Developmental Science

Developmental Science 12:3 (2009), pp 379–387 DOI: 10.1111/j.1467-7687.2009.00823.x

SPECIAL SECTION: COMPUTATIONAL PRINCIPLES OF LANGUAGE ACQUISITION

Comparing infants′ preference for correlated audiovisual speech with signal-level computational models

George Hollich[1] and Christopher G. Prince[2]

1. *Department of Psychological Sciences, Purdue University, West Lafayette, USA*
2. *Department of Computer Science, University of Minnesota Duluth, Duluth, USA*

Abstract

How much of infant behaviour can be accounted for by signal-level analyses of stimuli? The current paper directly compares the moment-by-moment behaviour of 8-month-old infants in an audiovisual preferential looking task with that of several computational models that use the same video stimuli as presented to the infants. One type of model utilizes only signal-level properties of visual motion whereas the other adds audiovisual integration (either through correlation or instantaneous addition of audio and visual signals). Together these models account for a significant portion of the variance in infant looking.

Introduction

From an early age, speech is processed through the eyes as well as the ears. For example, 4-month-olds prefer synchronized speech when watching a person talk (Dodd, 1979). Likewise, Pickens, Field, Nawrocki, Martinez, Soutullo and Gonzalez (1994) found that when given a choice between two videos of talkers, only one of which matches the audio source, 3-month-olds and 7-month-olds (but not 5-month-olds) will look at the face synchronized with the audio. Moreover, 7.5-month-olds appear able to use this correspondence between what they see and hear in order to focus their attention on a particular talker and hear better in noise (Hollich, Newman & Jusczyk, 2005). Such results suggest that infants are integrators of audiovisual speech from an early age.

Although such studies demonstrate that infants *are* sensitive to audiovisual information, the specifics of how they track the speaker remain a mystery. In the study by Dodd (1979) for example, trials were 30 seconds long, with differences in looking towards the matching display of less than 3 seconds. Clearly, infants did not spend all the allotted time looking at synchronized stimuli. Although such fluctuations in performance are certainly affected by high-level cognitive skills and/or individual differences in experience and speed of habituation, infant looking may also be affected by the nature of the stimuli. That is, when faced with a dynamic visual display involving multiple matches for audio, audiovisual correlations are not absolute. A visual display of a face does not provide an isomorphic match to the speech stream. As a direct result, the audiovisual correlation between sight and sound waxes and wanes. Likewise, extraneous factors, such as sudden visual movement, can temporarily make unsynchronized portions of the visual display highly salient. By knowing the specifics of the stimuli, we can be much more precise in relating the stimuli to infant behaviour and can come closer to understanding the actual mechanisms involved. One contemporary means to investigate these specifics is sensory-oriented modelling.

Sensory-oriented models are computational models that utilize, as inputs, the same stimuli as presented to infants, focusing on how signal-level details affect behaviour (Kleiner & Banks, 1987; Lovett & Scassellati, 2004; Prince, Helder & Hollich, 2005; Sirois, 2005). Sensory-oriented models provide an end-to-end explanation of infant behaviour – from raw sensory input to behavioural output. In the case of audiovisual speech, sensory-oriented models could allow us to determine the degree of correlation between the audio and visual streams at any given moment using the same stimuli as given to children. We might even find, using such an analysis, that audiovisual correlations can shift over time – to the point of having the visual motion of the 'unsynchronized' talker temporarily be more correlated with the audio than the actual speaker. In such a situation, an infant looking to the 'unsynchronized' talker is actually an indication of successful audiovisual integration. One might never know this without signal-level exploration of the stimuli.

Address for correspondence: George Hollich, Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, USA; e-mail: ghollich@purdue.edu

Invited target article for B. McMurray and G. Hollich (Eds.) 'Core computational principles of language acquisition: Can statistical learning do the job?' Special section in *Developmental Science.*

Cup Audio Condition

"The cup was bright and shiny. The clown drank from the red cup. The other one picked up the big cup. His cup was filled with milk. Meg put her cup back on the table. Some milk from your cup spilled on the rug."

OR

Dog Audio Condition

"The dog ran around the yard. The mailman called to the big dog. He patted his dog on the head. The happy red dog was very friendly. Her dog barked only at squirrels. The neighborhood kids played with your dog."

Figure 1 *The audiovisual preferential looking stimuli. Only one audio condition is played at a time, and each lasts 15 seconds. The task is to locate the synchronized talker.*

Furthermore, signal-level analysis could help to separate possible mechanisms of audiovisual integration. That is, although infants could be computing a running correlation, noticing how the precise patterns of visual and auditory changes correspond, infants could also be doing something considerably less sophisticated: perhaps infant audiovisual integration involves nothing more than focusing on large instantaneous changes such as the onset or offset of motion simultaneous with sound? If a mouth starts moving when speech begins, adults are likely to connect the mouth with the speech – even if the match isn't exact. This idea is central to the famous ventriloquist effect (Driver, 1996). Signal-level models can explicitly simulate what behaviour would look like for each of these cases.

