

Running head: SYNCING MODELS WITH INFANTS

Syncing Models With Infants:
A Perceptual-Level Model of Infant Audiovisual Synchrony Detection

Christopher G. Prince

&

George J. Hollich

Authors' note:

G. J. Hollich is with Purdue University, Department of Psychological Sciences, West Lafayette, IN 47907 USA. C. G. Prince is with the University of Minnesota Duluth, Department of Computer Science, Duluth, MN 55812 USA. Contact: C. G. Prince, chris@cprince.com

Abstract

Synchrony detection between different sensory channels appears critically important for learning and cognitive development. In this paper we compare infant studies of audiovisual synchrony detection with a model of synchrony detection based on Gaussian mutual information (Hershey & Movellan, 2000), augmented with methods for quantitative synchrony estimation. Five infant-model comparisons are presented, using stimuli covering a broad range of audiovisual integration types. While infants and the model showed discrimination of each type of stimuli, the model was most successful with stimuli comprised of (a) synchronized punctuate motion and speech, (b) visually balanced left and right instances of the same person talking but speech synchronized with only one side, and (c) two speech audio sources and a dynamic-face motion source. More difficult for the model were stimuli conditions with (d) left and right instances of two different people talking but speech synchronized with only one side, and (e) two speech audio sources and more abstract visual dynamics—an oscilloscope instead of a face. As a first approximation, this model of synchrony detection using low-level sensory features (e.g., RMS audio, grayscale pixels) is a candidate for a mechanism used by infants in detecting audiovisual synchrony.

Introduction

Developmental research has illuminated many remarkable abilities of infants to integrate sensory information (e.g., the ability to visually track a person talking or to learn a new word); however the mechanisms underlying these abilities are still largely unknown. Presumably, the mechanisms will involve both higher-level cognitive abstractions and low-level perceptual skills. In this research, we are interested in knowing how far low-level perceptual mechanisms can take us in understanding infant behavior. Our logic is that unless we know the limitations of low-level mechanisms we can't know where cognitive skills start and perceptual skills end. Our approach is to utilize perceptual-level computational models and compare these models against infant behavior (Hollich, Mislivec, Helder, & Prince, 2004; Prince, Helder, Mislivec, Ang, Lim, & Hollich, 2003; Prince, Hollich, Helder, Mislivec, Reddy, Salunke, & Memon, 2004). We are focusing on infant synchrony detection skills because they arise early and have been hypothesized to underlie a range of infant behaviors. Our prediction is that a general-purpose mechanism will enable the model to match infants' performance across a series of synchrony-detection tasks. More specifically, we are evaluating the extent to which infant behavior can be accounted for by a general-purpose synchrony detection mechanism that relies solely on low-level audiovisual features.

At the most basic level, synchrony detection abilities involve recognizing temporally-based relations within or between sensory modalities. For example, we notice the simultaneous audiovisual changes when we watch a person speak. Such synchrony detection mechanisms appear critically important for infant learning and cognitive

development, and have been strongly implicated in developments ranging from word-learning (Gogate & Bahrick, 1998, 2001), to speaker localization (Pickens, Field, Nawrocki, Martinez, Soutullo, & Gonzalez, 1994), to segmenting speech in a noisy environment (Hollich, Jusczyk, & Newman, 2001). Indeed, Bahrick and colleagues (Bahrick, 2001; Bahrick & Lickliter, 2000) have suggested that audiovisual temporal synchrony is one of the most consistent and early relations to which infants are sensitive.

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 audiovisual representation are necessary to recreate the synchrony detection abilities of infants.

In addition to building more specific psychological theories of developmental mechanisms, we want to create practical robotic systems that utilize knowledge of how infants develop their skills (e.g., Weng, McClelland, Pentland, Sporns, Stockman, Sur, & Thelen, 2001). We suggest that understanding the mechanisms related to infants' developing synchrony detection skills can assist us in designing algorithmic mechanisms for robots. More specifically, we propose that synchrony detection mechanisms can act as a kind of integrative "glue" to connect different senses in the perceptual systems of robots. In computational neuroscience, Krichmar, Nitz, and Edelman (2004) have proposed that "dynamic synchronization of neural activity mediated by reentrant connections among many dispersed neural areas" (p. 5) can solve problems of within-

modality visual binding. Psychological theory has forwarded the idea that *amodal* representations (i.e., representations sharing information common across sensory and/or motor systems, such as time or space) act as a basis for human cognitive representation (Gopnik & Meltzoff, 1997). In epigenetic robotics, the need for directly detecting synchrony has only recently been considered (e.g., Lungarella, Metta, Pfeifer, & Sandini, 2003). In a robotic implementation, Arsenio and Fitzpatrick (2003; Fitzpatrick & Arsenio, 2004) detected rhythmic audiovisual synchrony relationships. They computed the distribution of durations between signal features (using histograms) within the ongoing audio and visual stimuli in order to measure periodicity. They then used comparisons of these distributions to detect audiovisual synchrony. Sensory (and motor) systems in epigenetic robots need to act in concert, and synchrony detection algorithms are one class of mechanisms that can afford this integration.

In this paper we compare algorithmic methods that directly compute audiovisual synchrony relationships between low-level audiovisual features (e.g., RMS audio and grayscale pixels) to empirical studies of infants' synchrony detection abilities. The algorithm we use is based on that of Hershey and Movellan (2000; *HM* algorithm in the following), which detects audiovisual synchrony, defined as Gaussian mutual information. 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 audiovisual synchrony when someone speaks is approximately the lips (see also Nock, Iyengar, & Neti, 2003, 2004). 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.

We use for comparison empirical studies of synchrony detection in infants that capture a full range of cases of audiovisual synchronization and utilize a set of video stimuli named *A*, *B1*, *B2*, *C*, and *D* (see Table 1). Our first infant-model comparison (using Stimulus A for the model) looks at the case of integrating the punctuate visual movements of an object and synchronous audio presentations of a word. Infant results in this case are from Gogate and Bahrick (1998), and they capture one the earliest types of detected synchrony. The second infant-model comparison section looks at the presumably more difficult case of integrating the continuous visual movements of a face with the speech stream. The focus here is on speaker localization, namely discriminating which of two people is actually generating the perceived speech-audio. In this case, the perceptual model was run on Stimuli B1 and B2 (two talkers, variable speech audio). Infant results from Pickens et al. (1994) are reviewed in this section, and new infant behavioral results are also reported (Hollich et al., 2004), based on Stimulus B2. The last two infant-model comparisons (Stimuli C and D) consider the task of audio source separation and are based on infant results from Hollich et al. (2001). More specifically, the third comparison looks at the task of separating out an irrelevant speech audio source using the continuous visual movements of a face. Stimulus C (one visual speaker, variable speech audio) was used with both the infants and the model in this case. The final comparison is based on an audio source separation task (Stimulus D) that may be even harder for both the infants and the model. With Stimulus D we substituted the continuous visual movements of an oscilloscope for the speech movements of a face.

The goal across these infant-model comparisons was to see how well a single low-level model of synchrony detection could account for the infant results in order to evaluate the plausibility of a single perceptual synchrony detection mechanism in infants. Because this was the first attempt to use the Hershey and Movellan (2000) algorithm for quantitative synchrony estimation, a secondary goal was to compare the effectiveness of several techniques for preparing the input and analyzing the output of the algorithm. The remainder of this paper comprises a description of the video stimuli and algorithm, a series of five infant-model comparisons, and a closing discussion.

Stimuli and Algorithm

We constructed digital video clips comprising a broad range of types of temporally synchronous audiovisual stimuli. This video data was processed using our synchrony detection program, *SenseStream* (Mislivec, 2004; see also Vuppla, 2004), based on the HM algorithm. In *SenseStream*, we measured synchrony using either a centroid method (from HM) for spatial localization, or a *connected region* method and an *edge detection* method for quantitative temporal synchrony estimation. While the HM algorithm generates topographic (qualitative) representations of synchrony (see below), it does not provide scalar estimates of the synchrony implicit in those representations. We developed the connected region and edge detection methods to quantify the topographic representations of synchrony generated by the HM algorithm.

Stimuli Construction

MPEG-1 digital video files were used as inputs for the model, with 29.97 visual frames per second and a 44.1 kHz audio sampling rate. Video was rendered into MPEG

format with the highest settings for data rate (Adobe Premier 6.5) to reduce compression. Stimulus A had one punctuate sound source (speech) and one punctuate visual motion source (object motion). Stimuli B1 and B2 were designed for speaker localization tasks, with people talking to the left and right and one or two audio sources. Stimulus C and D were designed for speech stream separation tasks, with one visual motion source and one or two audio sources. These stimuli are summarized in Table 1, and are described in more detail in the infant-model comparison sections. In all cases, the infants and the model received “mono” audio streams (not separated into left and right channels).

