

Taking Synchrony Seriously: A Perceptual-Level Model of Infant Synchrony Detection

Christopher G. Prince, George J. Hollich*, Nathan A. Helder, Eric J. Mislivec, Anoop Reddy,
Sampanna Salunke, and Naveed Memon

Department of Computer Science
University of Minnesota Duluth, Duluth, MN USA
chris@cprince.com, nhelder@nerp.net, {mislivi001,
parl0020, salu0005, memo0005}@d.umn.edu

*Purdue University
Department of Psychological Sciences
West Lafayette, IN 47907 USA
ghollich@psych.purdue.edu

Abstract

Synchrony detection between different sensory and/or motor channels appears critically important for young infant learning and cognitive development. For example, empirical studies demonstrate that audio-visual synchrony aids in language acquisition. In this paper we compare these infant studies with a model of synchrony detection based on the Hershey and Movellan (2000) algorithm augmented with methods for quantitative synchrony estimation. Four infant-model comparisons are presented, using audio-visual stimuli of increasing complexity. While infants and the model showed learning or discrimination with each type of stimuli used, the model was most successful with stimuli comprised of one audio and one visual source, and also with two audio sources and a dynamic-face visual motion source. More difficult for the model were stimuli conditions with two motion sources, and more abstract visual dynamics—an oscilloscope instead of a face. Future research should model the developmental pathway of synchrony detection. Normal audio-visual synchrony detection in infants may be experience-dependent (e.g., Bergeson, et al., 2004).

1. Introduction

We are exploring formal models of infant synchrony detection (Prince, Helder, Mislivec, Ang, Lim, & Hollich, 2003) in order to further elucidate the mechanisms of infant development and to create practical robotic systems (e.g., Weng, McClelland, Pentland, Sporns, Stockman, Sur, & Thelen, 2001). Synchrony detection mechanisms appear critically important for young infant learning and cognitive development, and have been strongly implicated in developments ranging from word-learning (Gogate & Bahrick, 1998), to learning arbitrary intermodal relations (Bahrick, 2001; Slater, Quinn, Brown, & Hayes, 1999), object interaction skills (Watson, 1972), emotional self-awareness and control (Gergely & Watson, 1999), naive theory understanding (Gopnik & Meltzoff, 1997), and learning related to the self (Rochat & Striano, 2000). Because synchrony detection plays such a pervasive role in infant development, it seems important to increase our understanding of the mechanisms utilized by infants. It is one thing to tacitly acknowledge the importance of synchrony detection, but quite another to use formal

modeling to help us build more specific psychological theories of these developmental mechanisms (Shultz, 2003). To accomplish the formal modeling we must carefully consider what synchrony *is*, and what specifics of audio-visual representation are necessary to recreate the synchrony detection abilities of infants. Additionally, we want to create practical robotic systems that utilize knowledge of how infants develop their skills. We suggest that understanding the mechanisms related to infants' developing synchrony detection skills can assist us in designing algorithmic mechanisms. For example, understanding the kinds of synchrony detection skills possessed by infants should help us specify requirements for synchrony detection algorithms and using these algorithms in epigenetic robotics may enable the robots to develop synchrony detection skills in a manner analogous to infants.

In epigenetic robotics, the need for detecting synchrony and contingency (e.g., detecting relationships between the motor output of a robot and the ensuing sensory consequences) has recently been considered (Asada, MacDorman, Ishiguro, & Kuniyoshi, 2001; Fasel, Deak, Triesch, & Movellan, 2002; Lungarella, Metta, Pfeifer, & Sandini, in press). For example, in a robotic implementation, Arsenio and Fitzpatrick (2003) detected rhythmic audio-visual synchrony relationships using histograms of durations between signal features to measure periodicity within audio and within visual stimuli, and histogram comparisons to evaluate cross-modal synchrony.

In this paper we use algorithmic methods that directly compute audio-visual synchrony relationships between low-level audio-visual features (e.g., RMS audio and grayscale pixels). The algorithm we use is based on that of Hershey and Movellan (2000; *HM* algorithm in the following), which while detecting audio-visual synchrony, defined as Gaussian mutual information, may also be useful in contingency detection (Helder, 2003). The HM algorithm was originally applied to the problem of detecting synchrony between a stream of visual data and a stream of audio data in order to find the spatial position of a vocalizing person in the visual image dynamics. The point of highest audio-visual synchrony when someone speaks is approximately the lips (see also Nock, Iyengar, & Neti, 2003). The HM algorithm is relatively general, detecting temporal synchrony between two time-based input streams and thus makes an excellent starting point for modeling general synchrony detection mechanisms in infants. Infants are known to be skilled at

detecting temporal, sensory-relational (contingencies wherein the amount of sensory change is expected to be proportionate to the amount of motor force exerted), and spatial synchrony (Gergely & Watson, 1999).