In this paper, we compare the moment-by-moment performance of 8-month-olds in an audiovisual preferential looking task with several signal-level computational models. These computational models enable precise estimates of where an infant might look as a function of the experimental video stimuli at any given moment. Our question is how much of infant behaviour can be accounted for by such signal-driven factors? Put another way, how closely does infant behaviour map to the stimuli?

Infant audiovisual preferential looking dataset

We use an infant audiovisual preferential looking dataset from the control condition of a study of infant audiovisual synchrony detection in noise (Hollich, Prince, Mislivec & Helder, 2005). In this earlier study, infants were exposed to side-by-side video clips of talking heads while the audio alternatively matched the video on the infant's right or left (see Figure 1).

Participants

The participants were 20 eight-month-olds ($M = 8.12$ months, $SD = 0.52$, 9 females, 11 males) with no history of hearing problems or language delay. Four additional participants were excluded owing to fussiness (less than 65% total looking time) or fixated looking (looking more than 75% of the time towards one side).[1] Parents were contacted via birth records. Consistent with the local population distribution, the vast majority of subjects were from middle-class Caucasian homes, with less than 10% participation by ethnic or racial minorities.

Stimuli

The video clips for the models and infant data were taken from Hollich, Newman and Jusczyk (2005) and displayed a close-up of the face of a Caucasian female speaker of American English as she read two passages (regarding either a cup or a dog) in infant-directed speech (an exaggerated, excited manner of speaking that is known to attract infant attention). These video clips were trimmed and combined using Apple's Quicktime Player Pro, to create side-by-side clips that were each 15 seconds in length.[2] The video that matched the cup audio appeared on the left of the screen, and the video matching the dog audio appeared on the right. The audio from the movies was then adjusted so that only one of the vocal tracks

[1] These values for exclusion are standard for such studies (see Hollich, Hirsh-Pasek & Golinkoff, 2000).

[2] See Cup and Dog stimuli from http://stimbank.talkbank.org/Prince-Hollich/video-clips.html.

played. This resulted in two, 15-second, splitscreen video clips, the audio for which matched only one of the clips (see Figure 1). Each clip was played to each child once; however, the final stimuli played to infants included both cup and dog audio conditions as well as two 'noise' conditions (not included in our analysis), in which the cup and dog audio tracks were played along with a distractor audio (a monotone male voice reading the methods section of a paper) played at equal loudness. The order of presentation was counterbalanced across children such that half of the children heard these noise conditions first, and the other half heard these conditions second. Because initial analysis did not find any order effects (from having heard the noise conditions first), data regarding the cup and dog conditions (heard in the clear) from both orders were combined to produce the dataset used in this paper.

Apparatus and procedure

After explaining the procedure and having a legal guardian of the infant sign a consent form, the infant was seated on the caregiver's lap approximately 45 inches from a large white screen (65 inches along the diagonal). Black curtains covered all but the screen and the lens of the camcorder used to record infant responding. The image on the screen was displayed by an LCD projector attached to an Apple computer. The audio was played using an amplifier attached to the audio output of the computer driving the display. The stimuli were played monaurally[3] through a single speaker set in the centre between the two videos and were 72 dB in average amplitude. After the infant was seated comfortably, and the caregiver blindfolded, the video was played to completion regardless of infant looking, although infant looking was generally quite high. Infants' average looking time per each of the two passages was 9.48 seconds (out of 15 seconds; 63.2%) with a standard deviation of 2.72 seconds.

Microgenetic coding

Coding of infant looking time was conducted off-line using video captured from the camcorder (using Apple's Quicktime Broadcaster software) and a coding program written by the first author (Hollich, 2005). This program allowed coders, blind to the condition being run, to step through the videos frame-by-frame and mark the beginnings and ends of each left and right look for the entire video. These marks were then exported to an Excel spreadsheet for analysis. Because of the frame-by-frame nature of this process, this method is highly precise (to within one-thirtieth of a second). Inter-rater reliability (as tested by random re-coding of 20% of the data) was above .98.

Of particular interest to the current simulations is the proportion of infants looking towards the cup video by frame. This proportion is calculated by dividing the number of infants looking to the left (to the cup) for each frame of the video (1/30 s) by the sum of all infants looking (either left or right) for that frame. Random looking is thus .5; proportions greater than .5 indicate looking to the cup video, and proportions lower than .5 indicate looking to the dog video. One would expect these proportions to be different depending on the audio. More specifically, the higher the audiovisual correlation between the audio and the cup video, the higher the expected proportion for that frame. Although the fact that one of the videos was synchronized with the audio would seem to give it an advantage, as noted in the Introduction, it is entirely possible that, at some moments, the 'unsynchronized' video might have a higher audiovisual correlation. The computational models allow us to examine this possibility, as well as providing a theoretical baseline for where infants would look in the absence of sound.

