Brain & Language 139 (2014) 36–48

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Audiovisual integration for speech during mid-childhood: Electrophysiological evidence

Natalya Kaganovich [a,b,*], Jennifer Schumaker [a]

[a] Department of Speech, Language, and Hearing Sciences, Purdue University, Lyles Porter Hall, 715 Clinic Drive, West Lafayette, IN 47907-2038, United States
[b] Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2038, United States

ARTICLE INFO

Article history:
Accepted 30 September 2014

Keywords:
Audiovisual speech perception
Electrophysiology
Child language development

ABSTRACT

Previous studies have demonstrated that the presence of visual speech cues reduces the amplitude and latency of the N1 and P2 event-related potential (ERP) components elicited by speech stimuli. However, the developmental trajectory of this effect is not yet fully mapped. We examined ERP responses to auditory, visual, and audiovisual speech in two groups of school-age children (7–8-year-olds and 10–11-year-olds) and in adults. Audiovisual speech led to the attenuation of the N1 and P2 components in all groups of participants, suggesting that the neural mechanisms underlying these effects are functional by early school years. Additionally, while the reduction in N1 was largest over the right scalp, the P2 attenuation was largest over the left and midline scalp. The difference in the hemispheric distribution of the N1 and P2 attenuation supports the idea that these components index at least somewhat disparate neural processes within the context of audiovisual speech perception.

1. Introduction

In the majority of cases, our experience of the world is multisensory in nature, and as children mature, they learn to match, detect various correspondences between, and integrate perception from different senses. Accumulating research suggests that these various sub-components of what is commonly referred to as "multisensory processing" may rely on at least partially disparate brain regions and have different developmental trajectories (e.g., Burr & Gori, 2012; Calvert, 2001; Stevenson, VanDerKlok, Pisoni, & James, 2011).

On the one hand, some ability to match and combine different modalities appears to be present during the first year of life (for the purposes of this study, we will review only research on audiovisual processing because it is most relevant to speech perception; however, a number of studies examined development of interaction between other modalities as well (for a review, see Burr & Gori, 2012; Gori, Sandini, & Burr, 2012; Jaime, Longard, & Moore, 2014)). For example, 10–16 week old infants can detect temporal synchrony between lip movements and speech sounds (Dodd, 1979). By 3 months of age, infants can learn arbitrary pairings between faces and voices (Bahrick, Hernandez-Reif, & Flom, 2005; Brookes et al., 2001), and by 4–7 months of age, they are able to match faces and voices based on age (Bahrick, Netto, & Hernandez-Reif, 1998). Additionally, multiple studies reported the presence of at least some degree of audiovisual integration in infants as young as 4.5–5 months of age as revealed by their perception of the McGurk illusion (in which, typically, an auditory 'pa' or 'ba' is dubbed onto a visual articulation of 'ka' or 'ga', resulting in the perception of 'ta' or 'da') (Burnham & Dodd, 2004; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum, Schmuckler, & Johnson, 1997). By recording event-related potentials (ERPs) to the auditory pronunciation of a vowel that either matched or mismatched the earlier visual articulation, Bristow and colleagues reported that 10–12 week old infants appear to have a cross-modal representation of phonemes and integrate auditory and visual information during early stages of perception, similar to what had previously been reported for adults (Bristow et al., 2008). Indeed, some visual speech skills of young infants may even surpass those of adults. As an example, Weikum and colleagues have demonstrated that 4- and 6-month-old (but not 8-month-old) infants are able to discriminate between two languages based on visual speech cues alone (Weikum et al., 2007).

On the other hand, a number of studies point to a protracted developmental course of certain aspects of audiovisual processing. For example, Lewkowicz reported that, compared to adults, infants require significantly longer separations between the onsets of auditory and visual stimuli, both in speech and non-speech

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, Purdue University, Lyles Porter Hall, 715 Clinic Drive, West Lafayette, IN 47907-2038, United States. Fax: +1 (765)494 0771.
E-mail address: kaganovi@purdue.edu (N. Kaganovich).

http://dx.doi.org/10.1016/j.bandl.2014.09.011
0093-934X/

contexts, in order to detect temporal asynchrony (Lewkowicz, 1996; Lewkowicz, 2010). In fact, sensitivity to audiovisual temporal offsets remains immature even during mid-childhood (Hillock, Powers, & Wallace, 2011; Hillock-Dunn & Wallace, 2012; Kaganovich, Schumaker, Leonard, Gustafson, & Macias, 2014). Further, although some ability to perceive the McGurk illusion appears to be present early on, multiple studies have documented a reduced susceptibility to the McGurk illusion in children compared to adults, suggesting that the ability to fully integrate auditory and visual speech cues does not mature until late childhood and depends, at least in part, on children's experience with visual speech (Massaro, 1984; Massaro, Thompson, Barron, & Lauren, 1986; McGurk & MacDonald, 1976; Tye-Murray, Hale, Spehar, Myerson, & Sommers, 2014). Behavioral benefits of audiovisual processing, such as faster reaction time to audiovisual as compared to auditory only or visual only stimuli and better speech-in-noise perception, also either begin to manifest themselves or continue to improve during mid-childhood (Barutchu, Crewther, & Crewther, 2009; Barutchu et al., 2010; Ross et al., 2011).