In this paper we compare a model based on the HM synchrony detection algorithm to four empirical studies of synchrony detection in infants. Our hypothesis is: *Can a single general-purpose synchrony detection mechanism, estimating audio-visual synchrony from low-level signal features, account for infant synchrony detection across audio-visual speech integration tasks of increasing complexity?* Our first infant-model comparison looks at the case of integrating punctuate visual movements of an object and synchronous audio presentations of a word. The second infant-model comparison looks at the more difficult case of integrating the continuous visual movements of a face with the speech stream. The third comparison looks at the harder task of separating out an irrelevant speech stream using the continuous visual movements of a face. The final comparison uses stimuli that may be even harder to process — we substitute the continuous visual movements of an oscilloscope for the speech movements of a face.

The remainder of this paper comprises a description of the video stimuli and algorithm, a series of four infant and modeling comparison sections, and a closing discussion section. The comparison sections correspond to four sets of increasing complexity audio-visual data: A, B, C and D. These comparison sections first present results from work with infants on either comparable (Stimuli A & B) or identical stimuli (Stimuli C & D), and then present results from our perceptual model of synchrony detection based on the HM algorithm (Mislivec, 2004). The infant results for Stimulus A and B are from the general literature on infant psychological development (Dodd, 1979; Gogate & Bahrick, 1998). Infant results from Stimulus C and D are from the lab of the second author (Hollich, Newman, & Jusczyk, in press).

2. Stimuli and Algorithm

We constructed digital video clips containing various types and degrees of temporally synchronous audio-visual stimuli. We processed this video data as input to our synchrony detection program, *SenseStream* (Mislivec, 2004), based on the HM algorithm. In the *SenseStream* program, we measured synchrony using either a centroid method (from HM), a connected region method, or an edge detection method. We developed the latter two methods to quantitatively estimate the degree of synchrony represented by the outputs of the HM algorithm. While the HM algorithm generates a topographic representation of synchrony (see below), it does not provide a scalar estimate of the synchrony in that representation. An alternative approach to quantitative audio-visual synchrony estimation, based on canonical correlation, is given by Slaney and Covell (2001).

Stimuli. MPEG-1 digital video files were used as inputs for the model, with 29.97 video frames per second,

44.1 kHz audio, and rendering with the highest settings for data rate (Adobe Premier 6.5), to minimize compression. These stimuli are summarized in Table 1. Stimulus A had one sound source (speech) and one visual motion source. In the video, the word “modi” was uttered nine times during intervals when a suspended object was in vertical motion in front of a white background. This stimulus approximates the synchronous condition stimulus used by Gogate and Bahrick (1998), which was used by those authors to assess if syllables co-uttered with moving objects would enhance infants’ learning of audio-visual associations. The clip duration was 30s.

Stim.	Sound Source(s)		Visual Motion Source(s)	
	No.	Description	No.	Description
A	1	“Modi” word	1	Vertical object motion
B	1	Male voices alternating	2	Two males talking
C	2	Female voice & male voice	1	Face of female
D	2	Female voice & male voice	1	Oscilloscope: female voice

Table 1: Stimuli design for infant-model comparisons of synchrony detection. Video files are on the Internet: <http://www.cprince.com/PubRes/EpiRob04>

Stimulus B had one sound source at any one time (speech or background noise), but had two visual motion sources. In this clip, two adult males, fix-positioned on both sides of a split screen, were talking for 30s. For the first 5s, the audio source was from the right male, the next 5s from the left male, and the next 5s the audio source was background noise. The remaining 15s repeated this right, left, noise pattern with different video data. Two additional 30s clips, controls for the model only, were also used. Control 1 had the right speaker only (same position as before) talking for the clip duration (30s), and background visual stimuli on the left. Control 2 was analogous (and also 30s duration) but with the left person.

Stimulus C and D both had a single visual motion source and two audio sources. The audio sources had a female and a male speaking, while the visual motion source was synchronized with the female voice. The visual motion source for Stimulus C was the dynamic image of the female’s face as she was speaking, and for Stimulus D was a dynamic oscilloscope representation (amplitude analysis) of the female voice. Stimulus C and D each had two comparison video clips, in which only one of the male (Male-only condition) or female voice (Female-only condition) was present. Stimuli with both male and female voices are termed the *Both* condition. Stimulus C and D were from Hollich, et al. (in press). The duration of these clips was 22s.

Algorithm. The *SenseStream* program (Mislivec, 2004) implements Equation 3 of HM, which is repeated here as Equation 1.

$$I(A(t_k);V(x,y,t_k)) = \frac{1}{2} \log_2 \frac{|\sum_A(t_k)| |\sum_V(x,y,t_k)|}{|\sum_{A,V}(x,y,t_k)|} \quad (1)$$