Signal-level simulation of infant preference

All models begin with signal-level estimates of activity. They accept, as input, the same stimuli as presented to infants. The estimates of visual activity started with two visual streams (one for the cup display and one for the dog) in DV format (720 × 480 pixels per frame, at a rate of 29.97 frames per second). The audio streams (one for the cup audio and one for the dog) consisted of a single (mono) audio channel, and used a sampling rate of 44.1 kHz. All signal-level estimates, video or audio, were normalized estimates of change rather than raw scores. We chose to consider normalized changes in audio and visual because of the known normalizing properties of the cortex (Loritz, 1999).[4]

Specifically, for our estimate of visual change, we focused on obtaining an estimate of normalized visual change (NVC) based on changes in grey-scale pixel intensity (see Butz & Thiran, 2002).[5] This method computes the change of intensity of pixels across three successive visual frames by ignoring the middle frame and summing a region of 9 pixels surrounding a centre pixel in the starting and ending frames of this triple of frames. These values are then subtracted to arrive at an intensity-change value for each pixel. These intensity-change values were then summed to produce an estimate of frame-intensity change (FIC) for each triple of frames. These FIC values were extracted

[3] In this study, we presented the audio monaurally in order to limit the number of experimentally varied factors and to make it necessary for infants to use audiovisual integration to succeed. This is not as unrealistic as it might seem: infants' localization skills are rather poor and many objects that have noticeable visual separation would appear to have identical audio origination from an infant's standpoint (Muir, Clifton & Clarkson, 1989).

[4] Not coincidentally, preliminary analysis indicated that such normalized estimates produced models with a better fit to the infant data.

[5] Other estimates of visual change (optic flow, frame subtraction, etc.) may produce slightly different data. However, preliminary testing indicated a wide overlap between the estimates.

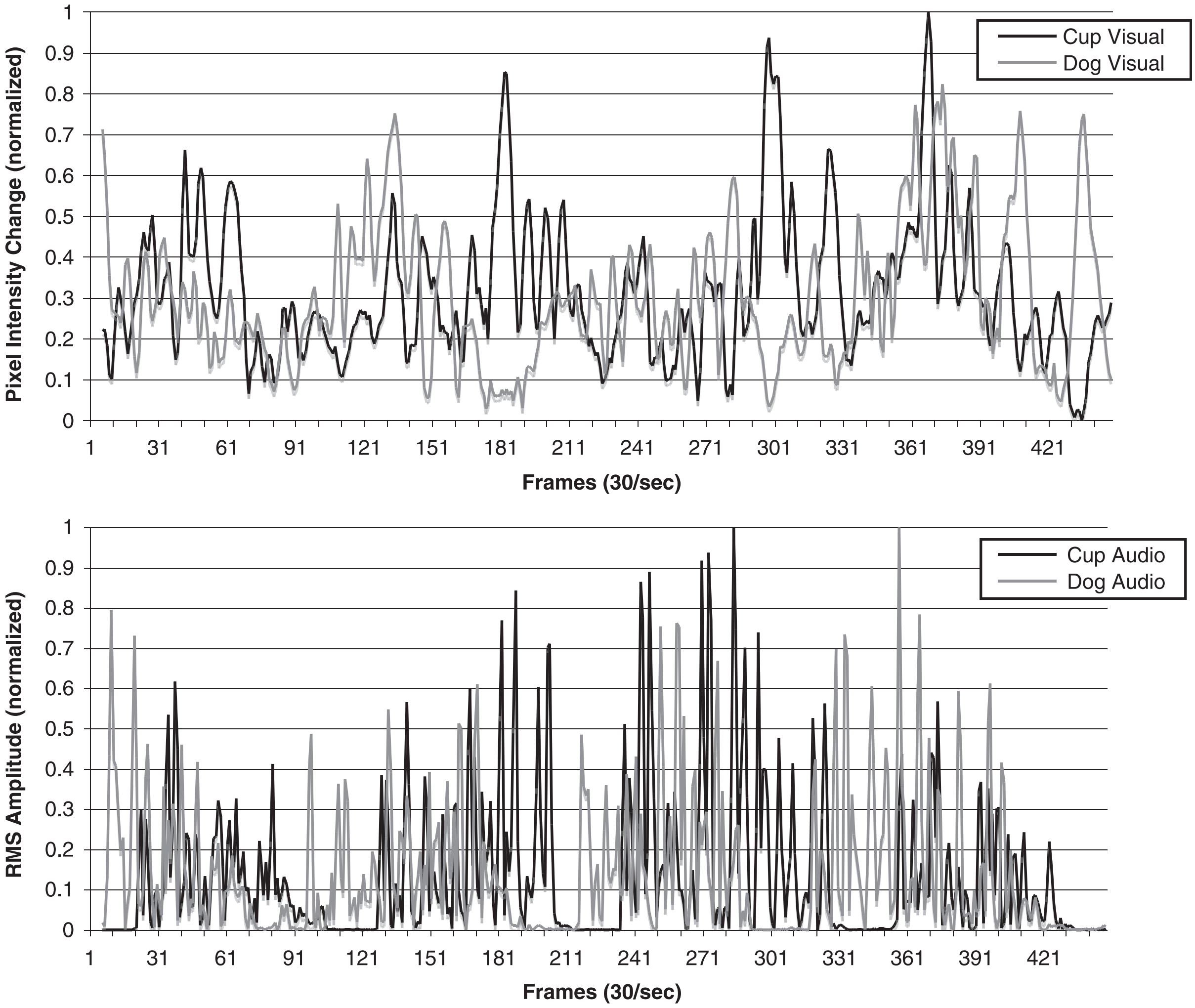

Figure 2 *Estimates of visual change (top panel) and audio change (bottom) across the trial.*

using the iSenseStream program[6] written by the second author (see Mislivec, 2004; Prince & Hollich, 2005) and then normalized on a scale from 0 to 1 by subtracting the minimum FIC value (across the entire video) from the raw score and then dividing by the FIC max and FIC min difference:

$$\mathrm{NVC} = \frac{(\mathrm{FIC}) - \min_{\mathrm{FIC}}}{\max_{\mathrm{FIC}} - \min_{\mathrm{FIC}}}. \tag{1}$$

