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

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 word-forms, 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.

their attention differently as time went on. This possibility is likely given previous studies of infant multi-sensory matching and multi-sensory sound discrimination, which have demonstrated declines in responsiveness across test trials (e.g., Bahrick, Hernandez-Reif, & Flom, 2005; Bruderer, Danielson, Kandhadai, & Werker, 2015; Lewkowicz, Minar, Tift, & Brandon, 2015). Indeed, a repeated-measures ANOVA with within-subject factors block (first four sequences vs. last four sequences) and word type (Familiar vs. Unfamiliar), revealed a main effect of block, $F(1, 17) = 26.77$, $p < 0.001$, and a word type x block interaction, $F(1, 17) = 6.34$, $p = 0.022$. Mean looking dropped significantly between blocks, from 17.6 s to 12.5 s. Because of the significant interaction we conducted an analysis of the two blocks separately. In the first block, there was a significant difference in looking times, $t(17) = -2.23$, $p = 0.039$, $d = 0.578$, with infants showing a preference for the unfamiliar words (with 15 out of 18 participants showing a preference for the unfamiliar words). A difference in looking time, regardless of the direction, demonstrates that infants have discriminated the two sets of words (and, therefore, recognized the familiar wordforms).[2] In the second block, in contrast, there was no difference in looking for the two word types, $t(17) = 0.832$, $p = 0.417$. Thus, infants differentiated the two word types in the first half of the test, but over time redirected their attention (Fig. 1).

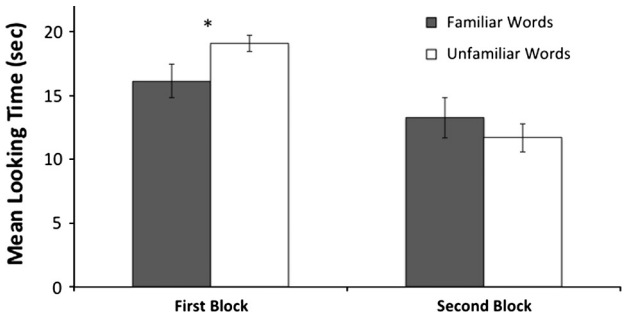

Fig. 1. Infants' mean looking time (in seconds) by word type, for the first and second block of Experiment 1. Error bars represent standard error.

3. Experiment 2

In Experiment 2, we investigated infants' processing of mispronounced familiar words. If visual speech impacts infants' word-form recognition, infants should recognize mispronounced words when the visual information is consistent, but not when it is inconsistent, with the correct pronunciation. Furthermore, if infants recognize the consistent words, the looking pattern in the visually consistent condition should resemble that of Experiment 1.

3.1. Participants

Thirty-six 12-to-13-month-olds (19 females, mean = 12 months 20 days) participated. An additional ten infants were tested but not included due to fussiness (1), parental interference (1), experimenter error (1), or an imbalance in the number of familiar word and unfamiliar word trials in each block (6). All participants were full-term monolingual English learners (not more than 3 weeks premature), and had no known hearing or vision problems.

3.2. Audio stimuli

The same sixteen words and sixteen unfamiliar words were used as in Experiment 1, but the first consonant of each stimulus was changed. For the visually consistent words, the place of articulation of the onset consonant remained constant but the voicing changed. For example, /b/ became /p/ (see Appendix A). For the visually inconsistent words, the voicing of the initial consonant remained constant, but the place of articulation changed. For example, /b/ became /d/. Importantly, the mispronunciations in both conditions were of the same magnitude (involved a single feature), and both voicing changes and place changes disrupt toddlers' word recognition (e.g., White & Morgan, 2008). The onset consonants in the unfamiliar words were altered in a corresponding manner to ensure that the sets of onset consonants were balanced across familiar and unfamiliar words. The same native English speaker from Experiment 1 produced the stimuli.

3.3. Audiovisual stimuli

The same procedure was used to create the test videos as in Experiment 1. Note that there is a critical audiovisual difference for the visually consistent and visually inconsistent mispronunciations. For the visually consistent words, the place of articulation did not change, and thus the articulatory gestures looked the same as the gestures in Experiment 1. For the visually inconsistent words, the place of articulation did change, and thus the articulatory gestures looked different from the gestures in the original items.