Equation 1 computes the Gaussian mutual information between a pixel at location (x, y) across a series of S consecutive frames of visual data (V ; dimension $h \times w$ pixels) and the audio data (A) co-occurring with those visual frames. For example, with $S=15$ and 30 frames per second video, Equation 1 gives the mutual information between a “column” of pixels and the audio source across 1/2 second of audio-visual data. Higher values of mutual information are interpreted as higher levels of synchrony; lower values of mutual information are interpreted as lower levels of synchrony. The mutual information minimum is 0 (no synchrony). In the model processing described here, $S=15$. Equation 1 is applied by our model to each pixel of data in each visual frame after the first $S-1$ visual frames of a clip. As in HM we used grayscale pixels (0...255), and RMS audio (one scalar per visual frame). We refer to each $I(A(t_k);V(x,y,t_k))$ value computed using Equation 1 as a *mixel*, for *mutual information pixel*, and refer to the entire output display (dimension $h \times w$ mixels) as a *mixelgram*. These mixelgrams can be interpreted as topographic representations of synchronization between the incoming data streams. More specifically, the occasions that these mixelgrams are classified by human raters as perceptually relevant (e.g., containing shapes) correlate strongly with the intervals of MPEG data in which the audio-visual signal is synchronized (Vuppla, 2004). Figure 1 gives an example of a perceptually relevant mixelgram from processing the Stimulus A data.



Figure 1: A mixelgram from an interval of synchronous audio-visual data in Stimulus A. Mixelgrams are typically perceptually relevant only when the two input streams (e.g., audio-visual) are synchronous (i.e., co-varying; see Vuppla, 2004).

HM used the centroid of the mixelgram to determine the (x, y) location of the peak of any synchrony existing between the audio and visual data. To quantitatively estimate the degree of audio-visual synchrony represented by the mixelgrams, we devised two additional methods to augment the HM algorithm, each of which operated as functions of mixelgrams and resulted in scalar estimates of synchrony. Our *connected region* method was based on the observation that in some cases of synchrony mixelgrams have groups of mixels

with similar values, some of which are large groups, some of which are small. That is, in these cases of synchrony, there are often connected mixel regions and often substantial variation in the sizes of these regions. We therefore computed the variance in the sizes of the connected regions per mixelgram. Connected regions were defined (recursively) as having eight-neighbor mixels with values within a factor of 1.125 of each other (edge mixels have fewer neighbors). The other method was based on *edge detection*, and used a similar observation to that above. In this case, we used Equation 2 as an estimate of the degree of synchrony.

$$\sum_{i=1}^{h-w} Sobel_{3 \times 3}(Gaussian_{15 \times 15}(M)) \quad (2)$$

We first blurred the mixelgram (M), reducing noise by convolution with a 15×15 Gaussian filter. Sobel edge detection was then applied, and the resulting values were summed over the matrix.

3. Stimulus A: Object Motion and Speech

Infant data and background. Bahrick and colleagues (Bahrick, 2001; Bahrick & Lickliter, 2000) have suggested that audio-visual temporal synchrony is one of the most consistent and early relations to which infants are sensitive. For example, Gogate and Bahrick (1998) found that 7-month-olds could learn the link between speech sound and object only if the sound was presented synchronously with object movement. In that study, 48 infants were tested across three conditions: a synchronous movement condition ($n=16$ infants), a static condition ($n=16$), and an asynchronous condition ($n=16$). In the synchronous condition, infants saw a hand move an unfamiliar object (either a toy crab or a porcupine) forward, synchronous with the vowel “ahhh” (for the crab) or “eee” (for the porcupine). On half of the conditions, the vowel-object pairing was switched. In the static condition the audio was the same, but the hand was not seen and the objects did not move. In the unsynchronized condition the movements were the same as in the synchronized condition, however the vowels were uttered between the forward movements.

Infants were familiarized to one of these conditions and then tested to see if they noticed if the vowel pairing was changed (as indicated by increased looking to the display when the pairing was “switched”). Only in the synchronized condition did infants increase their looking time over control trials (where no change was observed). Specifically, infants increased their looking by an average of 4.68s – a large effect in such experiments. Indeed, 11 out of 16 infants in the synchronous condition showed the predicted response. In contrast, the infants in the other two conditions actually looked more on the control trials than in the test trials when the vowel pairing was changed. Only 7 total out of the 32 infants in these conditions showed evidence of having noticed the switch in sound object pairings. Thus, it appears that 7-month-olds can use synchrony to learn a link between speech and object.

Word learning is an extremely complicated task involving multiple cues (Hollich, Hirsh-Pasek, Golinkoff,

2002), and including numerous social-pragmatic factors (Baldwin, 1993; Bloom, 2002). While one of our long-term goals is to incorporate these multiple factors in a model of word learning, the scope of the first simulation here was much more mundane. Given that infants must have detected the synchrony between punctuate movement and sound to have succeeded in this task (i.e., Gogate & Bahrick, 1998), the goal of the model was to do the same.

Model. To simulate this sound and object-motion synchrony detection, our model was exposed to stimuli similar to that of Gogate and Bahrick (1998), i.e., Stimulus A which contains utterances of the word “modi” co-occurring with vertical object motion. The connected region method was used to generate quantitative synchrony estimates for this data. Figure 2 shows the model processing results regarding the estimated degree of synchrony for the Stimulus A data. The model appears to have tracked the synchrony very well: periods of high synchrony closely match the periods of audio onset and offset for utterances of the word “modi.” Audio-visual synchrony in these cases results from the word being uttered at the same time as the object is moved.