These normalized FIC values are presented in the top panel of Figure 2.

For our estimate of auditory activity, we focused on normalized audio change (NAC) values.[7] These were calculated for each condition by extracting the RMS amplitude for each audio frame using the iSenseStream program, subtracting these values across each triplet of frames, and then normalizing the values similarly to above. These values are presented in the bottom panel of Figure 2.

Audiovisual correlation model

Given the signal-level estimates thus obtained, there are several possibilities for simulation of audiovisual integration. We considered two in our simulations. First, one could compute a running correlation between visual changes (as indexed by NVC values for the two sides) and changes in audio (as indexed by NAC values) over some time window. Higher correlations between the audio and visual on one side compared with the other should lead to greater looking. Notice that this model assumes that the child would have some memory for the changing patterns of stimulation over a particular time window.

For this audiovisual *correlation* model, we computed two running correlations across 10 or 20 prior frames: one correlated the NVC values for the left visual with the NAC values for that audio condition, and the other

[6] SenseStream is available on the Internet at http://www.cprince.com/PubRes/IKAROS/iSenseStream, and is implemented within the Ikaros modelling framework (http://asip.lucs.lu.se/IKAROS).

[7] We intentionally did not include frequency information in these models, although we suspect this may allow the models to separate streams of speech.

correlated those same NAC values with the NVC values from the right. Thus, on any given frame, the model had an estimate of the degree of correlation between the audio change (NAC) and the visual change (NVC) values for the previous third of a second (10 frames) or two-thirds of a second (20 frames). These correlations were added to 1 to produce a single, positive number ranging from 0 to 2, where higher numbers indicate greater positive correlations. We added one because we found it likely that a strong negative correlation would drive the infants to look away from that side. That is, if the audio were changing while the visual was not, that would signal infants to look elsewhere for the match.

Finally, in this and all other models we then divided the values for the right by the sum of the scores for the left and right to obtain a preference score. In this manner, we had a proportion roughly analogous to the proportion of infants looking to a given side per frame. Thus, the final formula for the correlation model was

$$\frac{1 + r_{cc}}{(1 + r_{cc}) + (1 + r_{cd})}, \qquad (2)$$

where r_{cc} is the correlation between the NAC values for the cup audio and the NVC values for the cup visual, and r_{cd} is the correlation between the NAC audio values for the cup and the NVC values for the dog visual.

Instantaneous additive model

Alternatively, one could simulate simple instantaneous additive firing where similar-sized changes in audio and visual activity lead to greater responding to that side. Such a model assumes no memory on the part of the child for changing patterns, and in this manner similar-sized increases or decreases in audio simultaneous with visual changes simply make the visual for that side more salient. This second, instantaneous additive, interpretation is more consistent with the known behaviour of superadditive neurons that have been found in the superior colliculus of monkeys (Stein, Wallace, Stanford & Jiang, 2002), behaviour which would suggest that audiovisual integration is an instantaneous activity.

For this additive model, we computed the absolute value of the difference between the NVC value for each side and the NAC audio values and subtracted from one. The less of a difference (i.e. the more the values 'match'), the higher the score per side; the more of a difference (i.e. the less the values 'match'), the lower the score per side. Again, we computed a proportion to obtain a simulation of preference. The final formula for the additive model was thus

$$\frac{1 - |\Delta_{cc}|}{(1 - |\Delta_{cc}|) + (1 - |\Delta_{cd}|)}, \qquad (3)$$

where Δ_{cc} is the instantaneous difference between the NAC values for the cup audio and the NVC values for the cup visual, and Δ_{cd} is the instantaneous difference between the NAC values for the cup audio and the NVC values for the dog visual.

Visual model

We included one additional model, the visual model, which considers only the visual in the estimate of looking preference. The rationale behind this model was a check to see how much of the infant data could be accounted for solely by visual change (as indexed by NVC values), given that infants have been shown to look towards sudden visual movement (Henderson, 2005). The formula for this visual model was simply the NVC values for the cup visual divided by the sum of the NVC values for the cup and dog visuals.

Finally, we smoothed the output of all models by using a running average across *X* frames, where *X* was 20 or 40.[8] This was done to limit the speed at which the model shifted its looking preferences, in order to be more consistent with the infant data. That is, although the model can shift looking preference quite suddenly, the proportion of infants never makes such a drastic shift from looking to the cup to the dog video.

