data

Kayoko Okada, Gregory Hickok*

Department of Cognitive Sciences University of California, Irvine, Irvine, CA 92697-3800, United States

ARTICLE INFO

Article history:
Received 25 August 2008
Received in revised form 21 January 2009
Accepted 23 January 2009

Keywords:
Planum temporale
fMRI
Sensory–motor integration
Lip-reading
Speech-reading
Visual speech
Audio-visual speech

ABSTRACT

Visual speech (lip-reading) influences the perception of heard speech. The literature suggests at least two possible mechanisms for this influence: "direct" sensory–sensory interaction, whereby sensory signals from auditory and visual modalities are integrated directly, likely in the superior temporal sulcus, and "indirect" sensory–motor interaction, whereby visual speech is first mapped onto motor-speech representations in the frontal lobe, which in turn influences sensory perception via sensory–motor integration networks. We hypothesize that both mechanisms exist, and further that previous demonstrations of lip-reading functional activations in Broca's region and the posterior planum temporale reflect the sensory–motor mechanism. We tested one prediction of this hypothesis using fMRI. We assessed whether viewing visual speech (contrasted with facial gestures) activates the same network as a speech sensory–motor integration task (listen to and then silently rehearse speech). Both tasks activated locations within Broca's area, dorsal premotor cortex, and the posterior planum temporal (Spt), and focal regions of the STS, all of which have previously been implicated in sensory–motor integration for speech. This finding is consistent with the view that visual speech influences heard speech via sensory–motor networks. Lip-reading also activated a much wider network in the superior temporal lobe than the sensory–motor task, possibly reflecting a more direct cross-sensory integration network.

In face-to-face conversations, we are sensitive not only to acoustic cues in the speech signal but also to the visual cues present in a speaker's face. Lip-reading or speech-reading, is the ability of humans to understand speech by observing the lip and mouth gestures of the speaker. Visual cues provide linguistic information beyond analysis of facial expressions and can facilitate speech perception when an auditory signal occurs in a noisy environment or is degraded by noise [9,14,42]. Further, when audio and visual speech information are mismatched (hearing "ba" while watching someone articulate "ka"), this can induce a perceptual illusion, the McGurk effect [27], whereby the acoustic speech information is incorrectly perceived. It is quite clear, therefore, that auditory and visual speech information interact in perception.

Recent neuroimaging work investigating the neural correlates of visual speech perception implicate auditory and language-related regions in the superior temporal lobe including the superior temporal sulcus (STS), and in some reports, primary auditory cortex [10,13,11,34]. While the STS is activated in virtually all studies, primary auditory areas are less reliably reported [4,32]. In most studies, speech-reading also elicits activity of left lateralized or bilateral inferior frontal gyrus (IFG) and premotor cortex [11,32,16,26,38,39] regions classically implicated in speech production.

A major conclusion coming from the work on the neural basis of audio-visual speech perception is that visual speech has its influence on the perception of heard speech via multisensory integration, and that the posterior STS is a critical region in this respect. For example, Calvert et al. [12] showed that a region in the posterior STS showed supra-additive responses to audio-visual speech (AV>A alone + V alone), which these authors considered to be a signature of multisensory integration [12], although this claim is controversial [25]. This view, that visual speech modulates auditory perception of speech via multisensory integration in the STS, aligns well with evidence from studies in both human and non-human primates that implicates the posterior STS in cross-sensory integration [2,3,1,29,41]. A recent study recording single unit and local field potentials in monkeys while they perceived audio-visual monkey vocalizations provided direct evidence for the influence of STS activity on responses in auditory cortex, at least in monkeys, which supports this hypothesis [18,24].

Although it seems clear that multisensory integration in the STS is a major contributor to audio-visual interactions in speech perception, it may not be the only source as suggested by both behavioral and neuroimaging evidence [32,39]. Behavioral evidence for this possibility comes from a study that found that a McGurk-like effect can be induced, not only by *viewing* incongruent speech gestures,

* Corresponding author. Tel.: +1 949 824 1409; fax: +1 949 824 2307.
E-mail address: greg.hickok@uci.edu (G. Hickok).

0304-3940/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.neulet.2009.01.060

but by the listener's own incongruent speech gestures [36]. Listeners silently articulated speech sounds that were either congruent or incongruent with the syllables they were listening to. The incongruent condition led to significantly more misperceptions of the heard speech (32% correct) than the congruent condition (95% correct) suggesting that motor representations of speech can influence sensory perception of speech sounds. It has been suggested that the source of this influence is via efferent copies of motor commands that are transmitted to auditory regions, and that this process may form a kind of predictive (forward model) mechanism that modulates the analysis of sensory input [39,36,35]. This purely motor effect, however, appears to be substantially weaker than visually-induced misperceptions, as the Sams et al. [36] study found that when subjects viewed an audio-visual mismatch (the standard McGurk effect), performance dropped to 6% correct. Viewing one's own incongruent articulations in a mirror produced intermediate results (17% correct), showing that the addition of visual information associated with the same self-articulation resulted in additional performance decrement, which in turn suggests that visual information has an added influence beyond efferent motor copies.