It is important to note that consistency here is defined with respect to the real words (e.g., bottle). In our stimuli, there was no mismatch between the visual and auditory information, meaning there was no ambiguity as to the identity of the consonant in either condition.

3.4. Procedure

Infants were randomly assigned to either the visually consistent condition or the visually inconsistent condition. The procedure was the same as Experiment 1. Participants were again presented with four familiar word and four unfamiliar word test sequences.

3.5. Results

To maintain consistency with Experiment 1, we report the results of the first four sequences. A repeated measures ANOVA with the within-subject factor of word type (Familiar vs. Unfamiliar) and between-subject factor of condition (Visually Consistent vs. Visually Inconsistent) revealed a condition × word type interaction, $F(1, 34) = 6.17$, $p = 0.018$, $d = 0.825$. No main effect of word type was found, $F(1, 34) = 0.001$, $p = 0.973$.[3]

To further explore the effect of word type within each condition, paired-sample t-tests were conducted. In the visually consistent condition, there was a difference in looking time between the two word types, $t(17) = -2.39$, $p = 0.029$, $d = 0.569$ (with 14 out of 18 participants showing a preference for the unfamiliar words; Fig. 2). In the visually inconsistent condition, there was no significant difference, $t(17) = 1.47$, $p = 0.160$ (with 10 out of 18 participants showing a preference for the unfamiliar words). Therefore, infants discriminated the mispronounced familiar words and the unfamiliar words only when the mispronunciations were visually consistent with the original pronunciations.

[2] The direction of preference could not be predicted, as this is the first study to test infants' word recognition with audio-visual stimuli. We return to this issue in the discussion.

[3] When all eight sequences were analyzed, there was a marginal main effect of word type, $F(1, 34) = 3.88$, $p = 0.057$ and no condition x word type interaction $F(1, 34) = 2.78$, $p = 0.105$. Paired sample t-tests revealed a significant difference between the two word types in the visually consistent condition, $t(17) = -3.04$, $p = 0.007$ (with 14 out of 18 participants showing a preference for the unfamiliar words), but not in the visually inconsistent condition, $t(17) = -0.19$, $p = 0.853$ (with 9 out of 18 participants showing a preference for the unfamiliar words).

To confirm that the pattern of results for the visually consistent condition resembled the pattern of results for the correctly pronounced words in Experiment 1, a repeated measures ANOVA with the within-subject factor of word type (Familiar and Unfamiliar) and between-subject factor of condition (Correct and Visually Consistent) was run. As expected, there was a significant main effect of word type, $F(1, 34) = 9.42$, $p = 0.004$, $d = 1.053$ and no condition x word type interaction, $F(1, 34) = 0.838$, $p = 0.366$. In contrast, a comparison between the visually inconsistent condition and Experiment 1 yielded no main effect of word type, $F(1, 34) = 0.57$, $p = 0.454$, and a significant condition x word type interaction, $F(1, 34) = 7.04$, $p = 0.012$, $d = 0.91$. Therefore, infants treated the visually consistent mispronunciations the same as the correct pronunciations, but treated the visually inconsistent mispronunciations differently.

4. Experiment 3

One possible concern in Experiment 2 is that visually consistent mispronunciations were treated like correct pronunciations, not because of support from the visual articulation, but because voicing changes are less salient than place changes. Although previous research suggests that this is not true for adults or infants (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Miller & Nicely, 1955; White & Morgan, 2008), a final experiment was run using auditory-only stimuli. In this experiment, infants heard the audio corresponding to the visually consistent condition. If infants' preference in the consistent condition of Experiment 2 was because voicing mispronunciations are not salient, then they should continue to distinguish the mispronounced and unfamiliar words. However, if infants' preference was due to the support of the articulatory gestures, then there should be no recognition of the mispronounced words.

