Cognition 160 (2017) 103–109

Contents lists available at ScienceDirect

Cognition

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

Read my lips: Visual speech influences word processing in infants

Drew Weatherhead *, Katherine S. White

Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada

ARTICLE INFO

Article history:
Received 10 December 2015
Revised 29 December 2016
Accepted 4 January 2017
Available online 11 January 2017

Keywords:
Audiovisual speech perception
Infant development
Word recognition

ABSTRACT

What do infants hear when they read lips? In the present study, twelve-to-thirteen-month-old infants viewed a talking face produce familiar and unfamiliar words. The familiar words were of three types: in Experiment 1, they were produced correctly (e.g., "bottle"); in Experiment 2, infants saw and heard mispronunciations in which the altered phoneme either visually resembled the original phoneme (*visually consistent*, e.g. "pottle"), or did not visually resemble the original phoneme (*visually inconsistent*, e.g., "dottle"). Infants in the correct and consistent conditions differentiated the familiar and unfamiliar words, but infants in the inconsistent condition did not. Experiment 3 confirms that infants were sensitive to the mispronunciations in the consistent condition with auditory-only words. Thus, although infants recognized the consistent mispronunciations when they saw a face articulating the words, they did not with the auditory information alone. These results provide the first evidence that visual articulatory information affects word processing in infants.

1. Introduction

Word recognition is surprisingly robust, despite the fact that listeners have to contend with a noisy, and sometimes degraded, signal. One source of information that contributes to the robustness of this process in adults is visual articulatory information (Sumby & Pollack, 1954). The observation of mouth movements during speech provides information about temporal and phonetic properties of the acoustic signal, which can be used by listeners to decode the speech signal more reliably (Yehia, Rubin, & Vatikiotis-Bateson, 1998; Grant & Greenberg, 2001; Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). In fact, the influence of visual information is so strong that viewing articulatory gestures that are incongruent with the acoustic signal can alter the auditory percept, even if the acoustic signal is clear (McGurk & Macdonald, 1976).

How and when does this influence develop? Whereas adults have considerable experience watching others' articulations and producing their own, young infants do not. Nonetheless, even young infants are sensitive to information from the mouth. Two-month-old infants look at the video of a talking face that corresponds to a heard vowel (Kuhl & Meltzoff, 1982; Patterson & Werker, 2003). Four-month-old infants detect audiovisual asynchrony during speech perception (Lewkowicz, 2010). Infants are so sensitive to mouth movements that they can discriminate languages simply by watching silent videos of a talking face (Weikum et al., 2007). And, like adults, infants are susceptible to the McGurk Effect (Burnham & Dodd, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Visible speech articulation has even been shown to influence infants' learning of phonetic categories (Teinonen, Aslin, Alku, & Csibra, 2008).

However, adults and younger listeners may differ in how visual speech information is used. In adults, visual information affects interpretation of more than just speech sounds – it also affects lexical access. For example, when auditory and visual signals conflict, participants' decisions about the identity of an initial consonant in a stimulus are biased in the direction of the modality consistent with a real word (e.g., auditory "besk"/visual "desk" produces more /d/ responses, while auditory "beg"/visual "deg" produces more /b/ responses; Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; Brancazio, 2004; see also Ostrand, Blumstein, Ferreira, and Morgan (2016) for evidence of visual influences on the processing of auditory non-words in a different task). Thus, analogous to the effects of lexical status on phonetic perception in the auditory domain (Connine & Clifton, 1987; Ganong, 1980; Pitt, 1995), visual lexical status influences phonetic perception. Therefore, in adults, lexical knowledge affects how auditory and visual input is combined.

In young children, the evidence suggests that lexical knowledge does not influence audio-visual integration. When 5-to-10-year-old children had to detect consonant targets within words and pseudowords presented in noise, children were better able to identify the target consonants when stimuli were presented audio-visually than auditorily (Fort, Spinelli, Savariaux, & Kandel, 2012). However, unlike adults (Fort, Spinelli, Savariaux, & Kandel, 2010), children

* Corresponding author at: Department of Psychology, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada

E-mail address: deweathe@uwaterloo.ca (D. Weatherhead).

http://dx.doi.org/10.1016/j.cognition.2017.01.002
0010-0277/

did not identify target consonants more successfully for audio-visual words than audio-visual non-words. This suggests that for young children, visual speech contributes primarily to phonemic, but not lexical, interpretation. Another possibility, though, is that visual speech can impact lexical processing even in very young children, but the impact is overshadowed when attention is focused on phoneme identification. To address this possibility, we took a very different approach, testing infants' recognition of mispronounced familiar words using a word preference procedure.

