Neuropsychologia 63 (2014) 85–91

Contents lists available at ScienceDirect

Neuropsychologia

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

Vision of tongue movements bias auditory speech perception

Alessandro D'Ausilio [a,*], Eleonora Bartoli [a], Laura Maffongelli [a], Jeffrey James Berry [a], Luciano Fadiga [a,b]

[a] Robotics, Brain and Cognitive Sciences Department, The Italian Institute of Technology, Via Morego, 30, 16163 Genova, Italy
[b] Section of Human Physiology, University of Ferrara, Via Fossato di Mortara, 17/19, 44100 Ferrara, Italy

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form
15 July 2014
Accepted 15 August 2014
Available online 27 August 2014

Keywords:
Speech
Audiovisual speech perception
McGurk
Sensorimotor map
Motor theories
Action perception

ABSTRACT

Audiovisual speech perception is likely based on the association between auditory and visual information into stable audiovisual maps. Conflicting audiovisual inputs generate perceptual illusions such as the McGurk effect. Audiovisual mismatch effects could be either driven by the detection of violations in the standard audiovisual statistics or via the sensorimotor reconstruction of the distal articulatory event that generated the audiovisual ambiguity. In order to disambiguate between the two hypotheses we exploit the fact that the tongue is hidden to vision. For this reason, tongue movement encoding can solely be learned via speech production but not via others' speech perception alone. Here we asked participants to identify speech sounds while matching or mismatching visual representations of tongue movements which were shown. Vision of congruent tongue movements facilitated auditory speech identification with respect to incongruent trials. This result suggests that direct visual experience of an articulator movement is not necessary for the generation of audiovisual mismatch effects. Furthermore, we suggest that audiovisual integration in speech may benefit from speech production learning.

1. Introduction

Stable and reliable multisensory representations can be achieved by the natural alignment of information, from different modalities, related to the same event. Asynchrony in audio-visual temporal alignment can be detected in a variety of multimodal stimuli (speech, music and object action; Vatakis & Spence, 2006), indicating that we are particularly sensitive to violations in the temporal correlations. Intriguingly, participants also infer causal relationships from temporal correlation between audio and visual events (Parise, Spence, & Ernst, 2012). During every day communication the auditory information produced by a speaker is often temporally coupled with the visual information arising from visible articulators, such as lips. In most cases, building such stable correlations between speech audio–visual signals can aid perception in ecological scenarios. For instance, vision of the articulators enhances accurate auditory perception in noise (Sumby & Pollack, 1954). More generally, visible speech influences perception both by integrating under-specified acoustic information and by making perception more robust through redundancy (Campbell, 2008).

Otherwise, perturbation of the normal spatio-temporal alignment between audio and visual cues can induce illusory percepts. For example, in the ventriloquism effect, when auditory and visual information come from different spatial sources we tend to illusorily displace sounds towards the visual source (Pick, Warren, & Hay, 1969). On the other hand, if auditory (i.e. /ba/) and visual (i.e. /ga/) information do not match, an illusory perception such as the McGurk effect (McGurk & MacDonald, 1976) may arise. In this case, participants perceive a third syllable (/da/ or /tha/). The McGurk illusion is generally seen as a landmark demonstration of how previous learning affects the analysis and integration of multimodal speech stimuli.

Generally speaking, this effect is due to a perturbation of the learned auditory and visual speech-related association. This illusion is quite robust to even large temporal asynchronies (Munhall, Gribble, Sacco, & Ward, 1996) or spatial manipulations (Jones & Munhall, 1997), and is elicited without participants being aware of the task (Alsius & Munhall, 2013). However, one key question since the pioneering work of McGurk and McDonald was how much this effect is a by-product of being exposed to a stable multisensory environment providing repeated and reliable audiovisual correlation. In fact, during development, the repeated co-occurrence and match of audio and visual information was thought to build reliable statistics of the environment. In this sense, different age-spans were investigated (McGurk & MacDonald, 1976; Massaro, 1984) leading to the finding that the effect in children is somewhat weaker than in adults. These initial observations suggested that the McGurk effect was at least partially driven by a form of experience-dependent learning of the audiovisual statistics of the environment.

[*] Corresponding author. Tel.: +39 10 71781975; fax: +39 10 7170817.
E-mail address: alessandro.dausilio@iit.it (A. D'Ausilio).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.08.018
0028-3932/© 2014 Elsevier Ltd. All rights reserved.

In the following years a series of studies showed that infants are indeed affected by the visual cues present during speech perception. Four-month-old infants, show preference for the face that matches an auditory vowel (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Similarly, two-month-old infants detect the correspondence between the auditory and visually perceived speech information (Patterson & Werker, 2003). This could be explained by the fact that audiovisual matching could arise from at least three partially independent feature sets, including temporal cues (Vatakis & Spence, 2006), energetic cues (Grant, van Wassenhove, & Poeppel, 2004) and phonetic cues (Kuhl et al., 2006). Infants do not seem to rely on phonetic cues (Baart, Vroomen, Shaw, & Bortfeld, 2014; Jusczyk, Luce, & Charles-Luce, 1994) whereas temporal and spectral ones might be employed for early audiovisual correspondence detection in speech. Nevertheless, all these studies confirm that some form of multimodal matching of audiovisual speech already exist in pre-linguistic children (Rosenblum, Schmuckler, & Johnson, 1997; Burnham & Dodd, 2004), thus suggesting a partial independence with respect to their linguistic environment.