4.1. Participants

Eighteen 12-to-13-month-olds (7 females, mean = 12 months 14 days) participated. An additional five infants were tested but not included due to fussiness (2), software error (1), or to an imbalance in the number of familiar word and unfamiliar word trials in each block (2). All participants were full-term monolingual learners of English (not more than 3 weeks premature), and had no known hearing or vision problems.

4.2. Audio stimuli

Audio stimuli were from the visually consistent condition of Experiment 2 – voicing mispronunciations of familiar words and the corresponding unfamiliar words.

4.3. Procedure

The procedure was identical to that of Experiments 1 and 2, with the exception that instead of viewing talking faces, infants saw a checkerboard on the screen, as is standard in many auditory-only speech perception studies with infants.

4.4. Results

For the first four sequences, a paired-sample *t*-test revealed that there was no difference in looking time across the two word types, $t(17) = -0.10$, $p = 0.922$ (with 7 out of 18 participants showing a preference for the unfamiliar words; see Fig. 2).[4] Infants did not recognize the consistent words when they were presented in an auditory-only context. This result is consistent with previous demonstrations of infants' failure to recognize mispronounced familiar wordforms in auditory-only procedures (e.g., Swingley, 2005). Thus, the difference in looking observed in the visually consistent condition in Experiment 2 was driven by the presence of the speaker's articulatory movements.

5. General discussion

The current study explored whether visual speech information impacts infants' wordform recognition. To test this, we used a word preference procedure with a talking face. For correctly pronounced words (Experiment 1), 12-to-13-month-olds showed a preference for unfamiliar words over familiar words. When the familiar words were mispronounced (Experiment 2), infants showed the same pattern of looking, but critically, only for mispronunciations that were visually consistent with the correct pronunciation. Finally, when the consistent words were presented in an auditory-only context (Experiment 3), infants did not distinguish them from unfamiliar words. Together, these experiments provide strong evidence that visual speech affects infants' recognition of familiar wordforms. Below, we first discuss the pattern of looking and then turn to potential implications for lexical processing.

Previous studies using auditory-only word preference procedures have found that 11- and 15-month-olds listen longer to familiar than unfamiliar words (Best et al., 2009; Hallé & de Boysson-Bardies, 1996; Swingley, 2005). We found the opposite: infants preferred to look longer for the unfamiliar words. What might explain this difference? Previous studies have suggested that infants pay more attention to the mouth at points in development when they are learning the sounds of their native language (Lewkowicz & Hansen-Tift, 2012; Pons, Bosch, & Lewkowicz, 2015). It may be the case that infants also attend more to the mouth when they encounter unfamiliar words, as the mouth movements reinforce the acoustic signal, providing a more robust learning situation. If infants attend longer to the mouth when they are trying to learn words, this would account for why infants in Experiment 1 and the visually consistent condition in Experiment 2 attended longer to the unfamiliar words – they recognized the familiar wordforms and could devote less attention to the visual information for those words. At this point, such a proposal is only speculative. Future work using eye-tracking technology to determine the precise nature of infants' visual attention will be needed to evaluate this possibility.

Many models of spoken-word recognition, such as Cohort (Marslen-Wilson & Welsh, 1978), TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994), and MERGE (Norris, Mcqueen, & Cutler, 2000), focus exclusively on auditory input, without considering a role for visual speech information. Our results have important implications for understanding the architecture of this system, as they suggest that visual speech information may play a role in "spoken" word recognition, even in infants.

There are at least three possible ways in which visual speech information could be incorporated into such models to account for our results. First, seeing the initial articulatory gesture of a word may lead to activation of the corresponding phonemes. For example, in the consistent condition, infants see an articulation consistent with /p/ and /b/, activating the representations of both phonemes. The activation of /p/ would be stronger, as it is present in the audio as well, but /b/ would nonetheless be active. Therefore, when infants see and hear the remainder of word, "-ottle", they recognize the word as *bottle*.[5] Under this interpretation, audiovisual

[4] When all eight sequences were analyzed there was also no difference in looking time across the two word types $t(17) = -0.12$, $p = 0.902$ (with 6 out of 18 participants showing a preference for the unfamiliar words).

