Please cite this article in press as: Komeilipoor N et al. Involvement of superior temporal areas in audiovisual and audiomotor speech integration. Neuroscience (2016), http://dx.doi.org/10.1016/j.neuroscience.2016.03.047

Neuroscience xxx (2016) xxx–xxx

INVOLVEMENT OF SUPERIOR TEMPORAL AREAS IN AUDIOVISUAL AND AUDIOMOTOR SPEECH INTEGRATION

N. KOMEILIPOOR, [a,b] **P. CESARI** [b] **AND A. DAFFERTSHOFER** [a]*

[a] *MOVE Research Institute Amsterdam, Faculty of Behavioural and Movement Sciences, Vrije Universiteit, Van der Boechorststraat 9, 1081BT Amsterdam, The Netherlands*

[b] *Department of Neurological, Biomedical and Movement Sciences, University of Verona, 37131 Verona, Italy*

Abstract—Perception of speech sounds is affected by observing facial motion. Incongruence between speech sounds and watching somebody articulating may influence the perception of auditory syllable, referred to as the McGurk effect. We tested the degree to which silent articulation of a syllable also affects speech perception and searched for its neural correlates. Listeners were instructed to identify the auditory syllables /pa/ and /ta/ while silently articulating congruent/incongruent syllables or observing videos of a speaker's face articulating them. As a baseline, we included an auditory-only condition without competing visual or sensorimotor input. As expected, perception of sounds degraded when incongruent syllables were observed, and also when they were silently articulated, albeit to a lesser extent. This degrading was accompanied by significant amplitude modulations in the beta frequency band in right superior temporal areas. In these areas, the event-related beta activity during congruent conditions was phase-locked to responses evoked during the auditory-only condition. We conclude that proper temporal alignment of different input streams in right superior temporal areas is mandatory for both audiovisual and audiomotor speech integration. © 2016 Published by Elsevier Ltd. on behalf of IBRO.

Key words: EEG, McGurk effect, multisensory integration, sensorimotor interaction, superior temporal gyrus.

INTRODUCTION

The brain receives a continuous stream of information from different sensory modalities. Proper integration of input is essential for accurate perception. The perception of speech sound is clearly affected by observation of facial motion: incongruent visual input caused sound perception to degrade, as the visual input may affect the perception of auditory syllable. This is referred to as the McGurk effect (McGurk and MacDonald, 1976). The McGurk effect has inspired many researchers investigating multisensory integration (Tiippana, 2014). The perception of a sound syllable can also be affected by tactile stimulation (Gick and Derrick, 2009; Ito et al., 2009).

The identification of auditory syllables can be either degraded or improved when the listeners silently articulate incongruent or congruent syllables, respectively, as well as when they observe others producing those syllables (Sams et al., 2005; Mochida et al., 2013; Sato et al., 2013). Sams et al. (2005) suggested that both effects may rely on the same neural mechanism and may be due to modulation of the activity in auditory cortical areas. Functional magnetic resonance imaging (fMRI) studies indicated that lip reading modulates activity of the auditory cortex (Calvert et al., 1997). Visual speech may hence affect the auditory perception by altering activation of auditory cortical areas. Likewise, magnetoencephalography (MEG) studies suggest a modulation of activity in the auditory cortex during both silent and loud reading (Numminen et al., 1999; Kauramäki et al., 2010; Tian and Poeppel, 2010) as well as silent articulation (Numminen and Curio, 1999) and lip reading (Kauramäki et al., 2010). Interestingly, the responses were weaker for covert speech as compared to silent reading (Numminen et al., 1999), in lip reading and covert speech compared with a visual control and baseline tasks (Kauramäki et al., 2010) and during silent articulation as compared to speech listening (Numminen and Curio, 1999). It has been suggested that the auditory suppression during speech might be due to the existence of an efference-copy pathway from articulatory networks in Broca's area to the auditory cortex via the inferior parietal lobe (Rauschecker and Scott, 2009). Thus, the effect of observing and articulating incongruent syllables on the perception of auditory syllables (Sams et al., 2005; Mochida et al., 2013; Sato et al., 2013) may be ascribed to their impact on alteration of activities in auditory areas, which interferes speech perception.

A further way to conceive the neuronal underpinning of multisensory perception is to consider it as a result of multimodal neurons activity processing inputs from different sensory modalities. In mammals, such multisensory cell assemblies are presumably located at

*Corresponding author. Address: MOVE Research Institute Amsterdam, Faculty of Behavioural and Movement Sciences, Vrije Universiteit, Van der Boechorststraat 9, 1081BT Amsterdam, The Netherlands.
E-mail address: a.daffertshofer@vu.nl (A. Daffertshofer).
Abbreviations: BEM, boundary element method; DICS, dynamic imaging of coherent sources; EEG, electroencephalography; EMG, electromyography; MEG, magnetoencephalography; MNI, Montreal Neurological Institute; PLV, phase-locking value; STS/STG, superior temporal sulcus/gyrus.

http://dx.doi.org/10.1016/j.neuroscience.2016.03.047

multiple neural levels in mammals, from midbrain to cortex (Stein and Stanford, 2008). Regarding the McGurk effect, neuroimaging revealed an involvement of the superior temporal sulcus/gyrus (STS/STG) (Calvert et al., 2000; Jones and Callan, 2003; Sekiyama et al., 2003; Bernstein et al., 2008; Irwin et al., 2011; Nath and Beauchamp, 2012, Szycik et al., 2012; Erickson et al., 2014). A number of recent papers considered the dynamic interplay of neural populations as a key to cross-modal integration (Senkowski et al., 2008; Arnal et al., 2009; Arnal and Giraud, 2012). The superior temporal area is considered a multisensory convergence site as it receives inputs from unimodal auditory and visual cortices and contains multisensory neurons (Karnath, 2001). However, what precisely happens in this area to accomplish multisensory integration and whether it is responsible for the reported effect of silent articulation on auditory perception (e.g. Sams et al., 2005) is still largely unclear.