Although a number of studies have begun charting developmental trajectories of various audiovisual skills beyond infancy, the absolute majority of studies on audiovisual processing have been done with either infants or adults, and relatively little is known about how its development unfolds during mid-childhood. Additionally, the bulk of studies on audiovisual processing during school years rely on behavioral paradigms, which may be an unreliable measure of audiovisual processing in this population for several reasons. First, even when children's behavioral responses on an audiovisual task are the same as those of adults and suggest the presence of multisensory processing, the neural circuitry engaged by the task may nonetheless be different in children. As an example, the left posterior superior temporal sulcus (STS) has been shown to play an important role in audiovisual speech perception and in the ability to perceive the McGurk illusion in particular (Beauchamp, Nath, & Pasalar, 2010; Calvert, 2001; Nath & Beauchamp, 2012; Nath, Fava, & Beauchamp, 2011). However, a study by Nath et al. (2011) reported that, compared to adults, some children had weak STS responses even when they perceived the McGurk illusion, thus deviating substantially from the adult pattern of neural activity associated with audiovisual processing of speech. Second, behavioral responses are the end result of many sensory, cognitive, and motor processes. Therefore, a lack of multisensory facilitation in reaction time, accuracy, or other behavioral measure may potentially be the result of the immature motor system, overall greater variability of responses in younger research participants (e.g., McIntosh, Kovacevic, & Itier, 2008; Williams, Hultsch, Strauss, Hunter, & Tannock, 2005), or the inability to apply audiovisual skills to a specific task. Lastly, audiovisual integration happens over different stages of speech processing, such as for example acoustic, phonemic, or lexical (Baart, Stekelenburg, & Vroomen, 2014; Hertrich, Mathiak, Lutzenberger, Menning, & Ackermann, 2007; Kaiser, Kirk, Lachs, & Pisoni, 2003). However, behavioral studies typically cannot provide information about the timing and nature of the neural mechanisms engaged by the task. Unlike behavioral measures alone, the ERP method, with its ability to track neural activity on a millisecond basis, allows one to focus on specific stages of audiovisual processing, often without a need for overt behavioral responses.

In the present study, we took advantage of a well-established electrophysiological paradigm in order to examine an early stage of audiovisual integration in school-age children and adults (Besle, Bertrand, & Giard, 2009; Besle, Fort, Delpuech, & Giard, 2004; Besle, Fort, & Giard, 2004; Besle et al., 2008; Giard & Besle, 2010; Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014). The paradigm is based on the fact that during sensory processing (i.e., within approximately 200 ms post-stimulus onset), ERPs elicited by auditory and visual stimuli sum up linearly. As a result, in the absence of audiovisual integration, the amplitude of the N1 and P2 ERP components (that are typically present during this early time window) elicited by audiovisual stimuli (AV condition) is identical to the algebraic sum of the same components elicited by auditory only (A) and visual only (V) stimuli (the A + V condition). Audiovisual integration, on the other hand, leads to the attenuation of the N1 and P2 amplitude and latency in the AV compared to the A + V condition, as long as visual movement precedes the auditory signal (Besle, Fort, & Giard, 2004; Giard & Besle, 2010; Stekelenburg & Vroomen, 2007; Van Wassenhove, Grant, & Poeppel, 2005).

Recent research has demonstrated that within the context of this paradigm, changes in the amplitude and latency of the N1 and P2 components to audiovisual stimuli may occur independently of each other and index different aspects of audiovisual processing. Attenuation of the N1 amplitude is thought to depend on how well visual movements can cue the temporal onset of the auditory signal, with the nature of the audiovisual stimuli – speech or non-speech – being irrelevant. This interpretation agrees with findings showing that N1 attenuation is not sensitive to the degree to which lip movements predict the identity of the phoneme (Van Wassenhove et al., 2005). In fact, N1 reduction was even reported for incongruent audiovisual presentations (Stekelenburg & Vroomen, 2007). However, it is typically absent when visual cues do not precede the auditory signal (e.g., Brandwein et al., 2011) or carry little information about the temporal onset of the latter (e.g., Baart et al., 2014). On the other hand, shortening of the N1 latency is greatest when lip movements are highly predictive of the articulated phoneme (such as in the articulation of bilabial sounds, for example) (Van Wassenhove et al., 2005).

In a recent study, Baart et al. (2014) proposed that changes in the P2 component elicited by AV speech reflect audiovisual phonetic binding (and not just a detection of audiovisual correspondences associated with the N1 component). The authors used sine-wave speech as stimuli, with some study participants perceiving them as speech and some as computer noises. The authors have demonstrated that while changes to N1 in the audiovisual condition were present when participants perceived sine-wave as both speech and non-speech, changes to the P2 component were present only in those research participants who perceived their stimuli as speech. While P2 attenuation is not speech-specific (Vroomen & Stekelenburg, 2010), it appears to reflect the binding of auditory and visual modalities that are perceived as representing a unitary audiovisual event.

Because almost all studies employing the above electrophysiological paradigm have been conducted with adults, very little is known about when in development changes in the N1 and P2 components to audiovisual stimuli can be reliably detected and, consequently, when different aspects of audiovisual processing indexed by such changes reach adult-like levels. A study by Brandwein et al. (2011) used simple non-speech stimuli – a pure tone and a red disk – and tested children ranging in age from 7 to 16 as well as adults. They found that the amplitude of the N1 component was larger in the AV compared to the A + V condition in two oldest groups – 13- to 16-year-olds and adults. The direction of the reported effect was, however, opposite to what had previously been reported for adults (namely, the N1 and P2 amplitude elicited by the AV stimuli is typically smaller, rather than larger, compared to that elicited by the A + V stimuli). One reason for this outcome may be that the onset of visual stimuli in this study was temporally aligned with the onset of sounds, instead of preceding it (Stekelenburg & Vroomen, 2007). As a consequence, although generally speaking Brandwein and colleagues' findings are in agreement with earlier reports on the protracted developmental course of audiovisual integration, they may not generalize to more ecologically valid