[5] This recognition process can also be thought of more probabilistically – the evidence supports "pottle" with higher probability than "bottle", but bottle is still compatible, with some probability, with the input, and this activation is sufficient to drive attention.

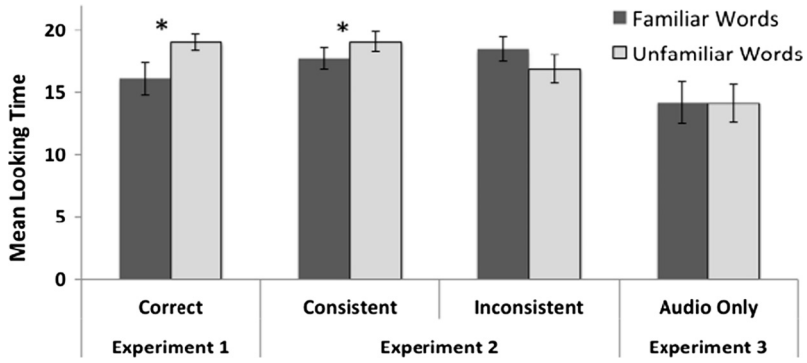

Fig. 2. Infants' mean looking time (in seconds) by word type, for the first block, in each Experiment. Error bars represent standard error.

integration would occur pre-lexically. Auditory and visual information would both impact phoneme identification and the combined phoneme information would be sent to the lexicon. This account is consistent with work suggesting that audiovisual integration during the McGurk phenomenon occurs prior to the perceptual selection of phonetic categories (e.g. Bernstein, 1989; Fowler, Brown, & Mann, 2000; Green, 1998; Massaro & Palmer, 1998).

Alternatively, the initial articulatory gestures may directly activate a pool of lexical candidates consistent with the visible articulations (Fort et al., 2013). For example, the onset of "pottle" may activate words with bilabial onsets (e.g., *baby*, *papa*, *bottle*, *puppy),* allowing infants to recognize the word *bottle* on the basis of the subsequent input (or at least for it to be active enough to drive attention). A final alternative is that the visual input is used post-lexically. For example, although both "pottle" and "dottle" would activate the lexical representation of *bottle,* because of the rhyme overlap in the auditory signal (e.g., Connine, Blasko, & Titone, 1993), only in the former case is the visual input compatible with *bottle*. This congruence between the activated word and visual articulation may lead to recognition.

Distinguishing between these alternatives is beyond the scope of the current study. Nevertheless, the present results add to the growing body of research demonstrating that viewing a speaker's articulatory movements significantly impacts word recognition. And, importantly, the current study is the first to show that visual speech impacts *infants'* word processing. Future studies should explore whether infants' visually induced recognition of these wordforms leads to semantic activation, as in adults (e.g., Ostrand et al., 2016). If so, this would suggest that the lexical processing architecture, however it may be structured, is relatively stable across development.

Demonstrating that visual articulatory information contributes to infants' wordform recognition has implications for (1) our understanding of the early lexical processing system, and (2) how infants recognize wordforms in less than ideal acoustic circumstances such as mispronounced speech, degraded speech, or a noisy environment. We suggest that infants combine visual and auditory speech during wordform recognition not only because mouth movements reinforce the acoustic signal, but also because mouth movements directly influence infants' word processing. Thus, the gateway to the infant lexicon may be through both their ears and their eyes.

Acknowledgements

The authors would like to thank Eiling Yee for helpful discussion, and all of the families and infants who participated. This work was funded by an operating grant from the Natural Sciences and Engineering Research Council of Canada awarded to K.S.W.