Previous studies with auditory-only stimuli demonstrate that 11-to-15-month-olds prefer familiar words (words known prior to arrival in the laboratory) over unfamiliar or nonsense words. However, they do not show a preference if the familiar words are accented or mispronounced, by even a single-feature, at least in stressed syllables (Best, Tyler, Gooding, Orlando, & Quann, 2009; Hallé & de Boysson-Bardies, 1996; Swingley, 2005). In the current study, 12-to-13-month-old infants viewed a talking face producing familiar and unfamiliar words. In Experiment 1, the words were pronounced correctly (e.g., "bottle"), to ensure that infants distinguish between familiar and unfamiliar words with audiovisual stimuli. In Experiment 2, the familiar words were mispronounced by either a voicing or place change in onset position (between subjects). Importantly, for the voicing mispronunciations, the altered phonemes were visually indistinguishable from the original phonemes (they were visually *consistent* with the correct pronunciation, e.g., "pottle"). In contrast, for place mispronunciations, the altered phonemes did not visually resemble the original phonemes (they were visually *inconsistent* with the correct pronunciation, e.g., "dottle"). In both mispronunciation conditions, the auditory and visual information matched. Finally, in Experiment 3, infants heard auditory-only versions of the consistent stimuli from Experiment 2. If visual speech impacts infants' ability to recognize wordforms, infants should recognize mispronounced words only when they are presented audiovisually and are visually consistent with the correct pronunciation.

2. Experiment 1

We first compared infants' preference for familiar vs. unfamiliar wordforms, to ensure that infants recognize familiar words when they are presented audiovisually.

2.1. Participants

Eighteen 12-to-13-month-olds (9 females, mean age = 12 months 16 days) participated. An additional five infants were tested but not included due to fussiness (3), software error (1), or an imbalance in the number of familiar word and unfamiliar word trials in each block (1). All participants were full-term monolingual English learners (not more than 3 weeks premature), and had no known hearing or vision problems.

2.2. Audio stimuli

Sixteen highly familiar words were chosen using the MacArther-Bates Communicative Development Inventories (Dale & Fenson, 1996; see Appendix A).[1] Sixteen unfamiliar words were created, matched in initial consonants and approximate lengths to the familiar words. The unfamiliar words consisted of primarily non-words, and a few very low-frequency words that infants this age do not know. A female native English speaker produced all thirty-two stimuli. Stimuli were recorded in a sound-treated booth at a sampling rate of 44,100 Hz and were later equated for amplitude in Praat (Boersma & Weenink, 2009). The audio stimuli were inserted into the videos described below.

2.3. Audiovisual stimuli

The speaker who produced the audio stimuli, a Caucasian 23-year-old female, was recorded against a plain, light-blue backdrop. Thirty-two videos were recorded, one for each of the 32 stimuli. The videos showed the speaker from the shoulder up, with her lips at the center of the video. The audio from the videos was replaced with the audio stimuli described above using Apple iMovie. To facilitate matching the speech rate of the video, the speaker viewed each video before recording the corresponding auditory stimulus.

The videos of the 16 familiar words were concatenated (with 600 ms separating each word) to create twelve pseudo-randomized sequences of 12 words each (each sequence approximately 24 s). To standardize the transitions between the words, the final frame of each individual video was frozen until the next word began. The twelve sequences were pseudo-randomized such that each of the sixteen words appeared equally often, and toward the beginning and end of the sequences equally often. Each infant saw four randomly chosen sequences. The same pseudo-randomized concatenation process was followed for the 16 unfamiliar word videos. Again, each infant was exposed to four of the 12 possible unfamiliar sequences.

2.4. Procedure

The participant sat on a parent's lap approximately 1.5 ft. from a 36 × 21-in. plasma screen television in a sound-treated testing room. Each participant saw eight unique test sequences (presented at 65–70 db): four familiar word sequences and four unfamiliar word sequences. Presentation of the video was contingent on the infant's looking behavior. Each sequence was presented as long as the infant fixated on the screen, up to a maximum of 24 s. The video stopped when the infant looked away, and the sequence ended when the infant looked away for 2 s. If the infant's looking time was less than 2 s, the sequence was repeated. A video of a baby laughing served as an attention getter between sequences.

Sequence order was pseudo-randomized with constraints: for half of the infants, the session began with a familiar word sequence; for the other half it began with an unfamiliar word sequence. Likewise, for half of the infants, the final sequence was a familiar word sequence; for the other half, it was an unfamiliar word sequence. The first four sequences were made up of two familiar word sequences and two unfamiliar word sequences, as were the last four sequences. The order of the sequences within each 4-sequence block was pseudo-randomized such that all possible sequence orders occurred. No more than two of each sequence type were played consecutively.

2.5. Results

A paired-sample t-test comparing average looking time for the two word types (Familiar and Unfamiliar) revealed no significant difference $t(17) = -0.51$, $p = 0.62$ (with 12 out of 18 participants showing a preference for the unfamiliar words). However, as this is the first word preference study using audiovisual stimuli, the optimal number of trials could not be predicted in advance. We therefore explored the possibility that infants looked differentially for the two types of words early in the experiment, but allocated

[1] We additionally asked parents in our experiments to report on their infants' familiarity with these words, using a scale of 1–4 (1 = child does not know word, 4 = child knows word very well). The average score for all 16 words across experiments was 3.04. There were no differences in parental reports across conditions and experiments (Wald X (df = 3, N = 64) = 1.68, $p = 0.641$). These reports confirm that infants in all of the experiments were familiar with the words prior to the testing session.