For the present study, we capitalized on the competition between auditory and visual inputs as well as between auditory and sensorimotor inputs to probe how cortical oscillations contribute to multisensory integration. We adopted a protocol recently introduced by Mochida et al. (2013), in which listeners are instructed to identify auditory syllables while silently articulating congruent/incongruent syllables, or observing videos of a speaker's face articulating congruent/incongruent syllables. Cortical activity was monitored using electroencephalography (EEG).

Consistent with the McGurk effect (McGurk and MacDonald, 1976), we expected, when dubbing the acoustic syllable /pa/ onto the visual presentation of articulatory gestures of /ta/, subjects to typically misperceive the sound. We also expected a similar result when subjects themselves silently articulated an incongruent syllable (Sams et al., 2005; Mochida et al., 2013; Sato et al., 2013). Furthermore, we expected source localization of EEG to reveal STS/STG as the area discriminating between proper and improper perception, in support with the aforementioned imaging studies. Finally, we hypothesized the phase dynamics in STS/STG to be essential for multisensory integration, as we believe that temporal alignment of distinct sensory streams is key to their integration.

EXPERIMENTAL PROCEDURE

Subjects

Twelve volunteers (mean age 26.1 years, five females) participated after giving their written informed consent. All were right handed and had normal hearing and normal or corrected-to-normal vision.

Protocol

The experimental protocol has been adopted from a recent study by Mochida et al. (2013). The ethics committee of the Faculty of Human Movement Sciences, VU University Amsterdam had approved it prior to conduction.

Task. Participants were asked to identify the syllables (/pa/ and /ta/) that they heard among the four possible alternatives (/pa/, /ta/, /ka/, or 'etc') displayed on the screen under the following subtask conditions: silently articulating congruent/incongruent syllables (*motor condition*), observing videos of a speaker's face articulating congruent/incongruent syllables (*visual condition*), and a condition without a subtask (*baseline condition* or *auditory only*); see Fig. 1 for overview. In the *motor condition*, participants were instructed to articulate the syllables with as little vocalization as possible while moving the lips and tongue as much as possible and to identify the syllables that they heard. Under the *visual condition*, subjects were required to indicate what they heard while they were presented with audiovisual stimuli. In the *baseline* (*auditory-only*) *condition*, participants were asked to listen to the syllables while watching a still frame of the video and choose the heard syllable after they were presented.

Stimuli. Stimuli had been produced by a Dutch male speaker. We recorded conventional videos at 50 Hz frame rate and they were edited in iMovie 10.0. Audio signals were digitized at a rate of 44.1 kHz. They were delivered at a level of 60 dB through paired speakers placed in front of the participants (distance 55 cm to the participant's torso) and were separated by approximately 30 cm. We superimposed white noise to the syllables (signal-to-noise ratio of 5 dB) to create ambiguity and reduce word recognition accuracy (Sato et al., 2013). Beginning and end of the noise were faded in and out, respectively (0.5 s duration). Syllables were preceded by four clicks (0.67 s inter-click interval) to provide a cue for silently articulating a syllable in the motor condition.

For the *visual conditions*, auditory syllables were paired with the videos of a speaker's face producing either congruent or congruent syllables yielding four different combinations: (i) congruent /pa/ (visual /pa/ auditory /pa/), (ii) congruent /ta/ (visual /ta/ auditory /ta/), (iii) incongruent stimuli (visual /pa/ auditory /ta/) and (iv) the converse incongruent stimuli (visual /ta/ auditory /pa/). Similar to *visual conditions*, in the *motor conditions*, silent articulation of congruent/incongruent syllables paired with the auditory syllables produced four conditions: (i) congruent /pa/ (articulation of /pa/ auditory /pa/), (ii) congruent /ta/ (articulation of /ta/ auditory /ta/), (iii) incongruent combination (articulation of /pa/ auditory /ta/) and (iv) the converse incongruent combination (articulation of /ta/ auditory /pa/).

In the *motor condition*, English characters representing /pa/ or /ta/ were presented on a front display (LCD monitor, frame rate 60 Hz, about 55 cm in front of the participant's nasion) until the participants pressed the space bar of a computer keyboard to start the trial. They were asked to silently articulate the indicated syllable in time with the clicks and the onset of the syllable while watching a still frame of the video.

For the *visual condition*, a video of the speaker's face articulating either /pa/ or /ta/ was presented on the front display. Prior to video presentation, the initial frame of

Please cite this article in press as: Komeilipoor N et al. Involvement of superior temporal areas in audiovisual and audiomotor speech integration. Neuroscience (2016), http://dx.doi.org/10.1016/j.neuroscience.2016.03.047