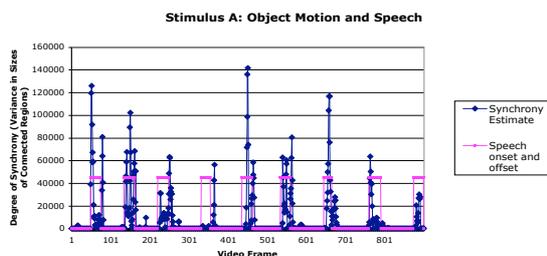


Figure 2: Stimulus A model estimates of quantitative degree of audio-visual synchrony. Synchrony estimates were computed using the connected region method. The speech onsets and offsets were obtained manually.

4. Stimulus B: Two Males Talking, Alternating Voice Source

Infant data and background. More difficult than detecting movement synchronous with single words (abrupt audio onsets and offsets) is detecting the synchrony between audio-visual stimuli with continuous speech and motion. We often are exposed to continuous sound and motion when observing someone speaking. Despite the potential perceptual processing difficulty, detecting audio-visual synchrony in this situation has obvious advantages. Infants’ abilities to localize sound are poor – sound sources must be at least 19 degrees apart for infants to notice the difference (Ashmead, Clifton, & Perris, 1987). Adults are more accurate and make extensive use of visual information in localizing talkers (Driver, 1996). If infants can spatially locate a speaker via visual information, they would have a powerful method to direct their attention past purely auditory means.

Dodd (1979) found that 10- to 16-week-old infants prefer to look at faces synchronized with speech as opposed to faces that do not match the audio. Infants 10 to 16 weeks of age watched a person in a sound-proof chamber reciting nursery rhymes. Every 60 seconds the audio played to the infant switched from being in

synchrony to 400ms out-of-synchrony (there was a 400ms audio delay). Infants (n=12) averaged 14.9% (2.9%-29.2%) inattention when speech was synchronous and 34.3% (1%-87%) inattention when out-of-synchrony. Thus, having the video 400ms out of synchrony incurred a 19.36% decrement in infants’ attention, implying that a face synchronized with the auditory stream helped infants direct their attention to the talker.

Model. Figure 3 shows the model processing results for Stimulus B, which consisted of two speakers talking alternately (top panel), one speaker on right (mid-panel), and one speaker on left (bottom panel). As shown in the top panel of Figure 3, the model did not successfully locate the horizontal position of the person talking when there were two motion sources. That is, the centroid position, averaged over the 5s intervals, was always on the left. The bias to the left occurred because the relative brightness of the left person was higher, and grayscale processing emphasized brightness in the visual component. The algorithm’s difficulty in discriminating amongst two motion sources, and correctly relating the sound to the motion source, may be due to the relative coarseness of the audio (or visual) features (i.e., RMS audio, grayscale pixels at 30fps). Better audio resolution (e.g., Mel-Frequency Cepstral Coefficients—MFCC’s), or better temporal visual resolution (e.g., 60 frames per second) might assist in this discrimination.

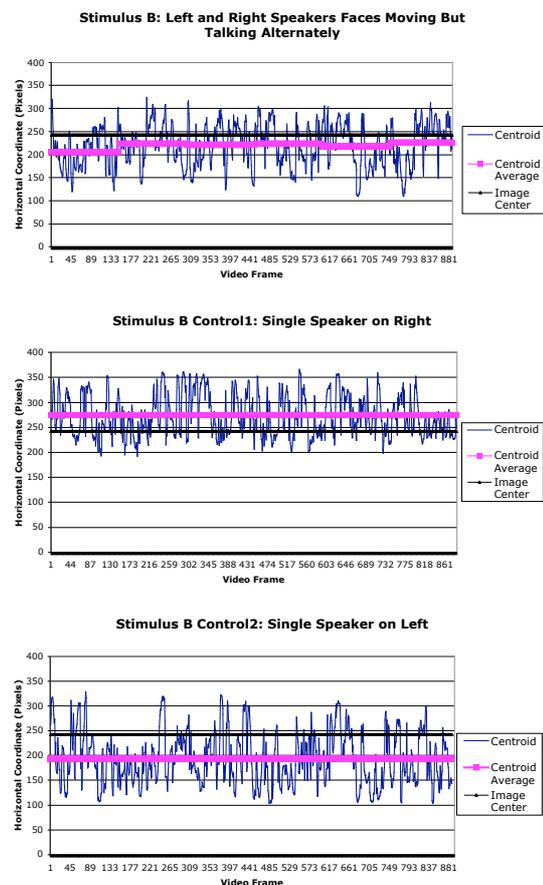


Figure 3: Stimulus B model results using mixelgram centroid to determine speaker location.