Appendix A

Familiar words			
Experiment 1	Experiment 2		Experiment3
Auditory word/visual word	Visually consistent Auditory mispro/visual word	Visually inconsistent Auditory mispro/visual non-word	Auditory mispro
baby	paby	daby	paby
ball	pall	gall	pall
bath	path	dath	path
book	pook	dook	pook
bottle	pottle	dottle	pottle
cookie	gookie	tookie	gookie
cup	gup	tup	gup
diaper	tiaper	biaper	tiaper
dog	tog	bog	tog
door	toor	boor	toor
foot	voot	soot	soot
keys	geys	teys	geys

(*continued on next page*)

Appendix A (*continued*)

Familiar words			
Experiment 1	Experiment 2		Experiment3
Auditory word/visual word	Visually consistent Auditory mispro/visual word	Visually inconsistent Auditory mispro/visual non-word	Auditory mispro
shoe	zhoe	foe	zhoe
sock	zock	hock	zock
telephone	delephone	pelephone	delephone
toy	doy	poy	doy

Unfamiliar words			
Experiment 1	Experiment 2		Experiment 3
Auditory non-word/visual non-word	Visually consistent Auditory non-word/visual non-word	Visually inconsistent Auditory non-word/visual non-word	Auditory non-word
bap	pap	dap	pap
beeg	peeg	deeg	peeg
boch	poch	doch	poch
boli	poli	doli	poli
boogle	poogle	doogle	poogle
caws	gaws	taws	gaws
copper	gopper	topper	gopper
dimper	timper	bimper	timper
dolp	tolp	bolp	tolp
doma	toma	boma	toma
dorso	torso	porso	torso
dith	gith	tith	gith
shomber	zhomber	homber	zhomber
sug	zug	fug	zug
tolempill	dolempill	polempill	dolempill
vick	fick	shick	fick

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.01.002.

References

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face-voice relations. *Developmental Psychology, 41*(3), 541.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., & Crewther, D. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology, 20*(1), 1–11. http://dx.doi.org/10.1080/09541440601125623.

Bernstein, L. E. (1989). Independent or dependent feature evaluation: A question of stimulus characteristics. *Behavioral and Brain Sciences, 12*(4), 756–757.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy toddlers' perception of native-and jamaican-accented words. *Psychological Science, 20*(5), 539–542.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1. 05) [Computer program].

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 30*(3), 445–463. http://dx.doi.org/10.1037/0096-1523.30.3.445.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences, 112*(44), 13531–13536.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*(4), 204–220.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7). http://dx.doi.org/10.1371/journal.pcbi.1000436.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32*(2), 193–210.

Connine, C., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13*(2), 291–299.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*(1), 125–127.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303–306.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *International Journal of Behavioral Development, 36*(6), 457–467. http://dx.doi.org/10.1177/0165025412447752.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication, 52*, 525–532.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes, 28*(8), 1207–1223. http://dx.doi.org/10.1080/01690965.2012.701758.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance, 26*(3), 877.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110.

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing.*

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. *Hearing by Eye II* (pp. 3–26). East Sussex, UK: Psychology Press.
Hallé, P., & Boysson-Bardies, B. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development, 17*, 463–481.
Kuhl, P., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.
Lewkowicz, D., & Hansen-Tift, A. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences, 109*(5), 1431–1436. http://dx.doi.org/10.1073/pnas.1114783109.
Lewkowicz, D. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66–77. http://dx.doi.org/10.1037/a0015579.
Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of experimental child psychology, 130*, 147–162.
Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*(1), 29–63.
Massaro, D. W., & Palmer, S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle* (Vol. 1) Cambridge, MA: MIT Press.
McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.
McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748. http://dx.doi.org/10.1038/264746a0.
Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27*(2), 338–352.
Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*(3), 189–234.
Norris, D., Mcqueen, J., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences Behavioral Brain Science, 23*(3), 299–325.
Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition, 151*, 96–107. http://dx.doi.org/10.1016/j.cognition.2016.02.019.
Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.
Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21*(4), 1037.
Pons, F., Bosch, L., & Lewkowicz, D. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science, 26*(4), 490–498. http://dx.doi.org/10.1177/0956797614568320.
Rosenblum, L., Schmuckler, M., & Johnson, J. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347–357.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.
Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science, 8*(5), 432–443.
Teinonen, T., Aslin, R., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*, 850–855.
Weikum, W., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J. (2007). Visual language discrimination in infancy. *Science, 316*(5828). http://dx.doi.org/10.1126/science.1137686. 1159–1159.
White, K., & Morgan, J. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language, 59*, 114–132. http://dx.doi.org/10.1126/science.1137686.
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1), 23–43.