Results and discussion

The proportion of infants looking towards the cup video by frame and audio condition is presented in Figure 3. On average, infants hearing the cup audio looked predominantly at the cup video (58%), whereas those hearing the dog audio looked predominately at the dog video (52%). However, this looking was not uniformly distributed throughout the trial. Notably, infants in both conditions looked towards the unsynchronized video at times (e.g. frame 181 or 241). How well, then, do the models capture these infant results?

Figure 4 presents a graph of the correlation model for each condition (top panel) and the visual model (bottom panel), with the smoothing for both models set at 20.[9] Interestingly, the audiovisual model estimates seem to line up with the infant data: the cup audio tends to be above the dog audio in similar places for both infants and the model. Thus, for example, a higher proportion of infants looked at the non-target on frames 181 and 241, similar to the models. Furthermore, the infant and model preferences for the matching video seem strongest at times when the audio for the other track was silent (e.g. frame 121 or 211; see Figure 2, bottom panel), suggesting a role for onset and offset of sounds in audiovisual integration. So at least qualitatively, the signal-level models appear to relate well to infant behaviour. The next section provides a statistical analysis of the amount of variance accounted for by each of the models at each of two different levels of smoothing.

[8] In order to maintain the time-lock with the infant data, this smoothing was always centred on the original frame. Thus, a smoothing of 20 actually included the prior 10 and the next 10 frames (for a total of 21 frames).

[9] Although the 'best fit' models are presented, the other models produced similar results. Detailed data from them is reported in the next section.

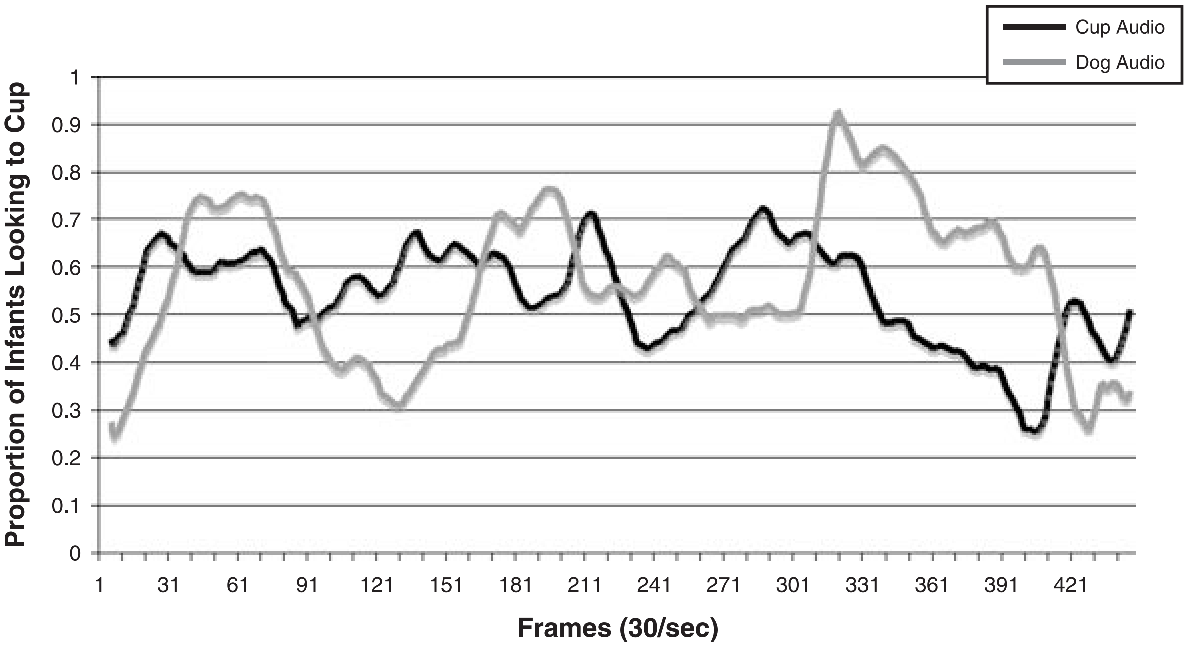

Figure 3 *Proportion of infants looking to the cup video across the trial and condition.*

Regression analysis including visual and audiovisual models

Table 1 gives the results of a two-step hierarchical regression analysis of the variance accounted for by each model. The first step of this analysis finds the amount of variance ascribable to the visual model at each level of smoothing. The second step examines the amount of variance accounted for by the combination of the visual model with each audiovisual model for that level of smoothing. Reported are values of R^2 as well as the change in variance (ΔR^2) accounted for by that step.

The overall results are small but significant, and all correlations are in the positive direction. They reveal that the visual model accounted for between 7.61 and 21.2% of the variance in infant performance, and the various incarnations of the audiovisual models accounted for an additional .91 to 5.56% of the variance, with the correlation models performing the best. Although the amount of variance accounted for is not large, recall that the effect sizes for infants are also quite small. Presumably the models should not do any better than the infants.