In both of the control data sets, the averaged centroid was positioned on the same side as the speaker (see lower two panels of Figure 3), as expected with detecting peak synchrony in the region of the speakers face. In these cases, the model did not have to deal with two motion sources, and was able to correctly relate the sound and motion. It is relevant to note that our Stimulus B task with two motion sources may be more difficult than the task posed by Dodd (1979) for infants, which had only one visual motion source and one audio source.

5. Stimulus C: Male, Female Voices and Face Visual

Thus far in our infant-model comparisons we have only considered tasks involving one audio source. Another relevant, and ecologically common, task involves audio source separation—for example being presented with two blended speech-audio sources and attending to only one of them. In the next two tasks, for infants, we are interested in the use of audio-visual synchrony for audio source separation. For the models, we focus on discrimination between the conditions on the basis of audio-visual synchrony.

Infant data and background. The task of separating one speech stream from another, using a visual motion source for assistance, is arguably more complex than detecting synchronization between speech and a face. Consider an infant sitting in a room with her family. Her mother might be speaking while her older sister is watching television and her two brothers are arguing nearby. In order to understand her mother, the infant must be able to separate her mother’s speech from that of the other voices in her environment.

Recent work by Hollich, Newman, and Jusczyk (in press) found that infants can use the visual synchronization between a talker’s face and the speech stream to help them focus on a particular speech stream and segment words from that stream in a noisy/blended stimulus. In these experiments, 7.5-month-old infants were familiarized with a visual display accompanied by an audio track of a blended stimulus (consisting of a female voice reciting a target passage and a distracting male voice). Importantly, the target audio (the female speaker) was the same average loudness as the distractor audio (the male speaker). This target stimulus was 5dB softer than the level at which infants had been shown to successfully segment speech streams using audio cues alone (Newman & Jusczyk, 1996).

As in the Gogate and Bahrck (1998) study, the type of visual familiarization differed across conditions. The first condition showed a synchronized video of the female speaker. The second and third conditions were controls that familiarized infants with a static picture of the female face or an asynchronous video of the female. While infants were expected to be unable to segment the speech in the control conditions, the controls insured that the effects seen were not the result of increased attention due to change in the visual stimulus or merely seeing a female face. Participants were 120 infants (30 in each visual condition) with a mean age of 7 months, 15 days (range: 7m 2d - 7m 28d).

Infants’ memory for target words presented during familiarization was tested using the Headturn Preference Procedure. Results indicated that only with synchronized video information did the infants succeed in this task. Infants looked reliably longer with the target words versus non-target words only when they saw a synchronized display ($t(29) = 4.39, p < .0001$). With a static display, or with asynchronous visual information infants did not show evidence of being able to complete the task ($t(29) = 1.16, p = n.s.$; $t(29) = 1.38, p = n.s.$). That is, they did not look longer with the target words versus the non-target words. Thus, 7.5-month-old infants can use synchronized auditory-visual correspondences to separate and segment two different streams of speech at signal-to-noise ratios lower than possible by merely auditory means. Infants did not succeed in this task if familiarized with a static or asynchronous video display of that speaker’s face, implying that it was specifically the synchronized video that produced this effect. These results suggest that infants gain a significant advantage by having synchronous visual information complement the auditory stream, especially in noise.

Model. Table 2 summarizes the results from the model synchrony analysis of the Stimulus C data (male, female voices and face visual), using the edge detection method to quantify synchrony. The mean for the Female-only condition differed from the Both condition and the Male-only condition at statistically significant levels ($p < 0.0000001$; two-tailed unpaired t-tests). The mean for the Male-only condition differed at a statistically significant level from the mean for the Both condition ($p < 0.009$; two-tailed unpaired t-test). These results are as expected—the Female-only condition should be the most synchronized, then the Both condition, followed by the Male-only (least synchronized). The HM algorithm, coupled with the edge detection method for quantitative assessment of synchrony, discriminated between these three conditions in the manner expected.

Stimulus		Female	Both	Male
C (Face)	Mean	26, 616.8	21, 010.6	19, 495.8
	Std Dev	16, 410.8	12, 139.8	13, 176.7
	Max	96, 302.4	74, 495.7	84, 144.0
	Min	3, 640.1	3, 189.2	3, 497.6
D (Oscilloscope)	Mean	8, 747.0	8, 200.8	8, 095.9
	Std Dev	3, 518.7	2, 945.9	3, 212.1
	Max	31, 107.0	29, 538.0	30, 015.5
	Min	2, 621.4	1, 219.5	1, 278.4

Table 2: Stimulus C and D model results. Table entries are averages of per mixelgram synchrony estimates from the edge detection method. *Female* is only the female voice. *Male* is only the male voice.

6. Stimulus D: Male, Female Voices and Oscilloscope Visual

Infant data and background. The results for Stimulus C, with male and female voices and the synchronized face of the female, demonstrate that infants *can* use synchronized visual information to help them segregate different streams of speech. The modeling results also demonstrate that the augmented HM algorithm makes an analogous discrimination. In retrospect, in a task using a dynamic-

face image, it should be clear that while infants may use knowledge specific to faces to perform their version of the task, the model did not use knowledge of faces. That is, while it is feasible to program face detection algorithms (e.g., Viola & Jones, 2001), our model did not incorporate such techniques.

