INFANTS' USE OF TEMPORAL AND PHONETIC INFORMATION IN THE ENCODING OF AUDIOVISUAL SPEECH

D. Kyle Danielson[*], Cassie Tam [†], Padmapriya Kandhadai[§], and Janet F. Werker [‡]
Department of Psychology, The University of British Columbia, 2136 West Mall, Vancouver BC V6T 1Z4, Canada

1 Introduction

Infants utilize both auditory and visual information to perceive the speech signal. From as early as two months after birth, infants detect correspondence in the *content* of seen and heard consonants in their native language(s) [1-2] and (until they are about nine months of age) in non-native languages as well [3]. Infants' auditory perception can also be modified by the imposition of matching or mismatching visual information from a dynamic talking face [4-5]. Infants are also sensitive to the *temporal* correspondence of heard and seen speech, detecting asynchrony greater than approximately 500 ms therein [6].

What remains unclear is how infants perceive audiovisual speech when the auditory and visual signals are incongruent. When the auditory and visual signals provide conflicting information to the infant, on which of the two signals does she rely more heavily? Second, given that infants are sensitive to the temporal correspondence of speech as well, does the temporal order of the auditory and visual components affect which of the two signals the infant uses? Relying on unfamiliar non-native speech to control for the effect of prior experience on infants' behaviour, we test these two questions in the current study.

2 Method

2.1 Sample

Sixty full-term six-month-old monolingual English-learning infants (mean age = 179 days; 30 females) were tested.

2.2 Materials

Audiovisual stimuli for this experiment were recorded from a native female speaker of Hindi (Figure 1). The speaker produced consonant-vowel (CV) syllables (duration ≈ 1 s) consisting of the Hindi voiced dental stop [d̪] and the Hindi long vowel [ɑ:] (dental syllables) and the Hindi voiced retroflex stop [ɖ] and the long vowel [ɑ:] (retroflex syllables). Incongruent syllables were constructed by splicing the auditory track from one syllable type with the visual track from a duration-matched syllable of the opposite type. This procedure was repeated, resulting in three unique, temporally synchronous tokens of each mismatching type. Temporally asynchronous tokens were constructed by offsetting the auditory and visual tracks of the mismatching syllables by 333 ms, resulting in two types of asynchronous syllable: visual-first or auditory-first.

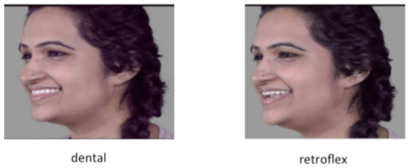

Figure 1 : Model producing dental and retroflex syllables

2.3 Procedure

Infants were seated on a caregiver's lap in a silent room facing a television screen equipped with a small camera. First, during the pre-test, infants listened to auditory-only exemplars of the Hindi dental and retroflex tokens while watching a black-and-white checkerboard. Then, infants were familiarized to 15-s sequences of Hindi audiovisual syllables in one of three conditions: synchronous, visual-first, and auditory-first. To ensure that infants only perceived audiovisual (not auditory-only) speech, stimuli were only presented when infants were looking at the screen. Each infant accumulated 120 s of looking time to the screen prior to advancing to the test phase of the experiment. Infants were tested using four auditory-only, checkerboard trials similar to those presented during the pre-test. Looking time was measured offline to determine whether infants exhibited a matching preference to the stimulus that they saw during familiarization (visual match) or to the stimulus that they heard during familiarization (auditory match).

3 Results

3.1 Main analysis

Looking time to auditory-match versus visual-match test sequences are visualized by condition as difference scores (Figure 2).

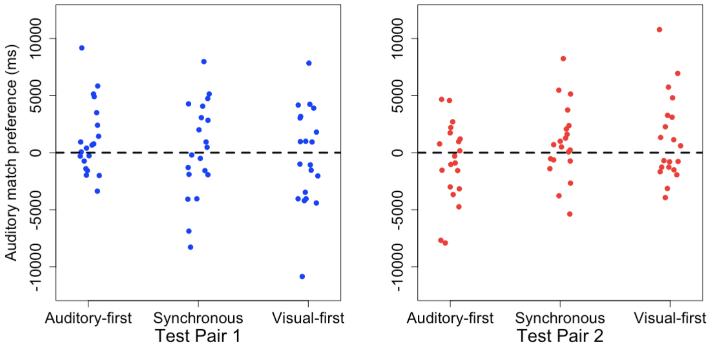

Figure 2: Difference scores (auditory match minus visual match in ms) for both pairs of test trials across three conditions (auditory-first, synchronous, visual-first). Positive scores indicate a preference for the auditory match sequences.