Table 1 *Linear regression: amount of variance in infant data accounted for by condition, model type, and amount of smoothing*

	Cup		Dog	
Model type	R^2	ΔR^2	R^2	ΔR^2
Step 1 – Visual				
Smoothing 10	.0761**	.0761**	.1235**	.1235**
Smoothing 20	.0924**	.0924**	.2121**	.2121**
Step 2 – Smoothing 10				
Additive	.1058**	.0328**	.1362*	.0094*
Correlation 10	.1080**	.0319**	.1507**	.0272**
Correlation 20	.0852*	.0091*	.1791**	.0556**
Step 2 – Smoothing 20				
Additive	.1184**	.0260**	.2251*	.0130*
Correlation 10	.1424**	.0500**	.2254*	.0133*
Correlation 20	.1019*	.0095*	.2361**	.0240**

* $p < .05$, ** $p < .001$.

In addition, the trials were quite long; it is likely that the models account for different amounts of variance depending on whether it is early or late in the trials. That is, infants may be swayed more by visual or audiovisual information early in the trial and then become habituated to that information as the trial moves on. Indeed, a weakness of the current models is that they do not habituate. Thus, a strong audiovisual correlation later in the trial has as much of an effect as one earlier in the trial, yet this may not be the case for infants, who become bored even with factors that were once highly interesting. For this reason, we split the data into 5-second non-overlapping blocks and conducted an analysis of variance (ANOVA) for the effect of block on each of the models. By including block in these analyses, the visual models plus either audiovisual model accounted for a far greater amount of the variance. For example, with smoothing at 20 and using the correlation model across a time window of 10, the whole analysis now accounted for 55.35% of the variance, $F(11, 396) = 44.62$, $p < .0001$, $R^2 = .5535$, in the cup condition and for 57.99% of the variance, $F(11, 396) = 49.70$, $p < .0001$, in the dog condition, with a peak of 72.33% of the variance accounted for in block 1. Besides the dramatic increase in the amount of variance accounted for, the most notable finding of this analysis was the significant block effect for both the cup, $F(2, 396) = 68.88$, $p < .0001$, and the dog, $F(2, 396) = 60.50$, $p < .0001$, condition, as well as the significant block by