Notice that while infants may have used knowledge of faces in the Stimulus C task, it is possible that their sensitivities to temporal synchrony are so strong that *any* synchronized visual stimulus would be sufficient to produce the benefit in related tasks. That is, perhaps infants' successful performance in the tasks reported above were not a result of their experience matching facial and vocal information, but were instead the result of a more general process of auditory-visual integration. A number of studies point to the idea that such integration in adults is not limited to feature-specific face information. For example, Rosenblum and Saldaña (1996) were able to get an improvement in phoneme recognition (over auditory alone) in adults by displaying point-light faces (in which one can only see the kinematics of movement).

For infants, too, auditory-visual integration has been shown for visual events other than faces. Some results in this regard were presented above in the section on Stimulus A, with object motion synchronized with vowels. Additionally, 4-month-old infants recognize the correspondence between the sight of a bouncing object and a sound (Spelke, 1976), and 6-month-old infants notice correspondences between a flashing picture and a synchronous pulsing sound (Lewkowicz, 1986). Indeed, according to Bahrick and Lickliter's (2000) "intersensory redundancy hypothesis," any redundant multi-modal information (also called amodal information) will attract significant infant attention. However, there has been no evidence to date in tasks involving continuous speech that infants will integrate the auditory speech signal with a visual signal other than a face. Continuous speech is a much more complicated acoustic event than are most of the signals tested in studies of infants' auditory-visual integration. Thus, skills in integrating a continuous speech signal with a visual stimulus may be the result of particular experience with auditory-visual correspondences.

The final infant experiment attempted to address this issue by changing the video familiarization to a moving oscilloscope pattern (Hollich, Newman, & Jusczyk, in press). The rationale was that the oscilloscope would preserve dynamic information while removing the visual shape of the face display, minimizing the chance that any effect seen would be the result of face-specific effects.

Participants were 27 infants with a mean age of 7 months, 10 days (range: 7m 1d - 7m 28d). The design, apparatus, and procedure were the same as in the previous infant experiment (Stimulus C). However, in the present experiment, a new display was created for the video familiarization. The oscilloscope waveform of the female passages across a 30ms running window was displayed on a computer monitor (using Harrier-Soft's Amadeus II software), video-recorded (via camcorder) and subsequently synchronized with the blended audio in the manner described in the first study (Stimulus C).

Importantly, this resulted in a video in which the oscilloscope display (a squiggly horizontal line) was synchronized only with the female voice. If amodal synchrony was partially responsible for the effects observed in the previous experiments, then the correlated motion of the oscilloscope would be expected to cue infants into the female talker's audio stream. If the effect in the previous experiment was the result of infants' particular experience with faces, however, they would be expected to fail on this task.

Infants listened significantly longer (1.40 seconds on average) to words that had occurred in the target passage than to words that had not, demonstrating successful segmentation of those words ($t(29) = 2.28, p < .05$). Thus, infants showed evidence of segmentation even when they were familiarized with a correlated oscilloscope pattern. In this manner, it appears that even the presence of such a correlated waveform pattern was sufficient to allow infants to succeed at this segmentation task. Without such visual information in this impoverished signal-to-noise ratio, infants would not be expected to succeed. This suggests that it was specifically infant sensitivity to amodal invariants that allowed them to correlate the patterns of visual change on the oscilloscope display with patterns of auditory change in the speech signal, and then to use this cue to help them separate that speech signal from other sound sources in their environment. Infant sensitivity to amodal invariants is enough to allow them to segment the speech stream in a noisy and often ambiguous acoustic environment.

Model. Table 2 summarizes the results from the model synchrony analysis of the Stimulus D data (male, female voices and oscilloscope visual), using the edge detection method to quantify mixelgrams. The results were similar to the model results for the Stimulus C data, except that the statistical comparison of condition Both to the Male-only condition was not significant ($p > .05$; two-tailed unpaired t-test). As in the Stimulus C modeling results, the Female-only condition differed from the Both and Male-only conditions ($p < 0.0006, p < 0.003$; two-tailed unpaired t-tests). The comparisons with the Female-only condition were as expected—this MPEG video, even with the oscilloscope visual motion, had the highest estimated degree of synchrony. Additionally, the numeric value of the Male-only condition was as expected—it indicated the lowest degree of estimated synchrony of all three.

Clearly the performance of the model was reduced from that observed with the face-visual data in the Stimulus C conditions. Using the oscilloscope visual representation resulted in a lack of a statistically significant difference between the Male-only condition and the condition in which both female and male voices were present. This may have occurred because there were smaller overall amounts of visual change (i.e., pixels), or perhaps because of the type and higher visual frequency of the changes. The oscilloscope changes were more discrete and rapid than the face motion. These differences may also be due to the technique we used for estimating synchrony. The edge detection method we used may detect edges better in the face case over the oscilloscope case.

