Neuropsychologia 109 (2018) 126–133

Contents lists available at ScienceDirect

Neuropsychologia

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

Electrophysiological evidence for Audio-visuo-lingual speech integration

Avril Treille[a], Coriandre Vilain[a], Jean-Luc Schwartz[a], Thomas Hueber[a], Marc Sato[b,*]

[a] GIPSA-lab, Département Parole et Cognition, Université Grenoble Alpes & CNRS, Grenoble, France
[b] Laboratoire Parole et Langage, Aix-Marseille Université & CNRS, Aix-en-Provence, France

ARTICLE INFO

Keywords:
Speech perception
Audio-visual integration
Visual tongue feedback
Action perception
EEG

ABSTRACT

Recent neurophysiological studies demonstrate that audio-visual speech integration partly operates through temporal expectations and speech-specific predictions. From these results, one common view is that the binding of auditory and visual, lipread, speech cues relies on their joint probability and prior associative audio-visual experience. The present EEG study examined whether visual tongue movements integrate with relevant speech sounds, despite little associative audio-visual experience between the two modalities. A second objective was to determine possible similarities and differences of audio-visual speech integration between unusual audio-visuo-lingual and classical audio-visuo-labial modalities. To this aim, participants were presented with auditory, visual, and audio-visual isolated syllables, with the visual presentation related to either a sagittal view of the tongue movements or a facial view of the lip movements of a speaker, with lingual and facial movements previously recorded by an ultrasound imaging system and a video camera. In line with previous EEG studies, our results revealed an amplitude decrease and a latency facilitation of P2 auditory evoked potentials in both audio-visual-lingual and audio-visuo-labial conditions compared to the sum of unimodal conditions. These results argue against the view that auditory and visual speech cues solely integrate based on prior associative audio-visual perceptual experience. Rather, they suggest that dynamic and phonetic informational cues are sharable across sensory modalities, possibly through a cross-modal transfer of implicit articulatory motor knowledge.

1. Introduction

Audio-visual speech perception is a specific case of multisensory processing that interfaces with the linguistic system. Like most natural perceptual events in which information from different sensory sources is merged, bimodal integration of the acoustic and visual speech signals depends on their perceptual saliency, their spatial and temporal relationships, as well as their predictability and joint probability to occur (Campbell and Massaro, 1997; Jones and Munhall, 1997; Green, 1998; Schwartz et al., 2004). When combined to the acoustic speech signal, visual information from the speaker's face is known to enhance sensitivity to acoustic speech information by decreasing auditory detection threshold, and to improve auditory speech intelligibility and recognition, notably when the acoustic signal is degraded/noisy (Sumby and Pollack, 1954; Benoît et al., 1994; Grant and Seitz, 2000; Schwartz et al., 2004). Audio-visual speech perception is also known to facilitate the understanding of a semantically complex statement (Reisberg et al., 1987) or a foreign language (Navarra and Soto-Faraco, 2005), and to benefit hearing-impaired listeners (Grant et al., 1998). Besides the studies demonstrating a perceptual gain for bimodal compared to unimodal speech perception, one of the most striking evidence for Audio-visual speech integration is the so-called McGurk illusion, when adding incongruent visual movements interferes with auditory perception and creates an illusory speech percept (McGurk and MacDonald, 1976).

Complementing these psychophysical and behavioral findings, a number of neurophysiological studies have provided new advances in the understanding of Audio-visual speech binding, its neural architecture and the time course of neural processing. One major finding is that activity within both unisensory auditory and visual cortices as well as the posterior superior temporal sulcus (pSTS) is modulated during Audio-visual speech perception when compared with auditory and visual speech perception (Calvertet al, 2000; Callan et al., 2003, 2004; Skipperet al, 2005; Skipper et al., 2007). Since the pSTS displays supra-additive and sub-additive haemodynamic responses during congruent and incongruent Audio-visual speech perception, it has been proposed that visual and auditory speech cues are integrated within this heteromodal brain region (Calvertet al, 2000; Beauchamp et al., 2004). Complementing this finding, it has been consistently shown that adding lip movements to auditory speech modulates activity quite early in the supratemporal auditory cortex, with the latency and amplitude of N1/M1 and/or P2 auditory evoked responses attenuated and speeded-up during Audio-visual compared to unimodal speech perception