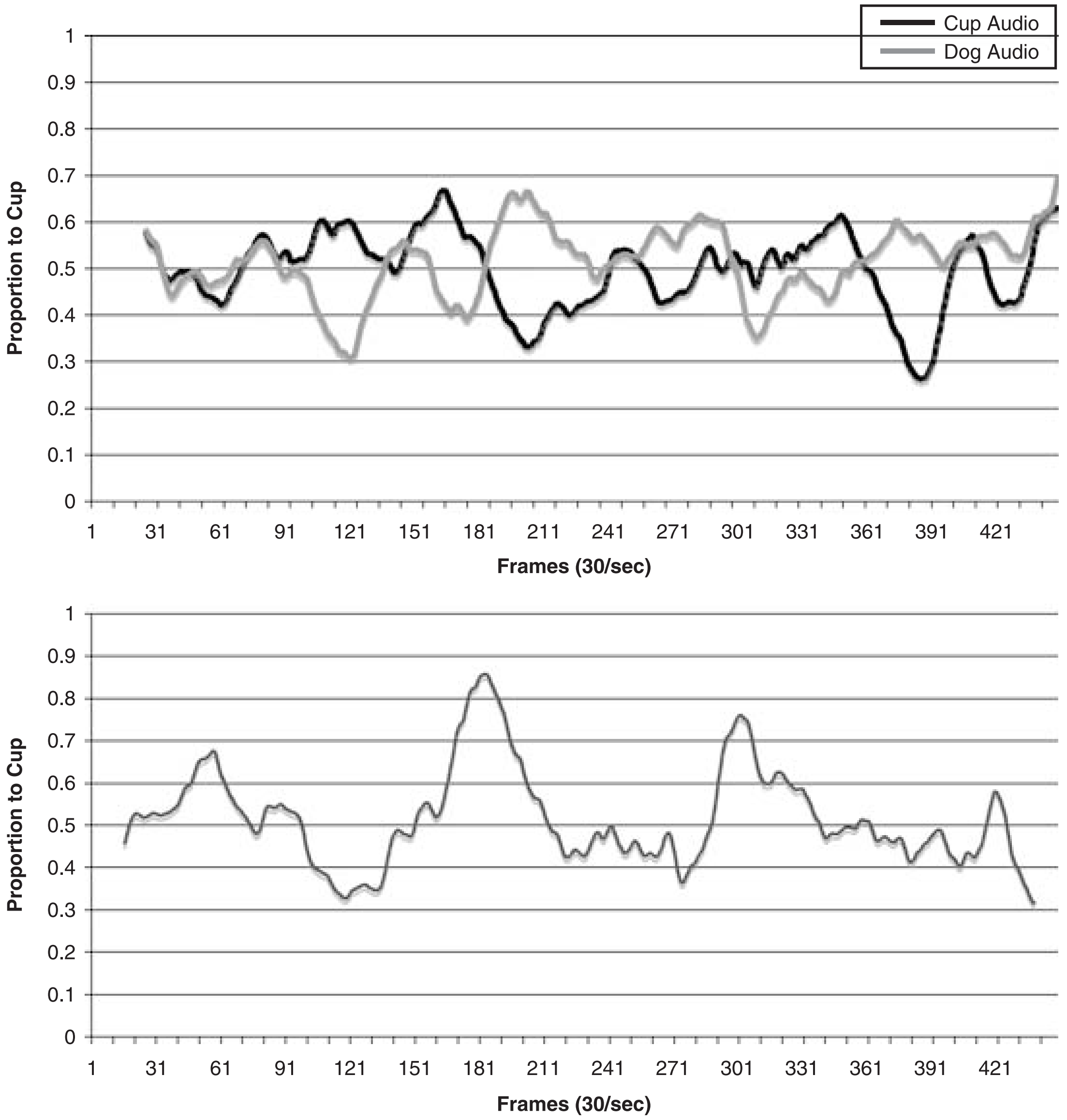

Figure 4 *Proportion looking to the cup video across the trial and condition. (Top) Correlation models (smoothing 20). (Bottom) Visual model (smoothing 20).*

model interactions, all *F*s > 30, *p* < .0001. Because of these large block effects across all the models, we repeated the previous analysis, separating by block (see Table 2).

As suggested by the block by model interaction, the amount of variance accounted for by each model varies widely over the course of the trials. For the dog audio conditions, the visual model accounts for the variance particularly well early in the trial (Block 1): 72.3% of the variance with smoothing 20. In contrast, for the cup audio, the visual estimate was much smaller early in the trial but accounted for 63.5% of the variance late in the trial (Block 3). Among the audiovisual models, the correlation model seemed the most successful, especially early on (Block 1), when it accounted for up to 28.6% (cup audio, smoothing 20) of the variance. The additive model, by contrast, was especially poor, accounting for no additional portion of the variance early in the model (Block 1). Finally, combining the visual with the audiovisual models accounted for up to 66.4% and 80.3% of the variance for the cup (Block 3) and dog (Block 1) conditions, respectively, with unique contributions for the audiovisual models of up to 28.6% (Block 1) of the variance.

General discussion and future directions

The principal finding of our work is that a significant portion of the variance in infants' average looking behaviour in an audiovisual speech task can be accounted for by signal-level perceptual analysis, particularly when time is included as a factor.[10] Thus, a model of visual salience,

[10] Because these data are at the group level, it is possible that no individual is actually following the audiovisual correlations. Instead, when the correlation is high, the likelihood of any given infant looking away is simply less than that when the correlation is low.

Table 2 *Linear regression: amount of variance in infant data accounted for by block, condition, model type, and amount of smoothing*

	Cup		Dog	
Model type	R^2	ΔR^2	R^2	ΔR^2
Block 1				
Step 1 – Visual				
Smoothing 10	.0034	.0034	.5332**	.5332**
Smoothing 20	.0024	.0024	.7232**	.7232**
Step 2 – Smoothing 10				
Additive	.0169**	.0135**	.6921*	.0094*
Correlation 10	.0623**	.0589**	.5341**	.0272**
Correlation 20	.0157*	.0123*	.5637**	.0556**
Step 2 – Smoothing 20				
Additive	.0024	.0000	.8027**	.0795*
Correlation 10	.1171**	.1147**	.7316**	.0084
Correlation 20	.2879**	.2855**	.7233**	.0001
Block 2				
Step 1 – Visual				
Smoothing 10	.0002	.0002	.2948**	.2948**
Smoothing 20	.0025	.0025	.3589**	.3589**
Step 2 – Smoothing 10				
Additive	.0002	.0000	.3972**	.1024**
Correlation 10	.0385*	.0383**	.4086**	.1138**
Correlation 20	.0407*	.0405**	.3766**	.0818**
Step 2 – Smoothing 20				
Additive	.0028	.0003	.4043**	.0454**
Correlation 10	.0373*	.0348*	.4971**	.1382**
Correlation 20	.0026	.0001	.4573**	.0984**
Block 3				
Step 1 – Visual				
Smoothing 10	.4351**	.4351**	.0122**	.0122**
Smoothing 20	.6353**	.6353**	.0712**	.0712**
Step 2 – Smoothing 10				
Additive	.4352**	.0001	.0130	.0008
Correlation 10	.4795**	.0444**	.0567**	.0445**
Correlation 20	.4387**	.0036	.0647**	.0525**
Step 2 – Smoothing 20				
Additive	.6354	.0001	.0832	.0120
Correlation 10	.6642**	.0289**	.2429**	.1717**
Correlation 20	.6516**	.0163*	.0827	.0115

* $p < .05$, ** $p < .001$.

based on nothing more than coarse visual change, was able to account for more than 72% of infants' performance, at times. Audiovisual models, based on the correlation between auditory and visual change or on instantaneous changes in the match between audio and visual activity, were similarly able to account for an additional 28.79% of the variation in infants' looking behaviour, at times.