7. Discussion

In this paper, we compared infant skills with a model of synchrony detection. Our goal was to assess the hypothesis: Can a single general-purpose synchrony detection mechanism, estimating audio-visual synchrony from low-level signal features, account for infant synchrony detection across audio-visual speech integration tasks of increasing complexity? The model we used was based on the Hershey and Movellan (2000; *HM*) algorithm, which we augmented with methods for estimating the degree of synchrony.

In this comparison we found some good results from the model and some notable exceptions. First, when faced with audio-visual stimuli comprised of a word spoken when an object was moved (Stimulus A), our model accurately generated estimates of synchrony (see Figure 2). Seven-month-old infants have been found to need such speech-object synchrony to learn speech-object relations (Gogate & Bahrick, 1998). Second, when faced with two motion sources, and one audio source (two people talking, and the audio source alternating), the model was unable to correctly indicate the location of the individual who was talking. In contrast to this, infants 10 to 16 weeks of age are better able to locate the sound of a person talking when the person's voice is synchronized with the person's facial motion. It should be noted, however, that in Dodd (1979), infants were only faced with a single audio and single motion source so their localization skills are still at issue. (Ongoing empirical studies are addressing this question.) Third, with one visual motion source and two speech audio sources (two people talking, but the dynamic face representation of only one voice), the model was well able, using a quantitative synchrony estimate, to distinguish between the three conditions—two voices, the voice of the person seen, and the noise (background voice). This parallels the results with infants—they perform better in learning words spoken by the person with a face-synchronized visual representation (Hollich, et al., in press). Fourth, in a variation of the previous work, the model was tested with a dynamic visual oscilloscope representation of one speaker's voice, still with two people talking. In this case the infants still learned the words from the oscilloscope-synchronized voice, but the model, while generally discriminating, had more difficulty.

In summary, we have shown that a model that directly estimates audio-visual synchrony from low-level features (i.e., RMS audio, and grayscale visual) across 0.5s intervals of time, but that does not utilize higher-level features or objects, detects audio-visual synchrony at levels in some cases similar to those of infants. We are not fully convinced that this model is a reasonable approximation of infant audio-visual synchrony detection, however, as there are cases in our own results where the model is not paralleling the infant results.

We see two main avenues for further exploration in modeling infant synchrony detection. First, there are a number of changes we can make to our present model based on the HM algorithm. For example, we can alter the kinds of audio and visual features used to compute the

synchrony relation. Using MFCC's as audio features, or using visual features based on pixel intensity changes (e.g., Nock et al., 2003) might improve the infant-model comparisons. Second, our models can start to take into account developmental changes in synchrony detection. For example, adults are more sensitive to audio-visual temporal asynchrony than infants (Lewkowicz, 1996). In addition, infants' audio-visual synchrony detection abilities get better with age and experience. For example, infants who are born deaf and given cochlear implants (CI) detect audio-visual synchrony better depending on the age at which the CI was initiated (Bergeson, et al., 2004). At present our synchrony detection model does not learn or develop. Some current algorithms for audio-visual synchrony detection incorporate training or learning (e.g., Slaney & Covell, 2001), and these methods deserve exploration in the present context. Other strategies include incorporating neuroscience findings (e.g., Calvert et al., 2000), and also performing synchrony detection in terms of higher-level features (e.g., faces). Potentially, once humans learn that faces (and other classes of objects) are distinctive parts of the environment, they may perform synchrony detection in terms of these higher-level entities and not just in terms of low-level sensory information (e.g., in the model, pixels and RMS audio). Just because infants can perform speech-based synchrony detection without using faces (e.g., Stimulus D above), doesn't mean that they don't use their knowledge about faces when faces are available.

Modeling the developmental trajectory for synchrony detection may enable closer approximations of the infant results, and this should also take us one step closer to utilizing synchrony detection methods in epigenetic robotic systems. An exciting robotic application for synchrony detection is *self-other* discrimination. Proprioceptive and visual inputs may be useful to help a robot in distinguishing between self-motion and motion in the world (e.g., Memon & Pollak, in progress). Learning or development integrated with synchrony detection in this context could enable robotic modeling of self-other issues such as infants' development from a preference for perfect contingency to a preference for imperfect contingency (e.g., Gergely & Watson, 1999; variations in this development also appear to relate to autism—Magyar & Gergely, 1998).

Acknowledgements

This research was funded in part by a donation to CGP from Digi-Key Corporation. We thank Kang James for her help with statistics, and Rocio Alba-Flores and Taek Kwon for their image processing advice. CGP thanks Roberta Golinkoff and Kathy Hirsh-Pasek and for their mentorship.