[*] Correspondence to: Laboratoire Parole et Langage, UMR 7309 CNRS & Aix-Marseille Université, 5 avenue Pasteur, 13100 Aix-en-Provence, France.
E-mail address: marc.sato@lpl-aix.fr (M. Sato).

https://doi.org/10.1016/j.neuropsychologia.2017.12.024
Received 27 September 2017; Received in revised form 21 November 2017; Accepted 13 December 2017
Available online 14 December 2017

(Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Huhn et al., 2009; Pilling, 2009; Vroomen and Stekelenburg, 2010; Winneke and Phillips, 2011; Frtusova et al., 2013; Schepers et al., 2013; Stekelenburg et al., 2013; Baart et al., 2014; Ganesh et al., 2014; Kaganovich and Schumaker, 2014; Treille et al., 2014a, 2014b, 2017a; Baart and Samuel, 2015; Hisanaga et al., 2016; Paris et al., 2016; for a recent review and discussion, see Baart, 2016). The latency facilitation of auditory evoked responses, but not the amplitude reduction, also appears to be directly function of the visemic information, with the higher visual recognition of the syllable, the larger latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009). In light of these studies, recent theoretical proposals postulate a fast direct feedforward neural route between motion-sensitive and auditory brain areas that helps tuning auditory processing to the incoming speech sound, thanks to the available information from the speaker's articulatory movements that precede sound onset in these studies (Chandrasekaran et al., 2009; but see Schwartz and Savariaux, 2014),[1] and a slower and indirect feedback pathway from the posterior superior temporal sulcus to sensory-specific regions that functions as an error signal between visual prediction and auditory input (Hertrich et al., 2007; Arnal et al., 2009).

The above-mentioned studies and theoretical proposals support the view that Audio-visual speech integration partly operates through visually-based temporal expectations and speech-specific predictions. This can be encompassed in a more general Bayesian perspective, with auditory and visual speech cues likely integrated based on their joint probability distribution derived from prior associative Audio-visual perceptual experience (for recent discussions, see van Wassenhove, 2013; Rosenblum et al., 2016). A number of experimental data however pose a challenge to this probabilistic perceptual account. Indeed, bimodal speech interaction has been shown to occur not only for well-known auditory and lipread, visuo-labial, modalities but also for other modalities with little, if any, associative perceptual experience.

One first example comes from a set of behavioral and electrophysiological studies showing that bimodal speech interaction can occur between auditory and haptic modalities, even with participants inexperienced with the haptic speech modality. In these studies, orofacial speech gestures were felt and monitored from manual tactile contact with the speaker's face. When the auditory and haptic modalities were presented simultaneously, a felt syllable affected judgment of an ambiguous auditory syllable, and vice-versa (Fowler and Dekle, 1991). In case of noisy/degraded acoustic speech signal, adding the haptic modality enhanced recognition of the auditory speech stimulus (Gick et al., 2008; Sato et al., 2010a). A similar perceptual gain was also observed when adding the haptic modality to lipreading (Gick et al., 2008). Further, audio–haptic McGurk-type illusion has been also observed (Fowler and Dekle, 1991; but see Sato et al., 2010a for inconclusive results). Finally, two recent electro-encephalographic studies showed that N1/P2 auditory evoked potentials are speeded up and attenuated not only during Audio-visuo-labial but also during audio-haptic speech perception, when compared to unimodal auditory perception (Treille et al., 2014a, 2014b). By providing evidence for cross-modal influences between auditory and haptic modalities, for a perceptual gain for audio-haptic compared to unimodal speech perception, and for cross-sensory speech modulation of the auditory cortex, these studies draw an exquisite parallel between Audio-visual and audio-haptic speech perception. Given that participants were inexperienced with the haptic speech modality, they clearly argue against the view that prior associative bimodal, and even unimodal, speech perceptual experience is needed for the two sensory sources to interact.