Algorithm

The SenseStream program (Mislivec, 2004) implements a modified form of Equation 3 from HM, which is given here as Part 1a of Equation 1.¹

$$M(x, y, t_k) = \frac{1}{2} \log_2 \frac{\overbrace{\left| \sum A(t_k) \right\| \sum V(x, y, t_k) \right|}^{(1a)}}{\underbrace{\left| \sum A, V(x, y, t_k) \right|}^{(1b)}} \left(1 - \frac{1}{2^{ra}}\right) \quad (1)$$

Part 1a of 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. t_k refers to the time associated with the visual and audio data, i.e., the k th frame of visual and associated audio data. The notations $A(t_k)$, $V(x, y, t_k)$, and $A, V(x, y, t_k)$ are interpreted respectively as the current audio data matrix ($S \times n$ in dimension; n is the number of audio features, e.g., $n=1$ for RMS audio), the current visual data matrix ($S \times m$ in dimension; m is the number of visual features, e.g., $m=3$ for RGB pixels), and the current audio data matrix

¹ In the context of Equation 1, the Σ symbol designates the covariance operation, and the notation $|X|$ designates the matrix determinant of X .

concatenated with the current visual data matrix ($S \times [n + m]$ in dimension). Each of these matrices has S rows: the first row is the oldest audio and/or visual sample (at time t_{k-s+1}) and the last row is the most recent audio and/or visual sample (at time t_k). Figure 1 illustrates the flow of processing in the HM algorithm. We extended Part 1a, the Gaussian mutual information from HM, with a further term, Part 1b, and our modified formula is given as Equation 1. In Part 1b, r was the maximum RMS audio value for the interval of audio data across which synchrony was analyzed (S), and $\alpha = 50$ was a fixed threshold. We included Part 1b in order to reduce the weight of smaller RMS values in the equation and to thus reduce sub-audible effects that were accidentally correlated with the visual stream. The HM method is sensitive to co-variations of any magnitude across the audio and visual streams, and we have observed low-magnitude RMS variation in our audio data (perhaps due to camera motor noise) causing perceptually relevant outputs (see below). An α of 50 was chosen because preliminary work indicated this level best counteracted the background noise for our video stimuli.

In the model processing described here, $S = 15$. We selected 15 frames as a likely value for S , because it roughly approximates the duration property with which temporal synchrony is perceived in humans (Lewkowicz, 1996)². With $S = 15$ and 30 frames per second video, Equation 1 relates a “column” of pixels (the pixels at the same x, y coordinate from 15 consecutive frames) and the audio source across 1/2 second of audiovisual data. Higher values of $M(x, y, t_k)$ from Equation 1 are interpreted as higher degrees of synchrony; lower values are interpreted as lower degrees of synchrony. The

² See Section 3, model sub-section, of this paper for the results of some variation in values of S .

$M(x, y, t_k)$ minimum is 0 (no synchrony)³. For visual features, we used *pixel intensity change* values or grayscale (0...255) values, and RGB (color) values to a lesser extent. Pixel intensity change (PIC) features are defined by Equation 2 (see also Butz & Thiran, 2002; Nock et al., 2003) where $F(x, y, t)$ is the grayscale pixel at location (x, y) in visual frame t (PIC features on edges were summed across fewer neighbors).

$$PIC(x, y, t) = \sum_{l, m=-1}^1 F(x + l, y + m, t + 1) - F(x + l, y + m, t - 1) \quad (2)$$

For audio, we relied on Root-Mean-Squared amplitude (RMS; one scalar per visual frame). As an example, the RMS audio values for Stimulus A (see Table 1) ranged between 0.002788 and 0.173405. For grayscale processing, Equation 1 was applied by our model to the pixels in the visual frames starting with visual frame S of a video clip. Equation 2 requires three consecutive frames to compute a PIC feature, and so for PIC processing, Equation 1 was applied to the pixels in the visual frames starting with visual frame $S+2$ of a clip, and ending one frame before the end of the video clip.

We refer to each $M(x, y, t_k)$ 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 qualitatively 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) roughly corresponds to the intervals of data in which the audiovisual signals are synchronized (Vuppla, 2004). Figure 2 gives an example of such a “perceptually relevant” mixelgram from processing the Stimulus A data. Of course, such qualitative estimates are of limited computational value. Quantitative

³ This results from the definition of mutual information, which is bounded below by 0.

methods for estimating synchrony are also important. For quantitative estimation, HM relied upon the centroid of the mixelgram to determine the (x, y) location of the peak of any synchrony existing between the audio and visual data.

For the current model, to quantitatively estimate the degree of audiovisual synchrony represented by the mixelgrams, we devised two additional methods to augment the HM algorithm, each of which operated as a function of mixelgrams and resulted in scalar estimates of synchrony per mixelgram. Our *connected region* method was based on the observation that in some cases of synchrony, mixelgrams have spatially-adjacent 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 (see Figure 2 for an example). We therefore computed the variance in the sizes of the connected regions per mixelgram. Nonzero mixels i and j are said to be *connected* when j is one of the eight-neighbors of i (edge mixels have fewer neighbors), and Equation 3 holds,

$$\max\left[\frac{M(i)}{M(j)}, \frac{M(j)}{M(i)}\right] \leq \text{Threshold} \quad (3)$$

where $M(\text{mixel})$ is the value of the *mixel*, and $\text{Threshold} = 1.125$. Again, the threshold level was an empirically determined constant, based on preliminary work. Connected regions are the spatial extent of pairs of mixels that are connected.

The other synchrony estimation method was based on *edge detection*. This method views the problem of quantitatively estimating synchrony from mixelgrams as a problem of enhancing the contrast of the mixelgrams (considered as images) and then estimating the amount of brightness in the contrast-enhanced mixelgrams. Looking at Figure 2, you can see marked contrast differences between the brighter and darker areas.

The brighter areas indicate audiovisual synchrony. The specific edge detection processing we used was inspired by the image processing literature and was a combination of Gaussian filtering followed by Sobel edge detection. Equation 4 depicts this process.

$$\sum_{i=1}^{h \cdot w} Sobel_{3 \times 3}(Gaussian_{15 \times 15}(M)) \quad (4)$$

Treating the mixelgram (M) as an image, we first blurred the mixelgram, reducing noise by convolution with a 15x15 Gaussian filter. Sobel edge detection was then applied. In order to estimate the amount of brightness in the resulting contrast-enhanced mixelgram, the resulting mixel values were summed.

Punctuate Object Motion and Speech

Infant Data and Background

One of the most basic cases of audiovisual integration concerns punctuate motion and audio – quick, visually circumscribed and acoustically very simple. Ecologically relevant examples of this kind of stimuli abound, from a hammer strike to a tree branch falling. Interestingly, young infants appear to use this kind of information in learning to link sight and sound – for example, to learn the link between word and object. Gogate and Bahrick (1998) found that 7-month-olds could learn the links between two speech sounds and two objects if the sound presentation occurred together with object movement. In that study, 48 infants were tested in one of three conditions: a synchronous movement condition ($n = 16$ infants), a static condition ($n = 16$), or an asynchronous condition ($n = 16$). In the synchronous condition, each infant saw a hand move one of two unfamiliar objects (a toy crab and a porcupine, or a lamb chop and a star), synchronous with the vowel “ahhh” (e.g., for the crab) or “eee” (e.g., for the porcupine).

Synchrony was defined by co-occurrence of the onset and offset of the motion and speech audio. That is, the audio and motion each started and stopped together. Each infant thus saw two objects, at separate times, each paired with its arbitrarily associated vowel sound. Vowel-object pairings were counterbalanced within the design. In the static condition the vowels were 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 object movements.

Infants were habituated under one of these conditions and then tested to see if they noticed when the vowel-object pairings were changed (as indicated by increased looking to the display when a pairing was “switched”). Only in the synchronized condition did infants look longer on the switched trials relative to 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-object pairing was changed. This is the opposite of what would be expected if the infants in the non-synchronous conditions would have learned the vowel-object pairings. Only 7 total out of the 32 infants in these non-synchronous conditions showed evidence of having noticed the switch in vowel-object pairings. Thus, it appears that 7-month-olds can use punctuate audiovisual synchrony to help them learn the link between words and objects. A follow-up study (Gogate, Bahrick, & Watson, 2000) indicated that mothers who spontaneously moved objects in synchrony in object naming situations had children with larger vocabularies.

We are not suggesting that this is all it takes to learn a word. Word learning is an extremely complicated task involving multiple cues (Hollich, Hirsh-Pasek, & Golinkoff, 2000), and including numerous social-pragmatic factors (Baldwin, 1993; Bloom, 2000). Indeed, one of our long-term goals is to incorporate these multiple factors in a model of word learning (for an AV-synchrony-only model: Prince & Mislivec, 2004). However, 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 kind of 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 contained utterances of the word “modi” co-occurring with vertical object motion (see also Table 1). 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. Both our Stimulus A and the synchronous condition stimulus of Gogate and Bahrick (1998) involved speech audio (the word “modi” and vowels respectively), and speech utterances co-occurred with object movement in both our Stimulus A and that of Gogate and Bahrick (1998). The Stimulus A clip duration was 30s. Only one word-object pairing was used in this analysis, as compared to the two pairings used by Gogate and Bahrick (1998), because the task was different for the model. The modeling goal was to assess discrimination between audiovisual synchrony vs. non-synchrony intervals in the video clip and so a single word-object pairing was sufficient.