References

- Arsenio, A. & Fitzpatrick, P. (2003). Exploiting cross-modal rhythm for robot perception of objects. The *2nd International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, Singapore, December 15 - 18, 2003.
- Asada, M., MacDorman, K. F., Ishiguro, H., & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new

- paradigm for the design of humanoid robotics. *Robotics and Autonomous Systems*, 37, 185-193.
- Ashmead, D. H., Clifton, R. K., & Perris, E. E. (1987). Precision of auditory localization in human infants. *Developmental Psychology*, 23, 641-647.
- Bahrnick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, 79, 253-270.
- Bahrnick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36, 190-201.
- Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20, 394-419.
- Bergeson, T., Houston, D., & Pisoni, D. B. (2004). Audiovisual speech perception in normal-hearing infants and hearing-impaired infants with cochlear implants. Presented at *The 14th Biennial International Conference on Infant Studies*, Chicago, IL, USA, May 5 – 8, 2004.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10, 649-657.
- 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 mislocation of speech sounds due to lipreading. *Nature*, 381, 66-68.
- Fasel, I., Deak, G., Triesch, J., & Movellan, J. (2002). Combining embodied models and empirical research for understanding the development of shared attention. In: *Proceedings of the 2nd International Conference on Development and Learning* (pp. 21-27). Cambridge, MA: IEEE Computer Society Press.
- Gergely, G. & Watson, J. S. (1999). Early socio-emotional development: Contingency perception and the social-biofeedback model. In: P. Rochat (Ed.), *Early Social Cognition: Understanding Others in the First Months of Life* (pp. 101-136). Mahwah, NJ: Lawrence Erlbaum.
- Gogate, L. J. & Bahrnick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, 69, 133-149.
- Gogate, L. J., Bahrnick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71, 878-894.
- Gopnik, A. & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
- Helder, N. A. (2003). *A Real-time, Computational Model of Perceptually-based Contingent Behavior Detection*. Honors Project, University of Minnesota Duluth, Department of Computer Science.
- Hershey, J. & Movellan, J. (2000). Audio-vision: Using audio-visual synchrony to locate sounds. In: S. A. Solla, T. K. Leen, & K. -R. Müller (Eds.), *Advances in Neural Information Processing Systems 12* (pp. 813-819). Cambridge, MA: MIT Press.
- Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65 (3, Serial No. 262).
- Hollich, G. J., Newman, R. S., & Jusczyk, P. W. (in press). Infants' use of visual information to segment speech in noise. *Child Development*.
- Lewkowicz, D. J. (1986). Developmental changes in infants' sensory response to synchronous durations. *Infant Behavior and Development*, 9, 335-353.
- Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1094-1106.
- Lungarella, M., Metta G., Pfeifer, R., & Sandini G. (2003). Developmental robotics: A survey. *Connection Science*, 15, 151-190.
- Magyar, J. & Gergely, G. (1998). The obscure object of desire: "Nearly, but clearly not, like me": Perception of self-generated contingencies in normal and autistic children. Poster presented at the *International Conference on Infant Studies (ICIS)*, April 1998, Atlanta, Georgia, USA.
- Memon, N. & Pollak, T. (in progress). *Detecting Environmental Synchrony Using a Robotic Camera*. Undergraduate Research Opportunity Projects, University of Minnesota Duluth.
- Mislivec, E. J. (2004). *Audio-Visual Synchrony for Face Location and Segmentation*. Undergraduate Research Opportunity Project, University of Minnesota Duluth.
- Newman, R. S., & Jusczyk, P. W. (1996). The cocktail party effect in infants. *Perception & Psychophysics*, 58, 1145-1156.
- Nock, H. J., Iyengar, G., & Neti, C. (2003). Speaker localization using audio-visual synchrony: An empirical study. In: E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, & X. S. Zhou (Eds.), *Image and Video Retrieval, Second International Conference, CIVR 2003*. Lecture Notes in Computer Science 2728, Springer.
- Prince, C. G., Helder, N. A., Mislivec, E. J., Ang, B. J., Lim, M. S., & Hollich, G. J. (2003). Taking contingency seriously in sensory-based models of learning in infants. *Poster presented at the 2003 Meeting of the Cognitive Development Society*, held at Park City, UT, USA, October 24-25.
- Rochat, P. & Striano, T. (2000). Perceived self in infancy. *Infant Behavior & Development*, 23, 513-530.
- Rosenblum, L. D. & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 318-331.
- Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.
- Slaney, M. & Covell, M. (2001). FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In: *Proceedings of Neural Information Processing Society 13*. Cambridge, MA: MIT Press.
- Slater, A., Quinn, P. C., Brown, E., & Hayes, R. (1999). Intermodal perception at birth: Intersensory redundancy guides newborn infants' learning of arbitrary auditory-visual pairings. *Developmental Science*, 2, 333-338.
- Spelke, E. S. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, 8, 553-560.
- Viola, P. & Jones, M. J. (2001). *Robust Real-time Object Detection*. Compaq Cambridge Research Laboratory, Technical Report Series CRL 2001/01, February 2001.
- Vuppla, K. (2004). *Evaluation of Two Synchrony Detection Implementations*. Masters Thesis, University of Minnesota Duluth, Computer Science Department.
- Watson, J. S. (1972). Smiling, cooing, and "The Game." *Merrill-Palmer Quarterly*, 18, 323-339.
- Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291, 599-600.