Generally speaking, the literature seems to suggest that some basic form of pre-linguistic audiovisual statistical association can be acquired very early in life. The critical problem concerns the nature of this audiovisual association. Specifically, the question is if active vocal exploration plays some role in the acquisition of these associations or if passive (rather limited) exposure to environmental audiovisual speech statistics is able to provide enough information. Along this line, a recent study used a sort of reversed McGurk effect. Adult participants heard speech sounds and at the same time had to judge the shape of “mouth-like” ellipses (Sweeny, Guzman-Martinez, Ortega, Grabowecky, & Suzuki, 2012). The clever use of ellipsoidal visual stimuli should in theory avoid the automatic and direct association with the visual representations of mouth shapes stored in memory. However, the authors suggest two alternative hypotheses for the biasing effect that auditory syllables had on shape judgments. One is that participants were able to grasp the statistical association that exists between speech sounds and the mouth visual shapes to produce that sound (from now on called audiovisual hypothesis). These audiovisual associations, favored by the perceptual similarities between mouth configurations and ellipse shapes, are substantially analogous to the previously outlined auditory and visual correlation, we are tuned to detect since infancy. Such a hypothesis thus predicts that passive exposure to environmental audiovisual speech statistics can cause the effect.

Alternatively, the effect could be due to the correspondence emerging from the automatic transformation of auditory and visual information, into articulatory movements in the motor system (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; from now on called sensorimotor hypothesis). The general mechanism could be that suggested by the analysis by synthesis approach (Stevens & Halle, 1967). This model proposed that the perception is derived from the computational re-creation of the input (Bever & Poeppel, 2010). Such a synthetic process regenerates the input by means of an abstract motor code without specifying detailed acoustic, visual or motor correlates. Liberman's motor theory, instead, insisted on a more specified motor program. The main difference being that the reconstruction envisioned by the motor theory, implies an internal representation of actual vocal movement. Here the driving factor might be the capability to extract, from abstract visual stimuli such as ellipses, basic sensorimotor primitives learned from active vocal production.

Unfortunately, in this study as well as in many audiovisual integration studies, both accounts are equally probable. No conclusion can be drawn in favor of either the audiovisual or sensorimotor hypotheses (Spence & Deroy, 2012). One possible solution to discriminate between the two hypotheses might instead be the study of adults' behavior on material for which no reliable audiovisual statistics is present. We propose that watching articulators, for which we have no visual experience such as the tongue could be the key aspect.

The tongue is indeed a critical articulator, hardly visible in its full motion and target configurations. During speech perception we at most barely see just the anterior tip of the tongue and thus a rather loose temporal association with the auditory effects of tongue movements. In fact, even if the most anterior part of the tongue can be partly exposed during speech production (if the jaw opening is somewhat exaggerated), the tongue back motion, which is the critical component for the /ga/ or /ka/ syllable (velar constriction), is in contrast always occluded. However, it is still possible that some correlation could be picked up between the posterior motion (inferred) and the anterior tongue information (partly visible). Nevertheless, such a correlation is almost absent since the anterior tip motion, for velar sounds, is much more variable than the tongue back, as for any non-critical articulatory feature (Papcun et al., 1992; Canevari, Badino, Fadiga, & Metta, 2013). During speech production, instead, we exploit tightly coupled proprioceptive, tactile, motor and auditory cues associated with tongue motor control. Therefore, we can get access to accurate tongue kinematics knowledge solely through tongue movement learning. It is important to stress that such sensorimotor knowledge does not necessarily need to be learned via speech production but rather can also emerge from non-speech tongue motor control.

In the present study, we capitalize on the fact that tongue motion is concealed from vision by running two behavioral experiments. More specifically, participants had to identify auditory syllables while we visually presented real tongue movements recorded with an ultrasound imaging technique. Visual stimuli showed a sagittal profile of a tongue producing a syllable that was either matching or mismatching with the auditory stimuli. Our prediction is that if the audiovisual hypothesis is true then we should see no difference between matching and mismatching audiovisual presentations. In fact, there is no statistical audiovisual association between speech sounds and visible tongue shapes. Furthermore, there is no perceptual similarity between visible mouth configurations and tongue movement (as it was the case for Sweeny et al., 2012). On the other hand, if the visual presentation of tongue movements induces a significant bias on auditory perception then, the sensorimotor hypothesis is more likely to be true. In fact, it would demonstrate that in principle, learning the statistics of the audiovisual environment cannot account for such a multimodal effect but rather we need to get access to knowledge that can be acquired solely via tongue movement control.

2. Materials and methods

2.1. General methods

Visual stimuli consisted of short video clips showing the sagittal profile of a tongue (See Fig. 1) articulating different syllables. Since ultrasound images can be very noisy, the tongue dorsal profile was enhanced by drawing a red line on top of it. Video clips were utterances of a female speaker producing /ba/, /ga/, /pa/, and /ka/ syllables. Frame-to-frame differences in pixel intensity (0 for black and 1 for white pixels) were measured to check for a possible bias in global visual motion between stimuli (sum of absolute differences: /ba/=52,105; /ga/=57,020; /pa/=61,185; /ka/=60,588). In fact, the ultrasound probe captures information (i.e. background or tongue body) that is not necessarily conveying information about articulatory gestures. In this sense, we computed whole image statistics about global motion, to control for spurious (non-gestural) movement differences and thus exclude the contribution of low-level global visual feature identification. Paired *t*-tests between anterior and posterior articulated stimuli were not significant (mean motion and standard deviation: /ba/=1914.72 std=694.62; /ga/=2097.03 std=853.9; /pa/=2225.83 std=710.67; /ka/=2220.84 std=782.16;