Both the connected region and edge detection methods were used to generate quantitative synchrony estimates for this data. In order to determine which of these two methods performed better, we used Pearson's correlation to compare the synchrony estimates (smoothed with a running average over a window of 11 points; see Figure 3) to the word onset/offset times. The word onset/offset times for Stimulus A were obtained manually using the Audacity sound editing program (<http://audacity.sourceforge.net>) on the audio track of the video data. The edge detection method synchrony estimates were more highly correlated with the word onset/offset times, $r = 0.719$, $t(872) = 30.7$, $p < .001$, than the connected region synchrony estimates, $r = 0.538$, $t(872) = 18.9$, $p < .001$. Figure 3 presents the synchrony estimates generated by the edge detection method. These results were obtained with pixel intensity change visual features; the grayscale visual feature analysis generated similar correlations, $r = .698$ and $r = .582$ for the edge detection and connected region methods. The model tracked the synchrony well: Periods of model-estimated high synchrony were well-correlated with periods of manually determined audio onset and offset for utterances of the word "modi." Audiovisual synchrony in these cases results from the word being uttered at the same time as the object is moved. In summary, the audiovisual synchrony inherent in punctuate object motion and speech is well-detected by the model using edge detection synchrony estimation and pixel intensity change visual features.

To illustrate the effects of our choice of processing window length in the HM algorithm ($S = 15$ or $1/2$ s with 30 frames per second video), we ran the synchrony analysis (edge detection estimates, pixel intensity change visual features) with two neighboring values of S . With $S = 14$, the synchrony estimates had a slightly lower

correlation with the audio onsets/offsets, $r = 0.698$, and with $S = 16$, they had a slightly higher correlation, $r = 0.730$. Small changes in the model parameter (i.e., S) resulted in only small changes in correlations of the synchrony estimates with the manually determined onsets/offsets of the words. Thus, the model in this case robust to small changes in S .

Speaker Localization

Infant Data and Background

More difficult than detecting movement synchronous with punctuate single words (abrupt audio onsets and offsets) is detecting the synchrony between audiovisual stimuli with *continuous* speech and motion. Further, we often are exposed to multiple sources of sound and motion when observing someone speaking. For example, a television may be on in the background at the same time as the caregiver is speaking to the infant. Despite the potential perceptual processing difficulty for an infant, detecting audiovisual synchrony in this situation has advantages. Audiovisual synchrony detection can assist in localizing a sound source. Infants' abilities to localize sound are poor when they have only audio information. Sound sources must be at least 19 degrees apart for 6.5-month-olds to notice the difference (Ashmead, Clifton, & Perris, 1987). While adults are more accurate in audio-only situations, they also can make extensive use of visual information in localizing talkers (Driver, 1996). If infants can likewise spatially locate a talker by integrating audio and visual information, they would have a powerful method to direct their attention past purely auditory strategies.

Dodd (1979) found that 10- to 16-week-old infants prefer to look at faces synchronized with speech as opposed to faces that are not synchronized with the audio. However, the method used by Dodd (1979) did not involve two visual motion sources and so was not directly comparable to a talker localization task. In more directly relevant work, Pickens et al. (1994) examined 3-, 5-, and 7-month-old infants ($N = 77$) using videos of two different female talkers side-by-side reciting speech passages in infant directed speech. The audio matching the talkers came from a central loudspeaker while the sound track alternated in matching the left or right talker over four trials (30 seconds each). Pickens found statistically higher rates of looking to sound-matching videos for 3- and 7-month olds, but not 5-month-olds.

Model

The video for this model (Stimulus B1) approximated the stimuli of Pickens et al. (1994) and had one sound source at any one time but two visual motion sources. In this clip, two adult males, fix-positioned on the sides of a split screen, were talking for 30s. For the first 5s, the speech audio was from the right male, the next 5s from the left male, and the next 5s the audio was background noise. The remaining 15s repeated this right, left, noise pattern with different video data. Two additional 30s clips, controls for the model, were also used. Control 1 had the prior video from 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 prior video from the left person.

Figure 4 shows the model processing results for Stimulus B1. The top panel shows two speakers talking alternately; the middle panel shows one speaker on the right,

while the bottom panel shows one speaker on the left. As shown in the top panel of Figure 4, the model did not successfully locate the horizontal position of the person talking in this stimulus when there were two motion sources. That is, the centroid position, averaged over the 5s intervals, was usually on the left (lower valued pixel coordinates), except for the last 5s, which was on the right (when the audio was only background noise). While there was some difference in the average centroids between the first two 5s intervals (see Figure 4, top panel), the difference was not in the expected direction. That is, the averaged centroid in first 5s was more towards the left, not the right as expected. 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 4), 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.

Subsequent examination of the model revealed that the difficulty with Stimulus B1 was primarily confined to salience differences between the left and right video stimuli. That is, the bias to the left occurred because the person on the left had more, and more variable, brightness features, and grayscale processing emphasized brightness in the visual component. Processing with RGB visual features instead of grayscale reversed this bias, as shown in Figure 5. The average centroids are now positioned on the right. The person on the right had more color features, resulting in centroids with a right bias. Processing with pixel intensity change visual features also generated centroids with a generally left-bias, not markedly different from Figure 4, top panel. Thus, the algorithm's difficulty in discriminating amongst two talkers, and correctly relating the speech-audio to the talker, was due to the low-level character of the audio and visual features used

(e.g., RMS audio, grayscale pixels at 30fps). Additionally, because infants were never exposed to these stimuli, it is possible that infants may have had similar biases. Infants are notorious for being affected by salience factors (e.g., Hollich et al., 2000).

Stimuli that were balanced for salience would likely result in better model performance. In addition, better audio resolution (e.g., Mel-Frequency Cepstral Coefficients—MFCC's), or better temporal visual resolution (e.g., 60 frames per second) might also assist in this discrimination.

Speaker Localization Revisited: Balanced Visual Salience

The model in the previous comparison did not show localization of the speaker when there were two motion sources, apparently due to left-right side variations in the salience of visual stimuli and characteristics of the model. Unfortunately, because infants were never tested on Stimulus B1, we have no means of comparison to know whether the model would have matched infant performance. In addition, both the infants (in Pickens et al, 1994) and the model were tested with a single audio source condition; it is possible that infants would be more likely to attend to synchrony if both the audio and the video provide multiple sources of variation (Hollich et al., 2004). That is, with multiple audio sources, infants may depend more strongly on synchrony information to help them attend to a particular speech stream. In noisy conditions, infants have to look at the visual display to figure out what is being said; in the clear, the need to watch the synchronous video is reduced.

To further assess speaker localization with the model (see also Nock et al., 2003), and to directly compare the model with infant performance, we created a new set of video

clips and ran these clips with both infants and the model. The new clips had a female speaking different passages on the left and right sides of the screen, but the audio source corresponded to her talking from only one of the sides. To correct for the salience imbalance in the previous comparison, every effort was made to ensure the videos were as similar as possible. The female speaker was video recorded on the same day, from the same distance, wearing the same clothes, and with the same lighting and audio conditions. Two clips, videoed without specific background noise, were called “Cup” and “Dog” because the fifteen-second passages spoken using child-directed speech in these clips emphasized those words. The Cup visual display was always on the left while the Dog visual display was always on the right. These clips provided a comparison and replication of the original Pickens et al. (1994) experiment with the exception that we used the same person on the left and right. Two additional clips, “CupNoise” and “DogNoise,” which blended in a background male speaker (with no corresponding visual motion source) with the respective source audio, were included as a test of the robustness of the model and infant performance. Presumably, sound source localization tasks performed by infants often take place in the audio context of other talkers, and thus such a manipulation not only makes the sound source localization task harder but more ecologically valid.