[*] kdanielson@psych.ubc.ca
[†] cassietam@alumni.ubc.ca
[§] priyak@psych.ubc.ca
[‡] jwerker@psych.ubc.ca

A 2 x 3 mixed-design ANOVA with match type as a within-subjects factor (auditory match, visual match) and familiarization condition (auditory-first, visual-first, synchronous) revealed no effects of match type ($F(1,57) = 1.40$, $p = .241$, $\eta^2_P = .02$) or of condition ($F(2,57) = 0.36$, $p = .703$, $\eta^2_P = .01$) on looking time. Importantly, the ANOVA also revealed no interaction between condition and match type ($F(2,57) = 0.11$, p = .893, $\eta^2_P < .01$), indicating that—regardless of familiarization condition—infants looked equally to the auditory match and visual match sequences at test.

3.2 *Post-hoc* analysis

Infants in this experiment were tested using unfamiliar Hindi dental and retroflex consonants. It is possible that infants treated these two types of consonants differently from one another, thus overshadowing the hypothesized effect. To explore this possibility, data were collapsed across all three temporally manipulated conditions, and then split based on whether the auditory information presented was dental or retroflex. These reformatted data are visualized as difference scores in Figure 3.

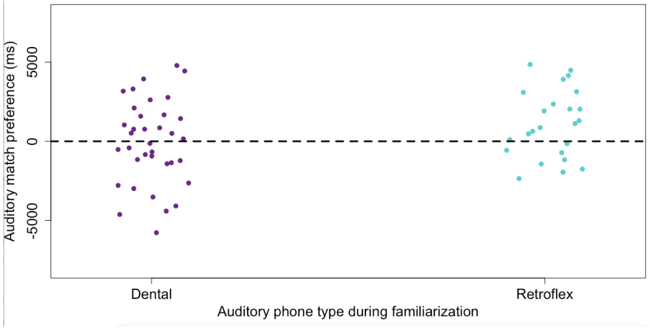

Figure 3: Difference scores (auditory match minus visual match in ms) for both pairs of test trials, split by auditory phone type during the familiarization phase.

A 2x2 mixed-design ANOVA was fitted to these looking time data, with match type (auditory match, visual match) as a within-subjects factor and auditory familiarization phone (dental, retroflex) as a between-subjects factor. A marginal interaction between match type and auditory familiarization phone emerged ($F(1,58) = 3.35$, $p = .072$, $\eta^2_P = .05$). Infants familiarized to stimuli in which the auditory component was a retroflex consonant exhibited a preference for retroflex test sequences.

4 Discussion

We familiarized infants to one of three types of audiovisual speech (visual-first, auditory-first, and synchronous), all consisting of tokens in which the visual and auditory information were incongruent. We then tested infants to determine whether they exhibited a matching preference for the syllables that they heard or that they saw. The main analysis did not reveal any interaction between familiarization condition and preference at test, indicating that infants' perception of the speech sounds was not affected by this manipulation. One possibility is that, when observing this type of audiovisually incongruent speech, infants' resulting percept is intermediate to the two syllables (as in the McGurk effect), matching neither of the two test items and resulting in no difference in preference. Another possibility is that infants perceived the incongruent visual and auditory syllables discretely, rendering both test items equally familiar. Alternatively, such a null result could be due to short exposure (120 s) or to the short temporal offset between the auditory and visual signals (333 ms), which was intentionally set below infants' hypothesized threshold for audiovisual integration [6].

However, a follow-up analysis revealed that infants familiarized to syllables in which the auditory component was a retroflex syllable exhibited a moderate preference for that syllable at test, regardless of familiarization condition. This result indicates that infants' sensitivity to the retroflex syllable may overshadow their perception of the visual dental consonant, regardless of whether the latter is presented before, simultaneously, or after the auditory retroflex consonant. Further research is required to determine why the retroflex auditory consonant, specifically, may attract more attention from English-learning infants than does the dental consonant.

5 Conclusion

The results of this study indicate that further research is necessary to determine whether infants preferentially utilize visual or auditory information to process speech sounds when the two signals provide conflicting information and when one precedes the other temporally. However, some preliminary evidence indicates that infants' speech perception may be primarily driven by auditory information from one type of unfamiliar speech sound over another.

Acknowledgments

We gratefully acknowledge the support of the National Institutes of Health (grant R21HD079260 to JFW), the Acoustical Society of America's Raymond H. Stetson Scholarship to DKD, and the Natural Sciences and Engineering Research Council of Canada's Undergraduate Student Research Award to CT.

References

[1] Kuhl, P. K., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138-1141.

[2] Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Dev Sci, 6*(2), 191-196.

[3] Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastian-Galles, N. (2009). Narrowing of intersensory speech perception in infancy. *PNAS, 106*(26), 10598-10602.

[4] Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants.

Percept Psychophys, 59(3), 347-357.

[5] Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850-855.

[6] Lewkowicz, D.J. (2010). Infant perception of audio-visual speech synchrony. *Dev Psychol*, *46*(1), 66-67.