Cognition 147 (2016) 100–105

Contents lists available at ScienceDirect

Cognition

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

Short Communication

Language familiarity modulates relative attention to the eyes and mouth of a talker

Elan Barenholtz [a,*], Lauren Mavica [b], David J. Lewkowicz [c]

[a] Department of Psychology/Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL, United States
[b] Department of Psychology, Florida Atlantic University, Boca Raton, FL, United States
[c] Department of Communication Sciences and Disorders, Northeastern University, Boston, MA, United States

ARTICLE INFO

Article history:
Received 6 March 2015
Revised 8 October 2015
Accepted 23 November 2015
Available online 30 November 2015

Keywords:
Multisensory
Speech
Language
Perception
Attention

ABSTRACT

We investigated whether the audiovisual speech cues available in a talker's mouth elicit greater attention when adults have to process speech in an unfamiliar language vs. a familiar language. Participants performed a speech-encoding task while watching and listening to videos of a talker in a familiar language (English) or an unfamiliar language (Spanish or Icelandic). Attention to the mouth increased in monolingual subjects in response to an unfamiliar language condition but did not in bilingual subjects when the task required speech processing. In the absence of an explicit speech-processing task, subjects attended equally to the eyes and mouth in response to both familiar and unfamiliar languages. Overall, these results demonstrate that language familiarity modulates selective attention to the redundant audiovisual speech cues in a talker's mouth in adults. When our findings are considered together with similar findings from infants, they suggest that this attentional strategy emerges very early in life.

1. Introduction

Speech processing depends on the rapid encoding and interpretation of a complex audiovisual signal. Fortunately, natural languages contain a high degree of structure at the phonetic, lexical, syntactic, and semantic levels and prior knowledge of these structures can facilitate processing. For example, under noisy conditions, perception of auditory speech is more accurate when the spoken language is familiar (Cutler, Weber, Smits, & Cooper, 2004; Gat & Keith, 1978; Lecumberri & Cooke, 2006; Mayo, Florentine, & Buus, 1997; Van Wijngaarden, Steeneken, & Houtgast, 2002). This suggests that language familiarity can reduce the amount of bottom-up information needed to successfully process auditory speech. Here, we asked whether language familiarity also affects responsiveness to audiovisual speech.

Typically, linguistic communication is multisensory in nature. People can both hear and see their interlocutor produce visual and auditory speech signals and they automatically integrate them (McGurk & MacDonald, 1976). Such integration produces a perceptually more salient signal (Meredith & Stein, 1986; Partan & Marler, 1999; Rowe, 1999). Indeed, studies show that concurrent access to redundant audible and visible speech cues enhances speech perception under noisy conditions (Middelweerd & Plomp, 1987; Rosenblum, 2008; Rosenblum, Johnson, & Saldana, 1996; Sumby & Pollack, 1954; Summerfield, 1979). Several recent studies have found that familiarity with a language modulates the perceived timecourse of audiovisual speech: when a language is familiar, the visual speech signal must lead the auditory speech signal by a larger time interval for simultaneity to be perceived compared with when the language is unfamiliar (Love, Pollick, & Petrini, 2012; Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010), perhaps because familiarity speeds up the auditory processing of speech.

Language familiarity may also modulate visual selective attention during speech encoding, a possibility supported by evidence from infant studies. Lewkowicz and Hansen-Tift (2012) presented monolingual, English-learning infants of different ages with videos of talkers speaking either in their native language or in a non-native language (i.e., Spanish). At 4 months, infants fixated the talker's eyes, whereas at 8 and 10 months of age—when infants enter the canonical babbling stage and begin to acquire spoken language—they fixated the talker's mouth. At 12 months of age, the infants no longer fixated the mouth more than the eyes when the talker spoke in the infants' native language but continued to fixate the mouth more when the talker spoke in a non-native language.

Lewkowicz and Hansen-Tift's (2012) findings indicated for the first time that selective attention to the audiovisual redundancy

* Corresponding author at: Department of Psychology, Florida Atlantic University, Boca Raton, FL 34331, United States.
E-mail address: elan.barenholtz@fau.edu (E. Barenholtz).

http://dx.doi.org/10.1016/j.cognition.2015.11.013
0010-0277/© 2015 Elsevier B.V. All rights reserved.

available in the mouth is modulated by language familiarity. If this early lip-reading behavior reflects a general encoding strategy in response to language familiarity, these differences in fixation behavior may persist into adult. Of course, as Lewkowicz and Hansen-Tift (2012) noted, lipreading in infancy may reflect acquisition of speech production capacity. If so, the selective deployment of attention to a talker's mouth in infancy may reflect infants' attempt to imitate and produce human speech sounds and, thus, may not generalize to adults. Indeed, Lewkowicz and Hansen-Tift (2012) found in a separate experiment with monolingual English-speaking adults that they looked longer at the eyes of a talker regardless of whether she spoke in their native language or not.