Infant Data and Background

Infants in this study were recruited from the West Lafayette community via advertisement and they ranged in age from 7.5 months to 8.5 months, with an average age of 8 months ($N = 20$). All infants were healthy and full-term and were only exposed to English in their caregiving environments. Infants of age 8 months were chosen because

this is approximately the age at which infants first demonstrate an ability to segment the speech stream (Jusczyk & Aslin, 1995), and this age was similar to the age of the infants in two of the studies reviewed in the previous sections (Gogate & Bahrick, 1998; Pickens et al., 1994). Infants were brought into the lab waiting room where they were acclimatized to the environment while parents signed a consent form. The infants were then brought into the testing room where infants sat on the lap of their parent, approximately 91 cm away from a 142 cm diagonal video projection (created by an InFocus L50 projector). The parents were then instructed to close their eyes in order to avoid influencing the child's looking. The video clips were played through in their entirety, each immediately after the other, while a digital camcorder hidden underneath the projection display recorded infant looking preferences at 29.97 frames per second. Each infant viewed all four video clips (Cup, CupNoise, Dog, and DogNoise); however, the order of presentation was counterbalanced across different infants. Because the total video stimulus was one minute long (15 seconds x 4), there was very little drop-off in infant attention, and only four infants failed to complete the study (and were replaced). Videos of the infants were subsequently transferred to a computer and analyzed in a frame-by-frame manner by coders blind to the condition being run. The coders transcribed all left and right looks by the infants. The coding was done on a frame-by-frame basis and thus any look within 1/30th of a second was coded. This resulted in a procedure that was highly reliable across coders. A random 25% of the video data was recoded by a second rater and indicated 99.8% reliability.

Table 2 presents the amount of time the infants, on average, looked to the left and right sides, and also the proportion of looks to the left side, all while viewing the

individual video clips. Our analysis here is in terms of the difference between the infant looking in the Cup versus Dog (and CupNoise versus DogNoise). While the amounts of looking at the left or right are relevant (and reported in Table 2), the difference in proportion of looking quantifies the discrimination of synchrony. The results from the clear conditions indicate a trend toward looking longer at the appropriate speaker, looking on average 6.59% more at the targeted side (see Table 2), $t(19) = 1.67, p = .11$. This percentage is almost identical to the results found by Pickens et al. (1994) for 7-month-olds. In the noise conditions (CupNoise and DogNoise), the infants' looking was also in the expected direction, but to a greater extent (10.27%), as predicted, $t(19) = 2.32, p = .03$. These results suggest that infants can locate a speaker using audio synchronization, and that this ability is easier to demonstrate in noise. It is worth noting that these noise results show a 4% improvement in speaker tracking over that found in Pickens et al. (1994) and our replication. This does not necessarily suggest that infants are any better at tracking synchrony in the noise condition, just that this more ecologically valid task taps their synchrony detection abilities more strongly. Indeed, for this reason, the remainder of the infant studies reviewed consider synchrony detection in multi-talker conditions.

Model

Table 2 also presents the results of running the model on the Stimulus B2 video clips. Results are presented from an analysis using grayscale visual features and also using pixel intensity change (PIC) visual features. If the model was successfully localizing the talker for the “Cup” clips, we should expect X-coordinate centroids on the left of the center position (360 pixels; the video clips were 720 pixels wide and visual

frames were not scaled). That is, we should expect X-coordinate centroids less than 360 pixels for the “Cup” clips. For the “Dog” clips, we should expect centroids greater than 360 pixels. Table 2 shows that the average centroids for both the “Cup” and “CupNoise” video clips were left of the center position, as expected, with both the grayscale and PIC analysis. Both of these values were also statistically significantly different from the center position (two-tailed t-tests; see Table 3 for p values). To provide better comparability with the infant results, we also analyzed the centroids generated per mixelgram by the model for the proportion of centroids to the left and right sides. If a particular centroid for a frame was ≤ 360 , then this was considered “attending” to the left, otherwise, this was considered “attending” to the right. The numbers of left and right frames attended from this analysis are reported as a proportion of N (the total number of mixelgrams for the clip). Clearly, as well as having average centroids on the left, the “Cup” and “CupNoise” videos both have the majority of centroids positioned on the left (see Table 2). We assessed the left/right frequencies for statistical significance using a cumulative binomial, with a chance probability of 0.5 of left vs. right centroid positioning. The left positioned centroids for “Cup” and “CupNoise” were statistically significant (see Table 2).

The results for the “Dog” and “DogNoise” clips were more variable across the grayscale and PIC analysis. For the “Dog” clip, we expected average centroids positioned on the right (≥ 361 pixels), but for the grayscale analysis, the average centroid was positioned on the left and the left vs. right proportion provided a left positioning as well—and further—both of these results were statistically significant. However, for the PIC analysis, the results were as expected—both the average centroid and the left vs.

right proportion was positioned towards the right speaker, and again both of these were statistically significant. The grayscale visual analysis for the “DogNoise” clip generated results as expected—both the average centroid and the left vs. right proportion were positioned on the right, at statistically significant levels. With the PIC analysis, while the average centroid was as expected (on the right), and statistically significant, the left vs. right proportion for the “DogNoise” clip was in the expected direction (right), but was not statistically significant different from random positioning.

Interestingly, this parallels the results from the infant data. As with the infants, it is important to compare the model performance on Cup to Dog, and CupNoise to DogNoise. This comparison quantifies the discrimination of synchrony; any other bias to a side could still be due to salience issues. For both the grayscale and PIC analysis, these comparisons were statistically significant based on two-tailed unpaired t-tests: Cup compared to Dog/grayscale, $p < .05$, $t(858) = 2.05$; CupNoise compared to DogNoise/grayscale, $p < .0001$, $t(859) = 5.08$; Cup compared to Dog/PIC, $p < .001$, $t(857) = 3.33$; CupNoise compared to DogNoise/PIC, $p < .05$, $t(855) = 2.51$. It is further interesting to note that the differences in proportion of looking to the left (in the “cup” direction) is more similar to the infants’ behavior for the grayscale analysis (6.8% for Cup-Dog vs. 17.1% for CupNoise-DogNoise) than for the PIC analysis (11.3% for Cup-Dog vs. 8.6% for CupNoise-DogNoise).

In summary, the PIC (pixel intensity change) analysis generated the most consistent results, with respect to “looking” to the left and the right. The average centroid was on the same side (left for Cup/CupNoise and right for Dog/DogNoise) as expected, and these averages were also statistically significantly different from the center position.

The left vs. right proportions were also in the expected directions (as above), and three out of four of these were statistically significant. The PIC analysis correctly localized the side of the speaker, even in noisy conditions. However, the grayscale analysis generated results that seem the most comparable to the infant behavior. The looking to the “Dog” stimuli by the infants and the model in this case was to the left, and both infants and the model with grayscale analysis discriminated better in the noise condition (CupNoise vs. DogNoise) than in the clear (Cup vs. Dog).

Audio Source Separation: Face Visual

The previous infant-model comparisons considered the task of speaker localization. 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 a sense this was tested to a degree in using the CupNoise and DogNoise stimuli of the last section, where distracting audio (a male voice) was blended with the speakers voice. However, just because infants looked at the matching video does not necessarily mean that they separated the two speech audio streams. In the next two tasks, for infants, we focus on the use of audiovisual synchrony for audio source separation. For the models, we focus on discrimination between the noise and clear conditions on the basis of audiovisual 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. This may involve not only localization, to look at her mother, but also separation of her voice from the other sounds, and identifying words and phrases in the mother's speech stream. In the speaker localization study reported above infants detected synchrony in noisy conditions; however, from that study it is not apparent whether infants are able to make use of that synchrony in analyzing the audio stream.

In work by Hollich, et al. (2001), they 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 signal to noise ratio was 5dB softer than the ratio at which infants had been shown to successfully segment speech streams using audio cues alone (Newman & Jusczyk, 1996).

As in the Gogate and Bahrick (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 in these control conditions were expected to be unable to segment the speech, the controls

ensured 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 90 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 (in comparison to words not presented during familiarization) was tested using the Headturn Preference Procedure. In this procedure, the infant sat on the caregiver's lap in the center of a three-sided enclosure, while a hidden observer coded infant headturns using a button box attached to a computer. A trial began with the flashing of a green light on the center panel of the enclosure. When the infant fixated on the green light, the experimenter pushed the button to begin the trial; the green light was turned off, and a red light on one of the side panels began to flash. When the infants' head turned at least 70 degrees towards the flashing light, the experimenter initiated the speech sample from a loudspeaker under that light (e.g., "cup, cup, cup"). If the infant turned away from the light for a period of at least two seconds, the computer would end the trial, and the green center light would begin to flash, signaling the beginning of a new trial. Information about the direction and duration of the head-turns and the total trial duration were stored in a data file on the computer. Any time the infant spent looking away was not included when measuring the total looking time. Both the experimenter and caregiver wore sound-insulated headphones that played continuous masking music to prevent them from hearing the stimulus materials throughout the duration of the experiment.

Results indicated that only with synchronized video information did the infants succeed in this task. Infants looked reliably longer (an average of 1.98 seconds) when the target words were played (as opposed to non-target words) only when familiarized with a

synchronized display, $t(29) = 4.39, p < .0001$. With the static display condition (an average of 0.4 seconds to the target display), or with the asynchronous visual condition (an average of 0.5 seconds to the non-targeted stimuli) infants did not show evidence of succeeding on the task, $t(29) = 1.16, p = \text{n.s.}; t(29) = 1.38, p = \text{n.s.}$ That is, they did not look longer when the target words were played than when the non-target words were played. 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

In order to enable a comparison of degree of synchronization for the model, the video stimuli from the infant studies were extended by adding two control video clips. In addition to the clip with both the male and female voice (Both condition), we used clips with only the male voice (Male-only condition), and only the female voice (Female-only condition). The motion visual component of these two additional clips was the same as in the Both condition, and we used the same voice audio that was originally blended in the Both condition clip.