The changing variance accounted for by these models over time would appear to indicate a change in the infants' weighting of factors throughout the trial. One audio condition (the dog condition) exclusively matched the visual model early on (21%) and then switched to matching the audiovisual model, whereas the other audio condition (the cup condition) exclusively matched the audiovisual model first and then switched to matching the visual model (64%), possibly as a result of habituation from so much initial time spent looking at the audiovisual match. Such results would suggest that infants' tendency to notice audiovisual synchrony may be determined by a complex non-linear interaction resulting from habituation of their sensitivity to visual motion and/or audiovisual correlations as the trial proceeds. Future models will explicitly include habituation and non-linear interactions between signal-level models.

Furthermore, given that issues with sensory integration seem to lie at the heart of many childhood disorders, including autism (Brock, Brown, Boucher & Rippon, 2002), and that attentional difficulties can lead to severe impairments in learning (Tsao, Liu & Kuhl, 2004), it would seem of critical theoretical and clinical importance to understand the basic mechanisms underlying individual differences in perceptual attention. Future simulations will explore the nature of such differences by modelling how infants who weigh visual or auditory factors heavily would behave as compared with infants who are particularly sensitive to audiovisual information.

Conclusion

These models are just the first step in our signal-level exploration of infants' audiovisual integration: they don't include habituation or examine the nature of individual differences. Nonetheless, these simple signal-level audiovisual models combined with a coarse estimate of visual change accounted for a significant portion of the variance in looking times. We now have a new means by which one can examine the mechanisms underlying audiovisual speech integration by infants.

References

Bahrick, L.E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, **13**, 99–102.

Brock, J., Brown, C.C., Boucher, J., & Rippon, G. (2002). The temporal binding deficit hypothesis of autism. *Development and Psychopathology*, **14**, 209–224.

Butz, T., & Thiran, J.-P. (2002). Feature space mutual information in speech–video sequences. *IEEE International Conference on Multimedia and Expo*, vol. II (pp. 361–364). Piscataway, NJ: IEEE Press.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, **11**, 478–484.

Driver, J. (1996). Enhancement of selective listening by illusory mis-location of speech sounds due to lip-reading. *Nature*, **381**, 66–68.

Henderson, J.M. (2005). Human gaze control during real-world scene perception. *Trends in Cognitive Science*, **7**, 498–504.

Hollich, G. (2005). *SuperCoder (Version1.5)* [Computer software] West Lafayette: Purdue University.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: an emergentist coalition model of word learning. *Monographs of the Society for Research in Child Development*, **65** (3, Serial No. 262).

Hollich, G., Prince, C., Mislivec, E., & Helder, N. (2005). Audiovisual synchrony in language learning. *Paper presented* *at the Xth International Congress for the Study of Child Language*.

Hollich, G., Newman, R., & Jusczyk, P. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, **76**, 598–613.

Kleiner, K.A. & Banks, M.S. (1987). Stimulus energy does not account for 2-month-olds face preferences. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 594–600.

Loritz, D. (1999). *How the brain evolved language*. New York: Oxford Press.

Lovett, A., & Scassellati, B. (2004). Using a robot to reexamine looking time experiments. In J. Triesch and T. Jebara (Eds.), *Proceedings of the Third International Conference on Development and Learning* (ICDL 04) (pp. 284–291). LaJolla, CA: UCSD Institute for Neural Computation.

Mislivec, E.J. (2004). *Audio-visual synchrony for face location and segmentation*. Undergraduate research opportunity project, University of Minnesota Duluth. Available at: http://www.cprince.com/PubRes/SenseStream

Muir, D.W., Clifton, R.K., & Clarkson, M.G. (1989). The development of a human auditory localization response: a U-shaped function. *Canadian Journal of Psychology*, **43**, 199–216.

Newman, R.S., & Jusczyk, P.W. (1996). The cocktail party effect in infants. *Perception & Psychophysics*, **58**, 1145–1156.

Pickens, J., Field, T., Nawrocki, T., Martinez, A., Soutullo, D., & Gonzalez, J. (1994). Full-term and preterm infants' perception of face-voice synchrony. *Infant Behavior and Development*, **17**, 447–455.

Prince, C.G., & Hollich, G. (2005). Synching models with infants: A perceptual-level model of infant audio-visual synchrony detection. *Journal of Cognitive Systems Research*, **6**, 205–228.

Prince, C.G., Helder, N.A., & Hollich, G.J. (2005). Ongoing emergence: a core concept in epigenetic robotics. In L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov, & C. Balkenius (Eds.), *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, 123* (pp. 63–70).

Siegler, R.S. (2006). Microgenetic analyses of learning. In W. Damon and R.M. Lerner (Series Eds.) and D. Kuhn and R.S. Siegler (Vol. Eds.), *Handbook of child psychology: Volume 2: Cognition, perception, and language* (6th edn, pp. 464–510). Hoboken, NJ: Wiley.

Sirois, S. (2005). Hebbian motor control in a robot-embedded model of habituation. *Proceedings of the International Joint Conference on Neural Networks* (IJCNN 2005) (pp. 2772–2777). Piscataway, NJ: IEEE Press.

Stein, B.E., Wallace, M.W., Stanford, T.R., & Jiang, W. (2002). Cortex governs multisensory integration in the brain. *The Neuroscientist*, **8**, 306–314.

Tsao, F.-M., Liu, H.-M., & Kuhl, P.K. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development*, **75**, 1067–1084.