Imaging evidence is consistent with the idea that some of the interaction between auditory and visual speech may be mediated by the motor system in that motor-speech related frontal structures, such as portions of Broca's area, typically activate during visual speech perception (see above). It is relevant that while both the posterior STS and Broca's area are activated during visual speech perception, their response properties are different under some conditions. For example, Miller and D'Esposito [28] have shown that the posterior STS responds more to audio-visual speech that is perceived as fused than audio-visual speech that is desynchronized and perceived as unfused. Broca's region showed the reverse pattern, responding more to unfused audio-visual speech, and with a later activity peak. Although it is not entirely clear how to interpret the details of these response patterns, it does suggest that pSTS and Broca's area are performing different kinds of computations on audio-visual speech stimuli.

Given the evidence reviewed above, we hypothesize the existence of two routes by which visual speech can influence auditory perception of speech sounds. One is via direct sensory–sensory integration in which visual speech information is integrated with auditory speech information in the STS via projections from sensory input systems [12,18]. The other route is via the motor system. Following previous authors [32,39,36,35], we hypothesize that visual speech gestures activate motor networks associated with articulating the visually perceived gestures, which in turn send efferent copies to sensory cortices via sensory–motor integration circuits, thus exerting an influence on perception. The behavioral evidence reviewed above suggests that this sensory–motor route is the weaker of the two.

The goal of the present study was to test one prediction of this hypothesis, namely that the perception of visual speech should activate networks known to be involved in sensory–motor integration for speech. A number of studies have investigated sensory–motor integration circuits for speech [6,23,7,8]. A network of brain regions has been identified that have both sensory and motor response properties in the speech domain, including portions of the STS, a region at the posterior most aspect of the Sylvian fissure at the parietal–temporal boundary (Spt), a portion of Broca's area (BA44 in particular), and a more dorsal premotor site in the frontal lobe.

The present study used standard paradigms for identifying sensory–motor activations for speech and for identifying cortical areas responsive to visual speech to determine if these two tasks activate partially overlapping networks. Such a finding would be consistent with the proposal that visual speech can influence heard speech via a sensory–motor mechanism.

Twenty six participants (10 females) between 18 and 44 years of age were recruited from the University of California, Irvine (UCI) community and received monetary compensation for their time. The volunteers were right-handed, native English speakers with normal or corrected-to-normal vision, no known history of neurological disease, and no other contraindications for MRI. Informed consent was obtained from each participant prior to participation in the study in accordance with guidelines from the UCI Institutional Review Board which approved this study. Two subjects were omitted from data analysis due to excessive head motion and one subject was omitted for failing to follow task instructions.

The data reported in this study were part of a larger experiment aimed at mapping responses to a range of sensory stimuli, including melodic sequences, noise bursts, auditory speech, and visual speech. Here we focus only on the speech conditions. The auditory speech stimuli were 3 s jabberwocky sentences (e.g., *It is the glandour in my nedderop*) in which content words were replaced with nonsense words, as used in previous experiments [23]. Visual speech stimuli were silent video clips of a male face articulating six visually distinguishable CV syllables (ba, tha, va, bi, thi, vi). Following Calvert et al. [10], we also presented video clips of six lower-face non-speech gestures that were used as a control to isolate speech-reading activations. The non-speech gestures were: partial opening of the mouth with leftward deviation, opening of mouth with rightward deviation, opening of mouth with lip protrusion, tongue protrusion, lower lip biting, and lip retraction.

Subjects were randomly presented with 15 s blocks of music, speech, noise bursts or videos in equal ratios. Subjects were instructed to monitor for "oddball" stimuli throughout the study, where oddball was defined relative to the type of stimuli that constituted a given block. Oddball stimuli for the speech trials were a change in speaker voice from male to female, and for videos, a change in the actor from male to female. Subjects pressed a button each time an oddball was detected. There were three oddball trials within each session and in total this comprised approximately 13% of the experiment. These trials were modeled as a regressor of no interest and excluded from the results.

On some trials, subjects were simultaneously presented with 3 s of jabberwocky sentences and a picture of either an ear or lips which stayed on screen for 15 s. If a picture of lips was presented, subjects were instructed to rehearse the jabberwocky sentence until the lips went off screen. If a picture of an ear was presented, they were instructed to simply pay attention to the 3 s jabberwocky sentence and "rest" during the remainder of the 15 s block. The listen–rehearse condition has been shown to drive activity in sensory–motor regions of the posterior planum temporale (Spt) [23], and the listen–rest condition served as a control for the effects of acoustic stimulation alone. Thus "sensory–motor" activations are defined as regions that respond both during the perception of speech (a sensory response) and during covert rehearsal of speech (a "motor" response). Previous studies have confirmed that area Spt activates during more conventional motor-speech tasks such as picture naming [19,30].

The experiment started with a short exposure session to familiarize subjects with all of the different experimental stimuli. Subjects were scanned during the exposure session to ensure they could comfortably hear the stimuli through the scanner noise, and to acclimatize them to the fMRI situation. This was followed by five experimental sessions (runs). Each experimental session contained an equal number of trials (blocks) of each condition. Each trial was 15 s in length and a single scanning session was approximately 6 min long. Auditory stimuli were presented through MR compatible headset and stimulus delivery and timing were controlled using Cogent software (http://www.vislab.ucl.ac.uk/cogent_2000.php) implemented in Matlab 6 (Mathworks, Inc, USA).