Table 3 summarizes the results from the model synchrony analysis of the Stimulus C data using grayscale visual features and 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, $t(1312) \geq 2.92$, $p < 0.005$; 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, $t(1312) = 3.61$, $p < 0.0005$; two-tailed unpaired t-test. These results are as expected—the Female-only condition was the most synchronized, followed by the Both condition, and then followed by the Male-only condition (least synchronized). The connected region method was also used to analyze the Stimulus C data, but the mean synchrony estimates (mean variances in sizes of connected regions) in this case did not show significant differences in all cases (the Female-Both condition comparison was not statistically significant). Analysis with pixel intensity change visual features was also carried out for both of the synchrony estimation methods. This improved the connected region method results (all three comparisons were in the expected direction, and were statistically significant), but performance on the edge detection method suffered (only two comparisons were now statistically significant).

In summary, the HM algorithm, using RMS audio and grayscale visual features, and coupled with the edge detection method for quantitative assessment of synchrony, discriminated between these three types of audiovisual synchrony in the manner expected. The video with only the female speaking, which matched the dynamic image of the visual information, was found to have the highest degree of synchrony, followed by the blended Male and Female, followed last by the video with only the sound of the male voice, which did not match the visual information.

Audio Source Separation: 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 was not a result of their experience matching facial and vocal information, but was 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 an 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 reviewed here attempted to address this issue by changing the video familiarization to a moving oscilloscope pattern (Hollich, et al., 2001). 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. If infants succeeded in this task, then this could be taken as evidence that any synchronized information would be useful.

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 (1.43 seconds on average) to words that had occurred in the target passage than to words that had not, demonstrating successful segmentation of those words, $t(29) = 2.28$, $p < .05$. Compared with the control conditions from the previous studies, 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

As with Stimulus C, in order to enable a comparison of degree of synchronization for the model, Stimulus D from the infant studies was extended by adding two control video clips. In addition to the clip with both the male and female voice (Both condition), we used clips with only the male voice (Male-only condition), and only the female voice (Female-only condition). The motion visual component of these two additional clips was the same as in the Both condition for Stimulus D, and we used the same voice audio that was originally blended in this Both condition clip.

Table 3 summarizes the results from the model synchrony analysis of the Stimulus D data, using grayscale visual feature analysis, and the edge detection method to quantify mixelgrams. The results differed from the Stimulus C data in that the mean Female-only synchrony estimate was numerically less than the mean Male-only estimate, the opposite direction to that expected for this pair. The statistical comparison of the Both condition to the Male-only condition was statistically significant, $p < .0005$, $t(1290) = 4.40$; two-tailed unpaired t-test, but the comparison of the Female-only to the Male-only condition was not, $p > .5$, $t(1291) = .266$, unpaired t-test. The direction of the difference of the Both condition as compared to the Male-only was as expected—this Both condition video, even with the oscilloscope visual motion, had the higher 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 video clips. The connected region method was also used to quantify the mixelgrams. In this case, none of the comparisons of the means of the synchrony estimates were significantly different at a 0.05 level. Using pixel intensity change (PIC) visual visual features did not

add to the analysis. For the connected region method, using PIC, none of the comparisons were statistically significant; for the edge detection method, two of the comparisons were statistically significant, but none were in the expected direction.

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 model synchrony estimate for one of the three videos (with only the female voice) that was less than expected. 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. Another possibility is that the model as constructed does not attend visually to only the oscilloscope motion. Perhaps if the visual features used in the model comprised only the oscilloscope motion the model would achieve synchrony estimates more similar to the performance of the infants.

Discussion

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

In this comparison we found some accurate results from the model and some notable exceptions. First, when faced with audiovisual stimuli comprised of punctuate

object motion and speech (Stimulus A: a word spoken when an object was moved), our model accurately generated estimates of synchrony. The correlation between the model synchrony estimate and the manually determined word onsets/offsets was: $r = 0.719$, $p < .001$ (see also Figure 3). Seven-month-old infants have been found to need such speech-object synchrony to learn vowel-object relations (Gogate & Bahrick, 1998, 2001).

Second, when faced with two different people talking, and the speech source alternating, the model was unable to correctly indicate the location of the individual who was talking (Stimulus B1; see also Figures 4 and 5). However, with two instances of the same person talking (i.e., controlled illumination and audio levels) the model, using pixel intensity change visual features, was able to determine the correct left vs. right side even under noisy conditions. In this video (Stimulus B2), two instances of the same talker, positioned on the left and right, were making face-speech movements but the audio was from only one of them. The model analysis with grayscale visual features was particularly similar to our findings with infants on the same stimuli (see Table 2). Infants eight months of age were able to localize the image of a person talking when the speech-audio was synchronized with the person's facial motion and, like the model with grayscale visual features, did particularly well in noisy conditions. Third, with one visual motion source and two speech audio sources (two people talking, but the dynamic face synchronized with only one voice), the model was well able, using a quantitative synchrony estimate, to distinguish in the expected manner between three conditions—the voice of the person seen, two voices, and the noise (see Table 3). 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., 2001). Fourth, in a variation of the previous comparison,

infants (Hollich et al., 2001) and the model were tested with a dynamic visual oscilloscope representation of one speakers' voice, with again one or two voices. In this case the infants still learned the words from the oscilloscope-synchronized voice, but the model had more difficulty (see Table 3). In these model analyses, the edge detection method for synchrony estimation was uniformly more accurate than the connected region method in modeling the infant results. Additionally, the use of pixel intensity change visual features somewhat more often resulted in a closer correspondences to the infant results than use of grayscale visual features.

In summary, we have shown that a model that directly estimates audiovisual synchrony from low-level features (i.e., RMS audio features, and grayscale or pixel intensity change visual features), across 0.5s intervals of time, detects audiovisual synchrony at levels similar to those of infants. While this model is in the beginning stages, it already accounts for a substantial portion of the variance associated with infant synchrony detection. Indeed, under some views of modeling this may be enough. That is, if one views modeling “as a tool to summarize experimental data provided by different approaches” (p. 772, Poggio & Bizzi, 2004), it might be argued that the goal of modeling is not to achieve “completeness” (e.g., see the strong equivalence of Pylyshyn, 1984). Our model could be just that—a basic abstraction of the system we are seeking to describe or explain. Nonetheless, there is still variation in infant’s synchrony detection behavior that we want to explain.

We see three main avenues for further exploration in modeling infant synchrony detection. First, we can extend the synchrony detection algorithm in ways similar to that carried out by HM. One issue along these lines is regarding the scalability and

neurological plausibility of the model; if the model must calculate correlations across a time window for all low-level features across several sensory modalities the computational load incurred may be impractical. A more realistic alternative might be to expect that further within-modality processing (e.g., dimensionality reduction and feature detection across longer time intervals and/or larger spatial regions) will occur prior to the process of synchrony detection. Thus, we can alter this model by changing the kinds of audio features used to compute the synchrony relations (e.g., encoding and computing changes in frequency), or by changing the kinds of visual features (e.g., orientation and speed of movement). If within modality processing takes some computational load away from algorithmically relating across the modalities (i.e., synchrony detection), this should make it more feasible to add other sensory modalities into the computation. The present algorithm is also restricted in that it may limit the synchrony detection to relatively short time intervals. Others have suggested that synchrony detection occurring over long time intervals could ultimately be linked to naïve theory understanding (Gopnik & Meltzoff, 1997). While at this stage of development in the model it is too early to test this hypothesis, examination of this idea is a long-term goal.

A second avenue is alternative algorithms to HM. For example, another approach to quantitative audiovisual synchrony estimation, based on canonical correlation across the audio and visual signals, has been forwarded by Slaney and Covell (2001). As well, as reviewed above, Arsenio and Fitzpatrick (2003; Fitzpatrick & Arsenio, 2004) have developed an algorithm specialized to detect rhythmic audio-visual synchrony relationships. Particularly promising in light of the adaptive long-term goals of epigenetic

robotics, are approaches based on learning algorithms. For example, de Sa and Ballard (1998) designed a multisensory algorithm that separately classifies audio and visual sensory information using neural networks, and then feeds those classifications back to the other modality, minimizing the disagreement between the classifiers to improve the performance of classification. This method relies on the intuition that each modality operates somewhat independently, classifying audio or visual information. However, each modality also receives inputs of the classification results from the other modality. This “self-supervised” method does not rely on externally provided category labels. More recently, Fisher, Darrell, Freeman, and Viola (2001) proposed an adaptive method for audio-visual synchrony detection. They used a pair of single-layer perceptrons (neural networks), one for audio and one for visual, which were trained by maximizing mutual information between the outputs of the visual-perceptron and the audio-perceptron. This method shows potential for incrementally adapting to environmental audio-visual synchrony dynamics.

