Are you synching what I'm synching? Modeling infants' real-time detection of audiovisual contingencies between face and voice

George Hollich,[*] Eric J. Mislivec,[**] Nathan A. Helder, [**] & Christopher G. Prince[**]

[*]*Department of Psychological Sciences*
Purdue University
West Lafayette, IN 47907 USA
gholich@purdue.edu

[**]*Department of Computer Science*
University of Minnesota Duluth
Duluth, MN 55812 USA
misli001@d.umn.edu, nhelder@nerp.net,
chris@cprince.com

Abstract

Audio-visual synchrony is one of the earliest and most salient properties to which infants are sensitive (Bahrick, & Lickliter, 2000). Furthermore, it is likely that detection of contingent relations in and across modalities is a critical beginning point for autonomous mental development (Fasel, Deak, Triesch, & Movellan, 2002). While there are numerous ecological examples of the need for contingency detection, one of the strongest is in learning to connect face and voice. Dodd (1979) demonstrated that infants looked longer to a face that was synchronized to speech than one that was asynchronous with speech, and Lewkowicz (1996) has tested the limits of this detection ability in infants.

The goal of the research reported here is to explicitly model infants' real-time detection of speech/face synchrony through direct comparison between detailed empirical data from infants and a formal model of audio-visual synchrony detection. The empirical data comes from the Purdue University Infant Laboratory. Infants, ages 4, 8, and 12 months, were tested using the splitscreen preferential looking paradigm. In this procedure, two moving faces were presented side-by-side on a large video screen, with audio alternately matching one of the faces. By following the developmental trajectory of infants' preference for the synchronous face, and by examining infants' reactions when the synchrony switches between faces, we gain a better understanding of the temporal sensitivities of infants at different ages. More importantly, we gain frame-by-frame coding of infant looking preferences that are directly comparable with the output of the formal model. In the model, we use methods that directly compute audio-visual synchrony relations between low-level audio-visual features (e.g., RMS audio and grayscale pixels) based on Gaussian mutual information across a time window of audio-visual information (Hershey & Movellan, 2000).

While the ability of the model to discover and localize sources of synchrony is still in its infancy, the model already shows similar levels of performance to infants using speech-object data with uttered words and object motion, and using two speech sources and one dynamic-face. Although the model has yet to fully capture the developmental trajectories of infant's audio-visual understanding, it is our hope that the basic principles of contingency detection illustrated by the model can scale to more general models of infant attention and autonomous development. Following this motivation, we are extending our model to utilize audio-visual synchrony to train within-visual-modality categorization and to bootstrap aspects of facial recognition.