Crucially, the adults in the Lewkowicz and Hansen-Tift (2012) study were only asked to passively watch and listen to the talker. Studies with adults have found, however, that the distribution of attention to the eyes and mouth is modulated by task. For example, findings show that the mouth attracts more attention when speech cues become relevant (Buchan, Paré, & Munhall, 2007; Driver & Baylis, 1993; Lansing & McConkie, 1999, 2003) and especially when the auditory signal is degraded (Driver & Baylis, 1993; Lansing & McConkie, 2003; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Conversely, when the task is to attend to social-reference, emotional, and deictic cues, the eyes attract more attention (Birmingham, Bischof, & Kingstone, 2008; Emery, 2000).

Given these findings, we asked whether speech in an unfamiliar language might cause adults to attend more to a talker's mouth if their explicit task is to process the speech. To test this possibility, we tracked selective attention in adults while they watched and listened to people speaking either in their native and, thus, familiar (English) language or in an unfamiliar (Icelandic or Spanish) language. The participants were explicitly required to encode the speech stimulus by subsequently being asked to perform a simple match-to-sample task. We expected that the participants would attend more to the mouth in the unfamiliar than in the familiar language condition.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Participants were 60 self-described English-speaking monolingual, Florida Atlantic University undergraduate students, participating for course credit. Separate groups of 30 participants, each, were randomly assigned to one of two Language groups (English/Icelandic or English/Spanish). Each group of 30 was further subdivided into two groups of 15 with the order of language presentation (i.e. familiar or unfamiliar first) counterbalanced across participants.

2.1.2. Stimuli

Stimuli consisted of movie recordings of two female models, recorded in a sound-attenuated room and presented on an infrared-based eye tracking system (T60; Tobii Technology, Stockholm, Sweden[1]) on a 17-in. computer monitor. Both models were fully bilingual speakers of both English (with no discernible accent) and one other native language (one Spanish, one Icelandic). Each model was recorded speaking a set of 20 sentences in English and the same 20 sentences in her other, native language. The models were recorded from their shoulders up and were instructed to speak naturally in an emotionally passive tone without moving their head. The face of the models measured approximately 6° visual angle width (ear to ear) by approximately 11° visual angle length. The recorded individual sentences averaged 2.5 s each for all three recorded languages.

2.1.3. Procedure

A single trial is schematized in Fig. 1. Participants were presented with sequentially presented pairs of video segments, each consisting of the same person audibly uttering a short sentence, followed by an audio-only clip of one of the two sentences. Participants had to choose which of the two previously presented audiovisual movie segments corresponded to the audio-only clip. For half the participants, the video sequences consisted of a bilingual female speaking English (familiar) sentences in one block and the same model speaking Icelandic (unfamiliar) sentences in a different block (English/Icelandic group). For the other half of the participants, the sequences consisted of a different model speaking English sentences in one block and the same model speaking Spanish sentences in a different block (English/Spanish group). Participants indicated whether the auditory-only clip was extracted from the first or second movie by pressing a key on the keyboard.

Each participant completed two experimental blocks, each consisting of ten pairs of sentences. In one block, all of the sentences were in English while in the other block they were all in an unfamiliar language, either Icelandic or Spanish. Each group was only presented with one model, speaking both English and Icelandic (Icelandic Group) or English and Spanish (Spanish Group). This ensured that the same visual features were present across the familiar and unfamiliar blocks for each participant. Block order (i.e. familiar or unfamiliar presented first or second) was counterbalanced across participants.

Participants' eye movements were recorded with an eye tracking system (T60; Tobii Technology, Stockholm, Sweden) and analyzed with the Tobii Studio 3.0.6 software. Gaze was monitored using near infrared and both bright and dark pupil-centered corneal reflection. Stimuli were presented on a 17-in. flat panel monitor with a screen resolution of 1280×1024 pixels and a sampling rate of 60 Hz. All participants were tested in a quiet room that was illuminated by the stimulus display and were seated ~60 cm from the screen. A standardized five-point calibration was performed prior to tracking as implemented in Tobii Studio software.

2.1.4. Fixation analyses

We defined three principal areas of interest (AOIs): the mouth, the eyes, and the whole face. For each condition, we calculated the time spent fixating the eye and mouth AOIs as a percentage of the total time spent fixating anywhere within the face AOI (Note that fixations within either the mouth or eyes AOI were counted toward the total fixation duration to the face). Fixation (as contrasted with saccades or other eye movements) durations were determined using Tobii Studio's fixation filter algorithm,[2] which distinguishes between time spent fixating within an AOI (which were the basis of our analyses) and time spent engaging in a saccade (which were not included in the analyses).

2.2. Results

Performance in the matching task was near ceiling (between 95% and 97%) across all conditions and language groups, with no significant difference between familiar and unfamiliar (all p-values >.1 by t-test). Fig. 2A shows the proportion of time spent

[1] Technical specifications are available at: http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_T60_T120_EyeTracker_UserManual.pdf.

[2] http://www.tobii.com/eye-tracking-research/global/library/white-papers/the-tobii-i-vt-fixation-filter/.