These learning approaches foreshadow our third avenue for future exploration: Incorporating developmental changes into the model. Based on the available empirical evidence at the time, Lewkowicz (2000) proposed that a succession of four infant skills develop regarding responsiveness to intersensory temporal relations: synchrony, duration/synchrony, rate/synchrony, and rhythm. Additionally, each of these perceptual skills likely exhibits further, ongoing development. For example, adults are more sensitive to disruptions in audiovisual temporal asynchrony than infants (Lewkowicz, 1996). Another example of the experience-dependent character of infants’ audiovisual synchrony detection comes from infants who are born deaf and given cochlear implants

(CI). These infants detect audiovisual synchrony better depending on the age at which the CI was initiated (Bergeson, Houston, & Pisoni, 2004)—the younger the better. At present our synchrony detection model does not learn or develop. 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 adaptive synchrony detection methods in epigenetic robotics. As synchrony detection assists infants in learning relations between audio and visual stimuli, sound source localization, and source separation; a similar synchrony detection mechanism that employs some form of learning should also assist robots to these ends.

Fisher, Darrel, Freeman and Viola (2001) have also suggested that learning methods are a suitable way to deal with possible violations of statistical assumptions in the HM method. More specifically, the HM method assumes Gaussian distributions for the matrices (with S rows) of the audio and visual features used in its processing. The issue for this assumption is regarding short time-intervals (of length S) of the audio and visual features, and not these features over the full time span of the video clips. To quantitatively assess for potential violations of the Gaussian assumption, we sampled 10 randomly selected grayscale pixels and the RMS audio, over the entire duration of a video clip (Stimulus B2, Cup: $N = 448$ frames). We then analyzed length 15 running windows (i.e., $S = 15$; the same data vectors used by SenseStream processing) of each of the pixels and the RMS audio using an Andersen-Darling test for normality. Only 28.9% of 4340 tests on the grayscale pixel intervals showed normality, $p < .05$. About the same percentage of RMS audio intervals showed normality (30.0% of 434 tests; $p < .05$). Despite this apparent frequent violation of the Gaussian assumption, the HM algorithm

has proven useful. Nonetheless, methods that adapt to the distributions of the input sensory data as opposed to making fixed (e.g., Gaussian) assumptions about the distributions may enable even more accurate modeling of infant behavior.

An area of application we find interesting for synchrony detection is that of bootstrapping within-modality learning. Synchrony detection acts, for an infant, in part as an attention directing mechanism. Infants find intersensory synchrony interesting, and it thus tends to capture their attention. But, given that infants are in the midst of multiple streams of incoming sensory information, which have redundant qualities, what else are they capable of doing with these streams of information? The Intersensory Redundancy Hypothesis (Bahrick & Lickliter, 2000) predicts that for infants, “detecting amodal properties in *unimodal* stimulation should be more feasible after [they] have had some experience detecting a given property when it was previously bimodally and synchronously available” (p. 191, our italics). On this basis, we speculate that various kinds of apparently within-modality categories, which have an amodal basis (e.g., visual shapes, audio rhythm), may initially be formed by virtue of incoming synchronized streams of intersensory information. If a child is holding an object, feeling the object’s shape, while seeing the object and having it named by their caregiver, they presumably have an excellent opportunity to learn the amodal shape property of the object. Additionally, once infants gain perceptual skills with these within-modality categories, they may then start to be capable of synchrony detection in terms of these higher-level entities (e.g., visual shapes, audio rhythm) 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. Similarly, it has been shown that "language-like sounds [have] some special status as covariates" for infants (quote: p. 140, Baldwin, 1995; research cited: Baldwin & Markman, 1989; see also: Vouloumanos & Werker, 2004). Speech signals may enter as higher-level and distinctive audio parts of the environment into the synchrony detection process.

Finally, whichever of these avenues of research we explore (modifying the algorithm, using different algorithms, incorporating learning and development into the model), in future work, we plan to instantiate synchrony detection models using an active vision system. That is, we plan to use these models in an actual robotic system. Applying audiovisual synchrony detection in an active vision system will, of course, raise at least one other interesting issue. When the active vision system moves its camera, this will result in visual change, and the system will need to distinguish between self-induced visual change, and visual change (along with audio change, in the case of audiovisual synchrony detection) resulting from non-self sources. This raises the need for robotic *self-other* discrimination, and synchrony detection methods may be useful here too.

Bahrnick and Watson (1985) found that 5-month-old infants can detect relations between proprioceptive and visual sensations to discriminate between self and other and "suggest that the contingency provided by a live display of one's body motion is perceived by *detecting the invariant intermodal relationship* between proprioceptive information for motion and the visual display of that motion" (p. 963, our italics). Detection of the proprioceptive-visual invariants that specify self may be facilitated by synchrony detection mechanisms, and these invariants may help robots distinguish between self-motion and other motion in the world. 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).

More generally, we propose that synchrony detection can provide a kind of natural mechanism that epigenetic robotic designers can use to integrate the sensory and sensorimotor systems of epigenetic robots. Robots are not born with a set of integrated sensorimotor systems. The robotics designer must provide these systems. When building an epigenetic robot, using synchrony detection in the service of learning audiovisual relations can act to “glue” together audiovisual information. The invariants (e.g., shapes, self-other) that can be detected with intersensory information may provide an amodal core for the architecture of epigenetic robotic systems. Clearly related to Hebbian learning, such integrative mechanisms are strongly needed in the area of epigenetic robots. This paper represents some first steps at the implementation and testing of these ideas with models and infants.

References

- Arsenio, A., & Fitzpatrick, P. (2003, December). *Exploiting cross-modal rhythm for robot perception of objects*. Paper presented at the 2nd International Conference on Computational Intelligence, Robotics, and Autonomous Systems, Singapore.
- 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.
- Bahrnick, L. E., & Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology, 21*, 963-973.
- Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*, 394-419.
- Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint Attention: Its Origins and Role in Development*. Hillsdale, NJ: Lawrence Erlbaum Assoc.
- Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development, 60*, 381-398.

- Bergeson, T., Houston, D., & Pisoni, D. B. (2004, May). *Audiovisual speech perception in normal-hearing infants and hearing-impaired infants with cochlear implants*. Paper presented at The 14th Biennial International Conference on Infant Studies, Chicago, IL.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Butz, T., & Thiran, J.-P. (2002). *Feature space mutual information in speech-video sequences*. Paper presented at ICME, Lausanne, Switzerland.
- de Sa, V. R. & Ballard, D. H. (1998). Category learning through multimodality sensing. *Neural Computation, 10*, 1097-1117.
- 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.
- Fisher, J. W. III, Darrell, T., Freeman, W. T., & Viola, P. (2001). Learning joint statistical models for audio-visual fusion and segregation. In T. K. Leen, T. G., Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press.
- Fitzpatrick, P. & Arsenio, A. (2004, August). *Feel the beat: Using cross-modal rhythm to integrate perception of objects, others, and self*. Paper presented at The Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Genoa, Italy.
- 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., & Bahrick, 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., & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy*, *2*, 219-231.
- Gogate, L. J., Bahrick, 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.
- 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., Jusczyk, P. W., & Newman, R. S. (2001). Infants' use of visual information in speech segmentation. *Journal of the Acoustical Society America*, *110*, 2703.
- Hollich, G. J., Mislivec, E. J., Helder, N. A., & Prince, C. G. (2004). Are you synching what I'm synching? Modeling infants' real-time detection of audiovisual contingencies

- between face and voice. Poster presented at the *International Conference on Development and Learning (ICDL 04)*, La Jolla, CA, 20-22 October 2004.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.
- Krichmar, J. L., Nitz, D. A., & Edelman, G. M. (2004). Object recognition, adaptive behavior and learning in brain-based devices. In *Proceedings of the Third International Conference on Development and Learning (ICDL)*. Held at La Jolla, CA, October 20-23, 2004.
- Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory 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.
- Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitation view. *Psychological Bulletin*, 126, 281-308.
- Lungarella, M., Metta G., Pfeifer, R., & Sandini G. (2003). Developmental robotics: A survey. *Connection Science*, 15, 151-190.
- Magyar, J., & Gergely, G. (1998, April). 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*, Atlanta, Georgia, USA.
- Mislivec, E. J. (2004). *Audio-Visual Synchrony for Face Location and Segmentation*. Undergraduate Research Opportunity Project, University of Minnesota Duluth.