Other tactile stimuli can also affect heard speech. When applying in synchrony a small, inaudible, puff of air to the skin of participant's hands, neck (Gick and Derrick, 2009), or ankles (Derrick and Gick, 2013), the auditory perception of aspirated and unaspirated syllables embedded in white noise is more often perceived as an aspirated syllable (causing participants to mishear /ba/as/pa/, or/da/as/ta/). These results suggest that perceivers integrate tactile-relevant information during auditory speech perception without prior training and even without frequent or robust location-specific experience. A final example comes from a study by Ito et al. (2009) who showed that the identification of ambiguous auditory speech stimuli can be modified by stretching the facial skin of the listener's mouth, thanks to a robotic device that induced cutaneous/kinesthetic changes, and that perceptual changes only occur in conjunction with speech-like patterns of skin stretch. A subsequent study showed the reverse effect, with the somatosensory perception of facial skin stretch modified by auditory speech sounds (Ito and Ostry, 2012).

Altogether, these haptic and tactile instances of multisensory speech perception provide strong support for a supramodal view on multisensory speech perception. They nicely exemplify the way lawful and speech-relevant information from many distinct sources, including one hardly uses at all, can be extracted to give rise to an integrated speech percept. From these findings, in an attempt to reconcile them with a Bayesian, associative probabilistic account of multisensory perception, speech theorists have argued that prior experience and learning should be sharable across modalities, and that dynamic and phonetic informational cues available across sensory modalities partly derive from the listener's knowledge of speech production (Fowler, 2004; Rosenblum et al., 2016). This appears in line with the longstanding, albeit debated, proposal of a functional coupling between speech production and perception systems in the speaking and listening brain, and a common currency between motor and perceptual speech primitives (Liberman et al., 1967; Liberman and Mattingly, 1985; Fowler, 1986; Liberman and Whalen, 2000; Galantucci et al., 2006; Skipper et al., 2007; Rauschecker and Scott, 2009; Schwartz et al., 2012; Skipper et al., 2016).

The present electroencephalographic (EEG) study capitalizes on these findings and theoretical proposals with the aim of determining whether visual tongue movements, which are audible but not visible in daily life, might integrate with relevant speech sounds. A second objective was to examine possible similarities and differences of Audio-visual speech integration between unusual Audio-visuo-lingual and classical Audio-visuo-labial modalities. To this aim, participants were presented with auditory, visual, and Audio-visual isolated syllables, with the visual presentation related to either a sagittal view of the tongue movements or a facial view of the lip movements of a speaker, with lingual and facial movements previously recorded by an ultrasound imaging system and a video camera. In line with previous EEG studies, Audio-visual integration was estimated using an additive model (i.e., $AV \neq A + V$; for a recent review, see Baart, 2016) by comparing the latency and amplitude of N1/P2 auditory evoked potentials in both the Audio-visual-lingual and Audio-visuo-labial conditions with the sum of those observed in the unimodal conditions.

Audio–motor association for tongue movements is frequently experienced in daily life (for instance, when speaking or eating). However, despite implicit articulatory motor knowledge on tongue movements, only a few recent studies explored the influence of visual tongue movements on heard speech. Using virtual tongue movements or ultrasound images of tongue movements, they showed that visual tongue feedback can strengthen the learning of novel speech sounds (Katz and Mehta, 2015) and enhance and/or speed up auditory speech discrimination when compared with unimodal auditory or incongruent

[1] One highly relevant assumption for lip-read-induced predictions is that the visual speech signal precedes the auditory one and helps to predict auditory onset variability depending on visual saliency. As a matter of fact, the material choice in almost all of these studies consisted on isolated syllables in which the visual speech signal preceded the acoustic speech signal by tens and even hundreds of milliseconds (Chandrasekaran et al., 2009), also leading to a maximal temporal bimodal integration window (van Wassenhove et al., 2007; Venezia et al., 2016). However, it should be noted that in more ecological and naturalistic situations, with continuous speech, the temporal relationship between auditory and visual speech onsets appears more variable and spans a range of 30–50 ms auditory lead to 170–200 ms visual lead (Schwartz and Savariaux, 2014).