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 **
* *Purdue University* ** *University of Minnesota Duluth*

PURDUE UNIVERSITY Infant Labs

UNIVERSITY OF MINNESOTA UMD DULUTH

Abstract

What mechanisms underlie infants' abilities to detect and locate synchrony? In a preferential looking procedure 7.5-month-olds tracked one of two talking faces. Frame-by-frame coding of infant looking behavior was shown to be similar to tracking of the same stimuli by a computational model of synchrony, defined as mutual information between low-level audio and visual features. These results suggest infants track synchrony, in part, using low-level audio-visual correlations.

Introduction

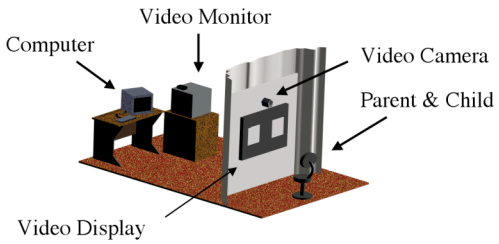
Audiovisual synchrony is one of the earliest and most salient properties to which infants are sensitive. Furthermore, the detection of contingent relations in and across modalities is likely a beginning point for autonomous mental development. While there are numerous ecological examples of the need for contingency detection, one of the strongest is connecting face and voice. Dodd (1979) demonstrated that infants would look longer to a face that is synchronized to speech than one that is asynchronous with speech (see also Pickens et al., 1994). The current studies examine whether infants succeed in this task by using low-level auditory-visual correspondences. Specifically, we compare 7.5-month-old's performance with a computational analysis of the moment-by-moment location of highest audiovisual synchrony.

Hypotheses

Infants can use the low-level synchronization between a talker's face and a target speech stream to focus on that face.

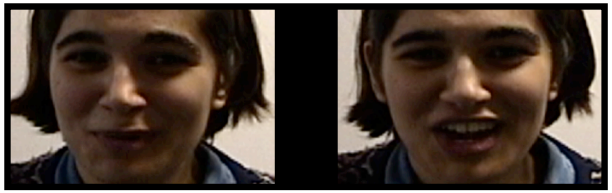
Moment-by-moment analysis of infant's performance should closely track computational estimates of low-level synchrony.

Method



The splitscreen preferential looking paradigm. Infants sit on a parent's lap 1m from a 1.5m projection display.

Procedure



"The cup was bright and shiny. A clown drank from the red cup. The other one picked up the big cup…"

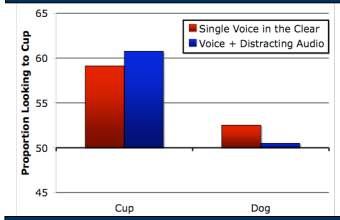
OR

"The dog ran around the yard. The mailman called to the big dog. He patted his dog on the head…"

Subjects

Subjects were 20 infants with ages ranging from 7 to 8 months (mean age = 7.54). All were from monolingual English-speaking homes with no history of hearing problems or language delay.

Results

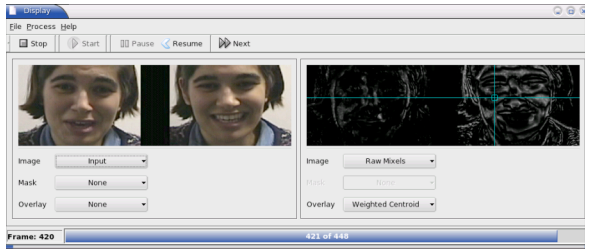


Despite a salience preference for the cup video, infants were able to track audio-visual synchrony, even in noise. They looked less at the cup video when the dog audio was played.

Model

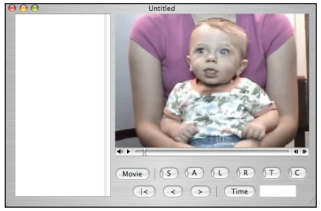
The model uses a modified version of an algorithm by Hershey & Movellan (2000; *HM* below). This algorithm defines synchrony as the Gaussian mutual information between a pair of time-based sensory streams, in our case an audio stream and a visual stream. The audio stream used a sampling rate of 44.1 kHz, and the visual frame rate was 29.97 frames per second. We used a time window of ½s (S = 15) in the mutual information computation. Audio features used were RMS (Root-mean squared) amplitude; visual features were grayscale and pixel-intensity change. The output of the HM algorithm is a *mixelgram* comprised of a matrix of *mixels* (*m*utual *i*nformation pix*els*). To some degree, the visual highlights of the mixelgrams correspond to audio-visual synchrony (Vuppla, 2004). Moment-by-moment estimates of the central location of audio-visual synchrony were computed as the mathematical centroid of the mixels in each mixelgram.

Mixelgram

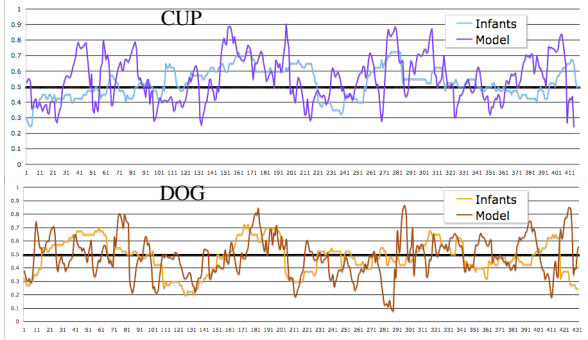


Frame-by-Frame Coding

Frame-by-frame coding of looking preference allows for moment-by-moment analysis of the proportion of infants looking toward the cup when the cup passage was read and when the dog passage was read.



Frame-by-Frame Comparison



Conclusions

- Infants (as a group) are tracking low-level synchrony (r=.30).
- We have the beginnings of a model that can quantitatively predict this preference.
- ESMA - An *E*pigenetic *S*ensory-level *M*odel of *A*ttention during early word learning and audio-visual segmentation.

Support: Purdue Research Foundation Grant to GH; donation to CGP from Digi-Key Corp., & University of Minnesota Duluth UROP grants.
Software: http://www.cprince.com/PubRes/SenseStream
Correspondence: ghollich@purdue.edu