Internet: <http://www.cprince.com/PubRes/SenseStream>

- 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.
- Nock, H. J., Iyengar, G., & Neti, C. (2004). Multimodal processing by finding common cause. *Communications of the ACM*, *47*, 51-56.
- 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.
- Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, *431*, 768-774.
- Prince, C. G., Helder, N. A., Mislivec, E. J., Ang, B. J., Lim, M. S., & Hollich, G. J. (2003, October). *Taking contingency seriously in sensory-based models of learning in infants*. Poster presented at the 2003 Meeting of the Cognitive Development Society, Park City, UT.
- Prince, C. G., Hollich, G. J., Helder, N. A., Mislivec, E. J., Reddy, A., Salunke, S., & Memon, N. (2004, August). *Taking Synchrony Seriously: A Perceptual-Level Model of Infant Synchrony Detection*. Paper presented at The Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Genoa, Italy. Internet: <http://www.cprince.com/PubRes/EpiRob04>

- Prince, C. G., & Mislivec, E. J. (2004). Investigating the sensory grounding hypothesis: A sensory-augmented connectionist word-learning model. Unpublished manuscript.
- Pylyshyn, Z. W. (1984). *Computation and Cognition: Towards a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.
- 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.
- 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.
- Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, 7, 270-276.
- Vuppla, K. (2004). *Evaluation of Two Synchrony Detection Implementations*. Masters Thesis, University of Minnesota Duluth, Computer Science Department.
- Internet: <http://www.cprince.com/PubRes/VupplaThesis04>

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.

Acknowledgements

This paper is an extended version of Prince, Hollich, Helder, Mislivec, Reddy, Salunke, & Memon (2004). This research was funded in part by a Purdue Research Foundation grant to GJH, by a donation to CGP from Digi-Key Corporation, and by UROP grants from the University of Minnesota Duluth. We thank Kang James for her help with the HM statistics, and Rocio Alba-Flores and Taek Kwon for their image processing advice. We also thank Roberta Golinkoff and Kathy Hirsh-Pasek for their mentorship. Lakshmi Gogate provided valuable comments on an earlier draft, and discussions with David Lewkowicz have also strengthened our thinking. Three anonymous referees also provided very useful comments. This work has benefited greatly by the contributions of the KidCause research team at the University of Minnesota Duluth, and CGP would like to make specific thanks to Nathan Helder and Eric Mislivec. We are indebted to Jeremy Friesner (Level Control Systems) for his open source 'Muscle' software.

Table 1

Stimuli design for infant-model comparisons of synchrony detection. Video files used with model are on the Internet: <http://www.cprince.com/PubRes/JCSR04>

Stim.	Sound Source(s)		Visual Motion Source(s)	
	No.	Description	No.	Description
A	1	“Modi” word	1	Vertical object motion
B1	1	Male voices alternating	2	Two males talking
B2	1 or 2	Female voice (+ male “noise” voice)	2	Same female talking on left and right
C	2	Female voice & male “noise” voice	1	Face of female
D	2	Female voice & male “noise” voice	1	Oscilloscope: female voice

Table 2

Infant and Model Results for Stimulus B2. The same person is speaking on the left and right. The audio is from the left or right and in the noise conditions includes an additional (background) male voice. The horizontal of the video was centered at 360.

Infant Results				Model Results					
Audio Stimulus [†]	Looking to		Proportion to Left	Grayscale			Pixel Intensity Change (PIC)		
	Sides (s)			Average	N	Proportion	Average	N	Proportion
	Left	Right	X-Coord Centroid		to Left	X-Coord Centroid		to Left	
Cup	5.63	4.26	.591*	333.79***	428	0.631***	345.58***	427	0.555*
CupNoise	6.99	4.40	.608*	338.90***	427	0.583***	339.64***	425	0.565**
Dog	5.43	4.52	.525	349.35*	432	0.563*	371.01**	432	0.442*
DogNoise	5.61	5.31	.505	373.15**	434	0.412***	367.64*	432	0.479

Notes: * = $p < .05$; ** = $p < .005$; *** = $p < .0005$. N = the number of mixelgrams analyzed for centroids. Some N's are less than expected as some centroids were not computed due to zero centroid mass (i.e., blank mixelgrams). † = The visual stimulus in each case was the same, only the audio stimulus was varied per video clip.

Table 3

Stimulus C and D model results. The audio was synchronized (Female Voice), partially synchronized (Both Voices), or not synchronized (Male Voice condition) with either a Face or an Oscilloscope (see text for further description). Table entries are averages of per mixelgram synchrony estimates from the edge detection method.

Stimulus		Female	Both	Male
		Voice	Voices	Voice
C				
(Face)	Mean	21, 139.1	19, 438.8	17, 444.4
	Std Dev	14, 973.8	11, 220.5	12, 116.6
	Max	77, 518.4	64, 085.1	78, 751.6
	Min	3.9	3.4	768.1
D				
(Oscilloscope)	Mean	7, 287.2	7, 971.9	7, 233.3
	Std Dev	4, 077.7	2, 903.4	3, 138.7
	Max	29, 822.7	28, 309.5	28, 056.0
	Min	9.2	1038.2	535.0

Notes: ** = $p < 0.005$; *** = $p < 0.0005$.

Figure Captions

Figure 1. Processing Flow of the HM Algorithm. Data from one channel consists of single vectors of n -elements, here processed audio features. Data from the other channel consists of $h \times w$ m -element vectors, here visual features from frames of digital motion video. Each feature is a real number. After the audio and visual features have been computed, the joint audiovisual covariance matrix is computed—which contains as sub-covariance matrices the audio and visual covariance matrices. Audio attenuation is not shown here.

Figure 2. A mixelgram (upper panel) from an interval of synchronous audiovisual data in Stimulus A. One representative visual frame in the same interval is shown in the lower panel. Mixelgrams are typically perceptually relevant only when the two input streams (e.g., audiovisual) are synchronous (i.e., co-varying; see Vuppla, 2004). The thicker lines around and on the object in the mixelgram have larger size connected regions than some of the other regions (see text).

Figure 3. Model estimates of the quantitative degree of audiovisual synchrony for Stimulus A (punctuate object motion and speech). Synchrony estimates were computed using the edge detection method, and pixel intensity change visual features, and are presented as a running average of estimates—5 neighbors on either side (11 total). The speech (word) onsets and offsets were obtained manually.

Figure 4. Stimulus B1 model results using mixelgram centroid to determine speaker location. Video clips were 720 pixels wide, but scaled by $2/3$ for processing—resulting in a midpoint of 240 pixels. Higher numbered pixels are to the right; lower

numbered are to the left. The labels above the top panel indicate the source of the audio: Either the talker on the right, the talker on the left, or background audio (noise) only.

Figure 5. Stimulus B1 (Both Speakers, But Talking Alternately) Processed with RGB (Color) Visual Features. The centroids are all on the right side (higher coordinates), which differs from processing with grayscale visual features, in which the centroids were primarily on the left (see Figure 4, top panel). The labels above the graph indicate the source of the audio: Either the talker on the right, the talker on the left, or background audio (noise) only.

Figure 1. Processing Flow of the HM Algorithm. Data from one channel consists of single vectors of n -elements, here processed audio features. Data from the other channel consists of $h \times w$ m -element vectors, here visual features from frames of digital motion video. Each feature is a real number. After the audio and visual features have been computed, the joint audiovisual covariance matrix is computed—which contains as sub-covariance matrices the audio and visual covariance matrices. Audio attenuation is not shown here.

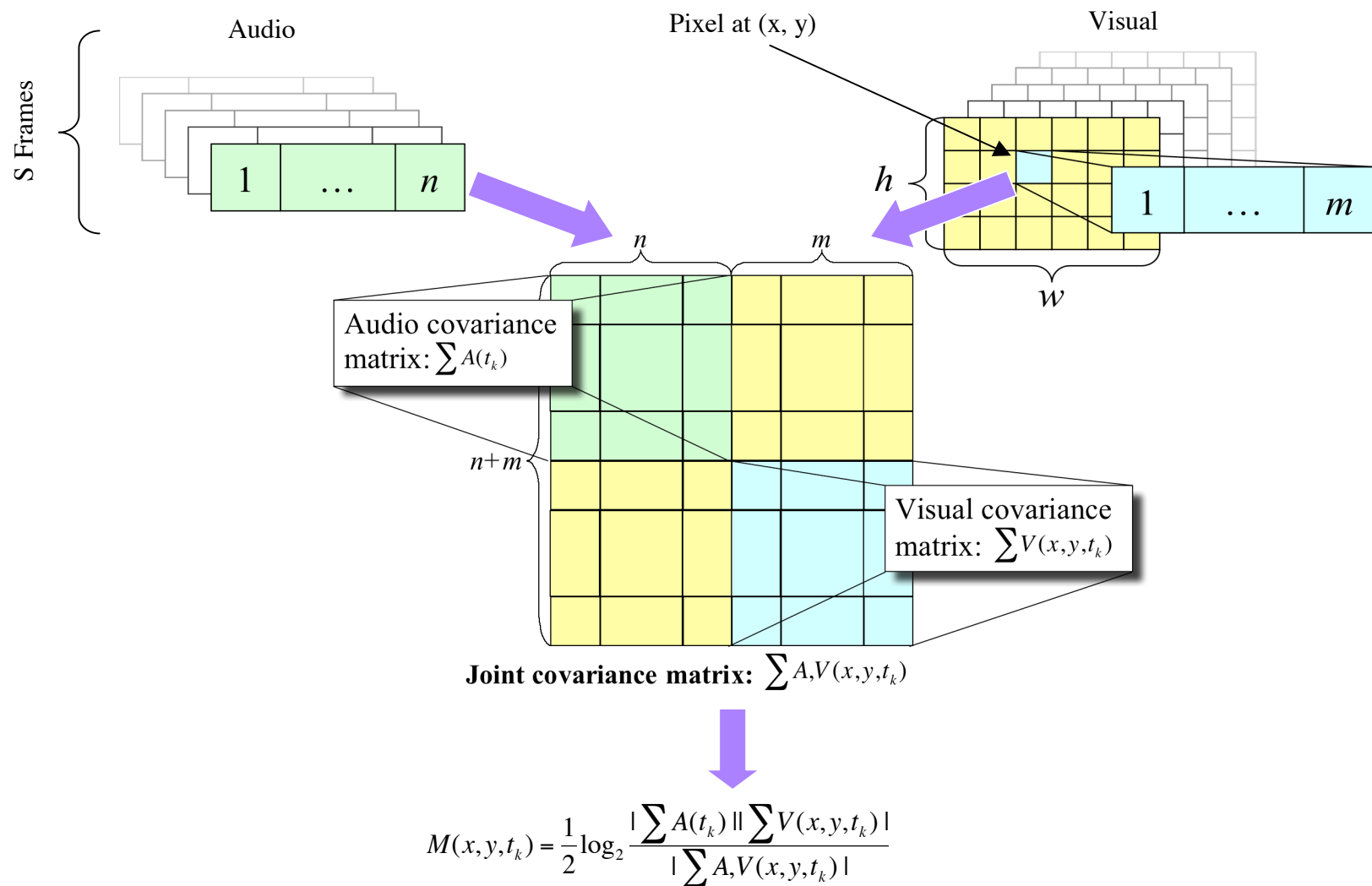


Figure 2. A mixelgram (upper panel) from an interval of synchronous audiovisual data in Stimulus A. One representative visual frame in the same interval is shown in the lower panel. Mixelgrams are typically perceptually relevant only when the two input streams (e.g., audiovisual) are synchronous (i.e., co-varying; see Vuppla, 2004). The thicker lines around and on the object in the mixelgram have larger size connected regions than some of the other regions (see text).



Figure 3. Model estimates of the quantitative degree of audiovisual synchrony for Stimulus A (punctuate object motion and speech). Synchrony estimates were computed using the edge detection method, and pixel intensity change visual features, and are presented as a running average of estimates—5 neighbors on either side (11 total). The speech (word) onsets and offsets were obtained manually.

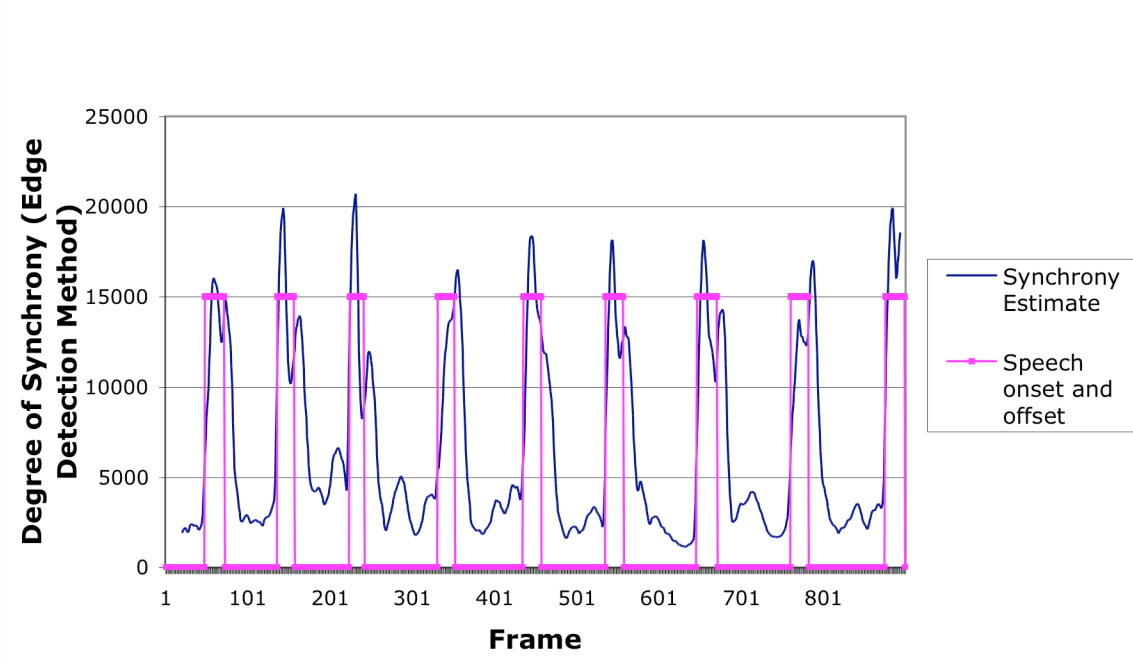


Figure 4. Stimulus B1 model results using mixelgram centroid to determine speaker location. Video clips were 720 pixels wide, but scaled by 2/3 for processing—resulting in a midpoint of 240 pixels. Higher numbered pixels are to the right; lower numbered are to the left. The labels above the top panel indicate the source of the audio: Either the talker on the right, the talker on the left, or background audio (noise) only.

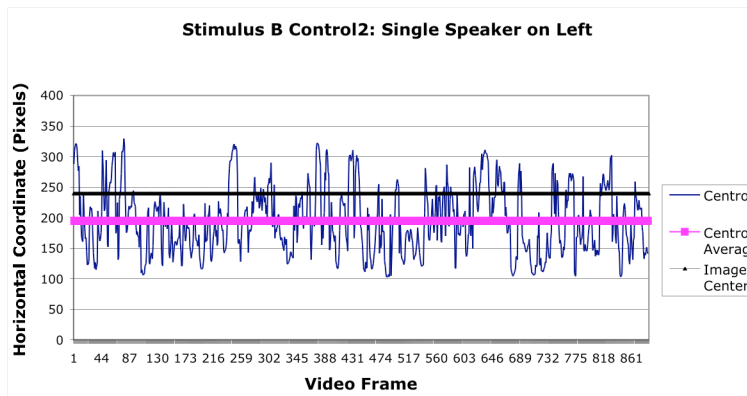
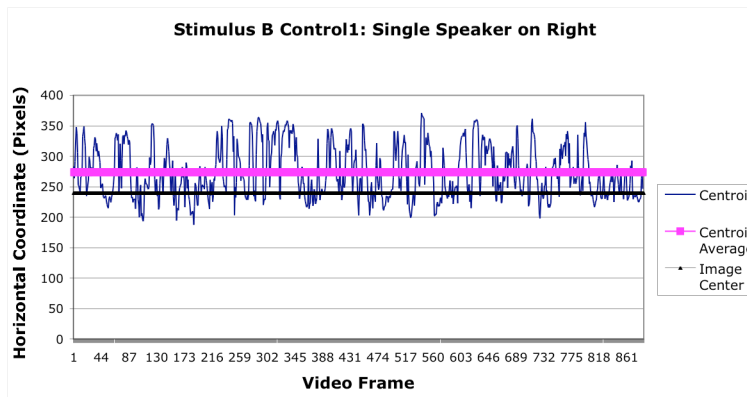
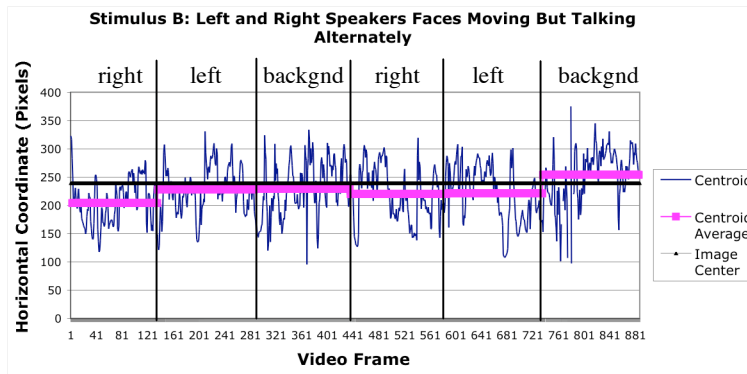


Figure 5. Stimulus B1 (Both Speakers, But Talking Alternately) Processed with RGB (Color) Visual Features. The centroids are all on the right side (higher coordinates), which differs from processing with grayscale visual features, in which the centroids were primarily on the left (see Figure 4, top panel). The labels above the graph indicate the source of the audio: Either the talker on the right, the talker on the left, or background audio (noise) only.

