Developmental Psychology
2010, Vol. 46, No. 1, 66–77

0012-1649/10/$12.00 DOI: 10.1037/a0015579

Infant Perception of Audio-Visual Speech Synchrony

David J. Lewkowicz
Florida Atlantic University

Three experiments investigated perception of audio-visual (A-V) speech synchrony in 4- to 10-month-old infants. Experiments 1 and 2 used a convergent-operations approach by habituating infants to an audiovisually synchronous syllable (Experiment 1) and then testing for detection of increasing degrees of A-V asynchrony (366, 500, and 666 ms) or by habituating infants to a detectably asynchronous syllable (666 ms; Experiment 2) and then testing for detection of decreasing degrees of asynchrony (500, 366, and 0 ms). Following habituation to the synchronous syllable, infants detected only the largest A-V asynchrony (0 ms vs. 666 ms), whereas following habituation to the asynchronous syllable, infants detected the largest asynchrony (666 ms vs. 0 ms) as well as a smaller one (666 ms vs. 366 ms). Experiment 3 investigated the underlying mechanism of A-V asynchrony detection and indicated that responsiveness was based on a sensitivity to stimulus-energy onsets rather than the dynamic correlation between acoustic and visible utterance attributes. These findings demonstrated that infant perception of A-V speech synchrony is subject to the effects of short-term experience and that it is driven by a low-level, domain-general mechanism.

Keywords: intersensory, infants, audiovisual speech, audio-visual synchrony

When infants interact with the people in their world, they usually can hear and see them talking. This means that during most of their daily social interactions, infants have access to concurrent and highly redundant auditory and visual speech information. Indeed, most of our real-world perceptual experiences are usually specified by synchronous and highly redundant multisensory perceptual attributes, and as a result, we typically have access to multiple and redundant multisensory attributes. This is highly advantageous because it increases perceptual salience and, as a result, enhances perception, learning, and discrimination (Bahrick, Lickliter, & Flom, 2004; Lewkowicz & Kraebel, 2004; Partan & Marler, 1999; Rowe, 1999; Stein & Meredith, 1993; Stein & Stanford, 2008).

Audiovisual speech is a particularly rich source of multisensory redundancy because the concurrent streams of auditory and visual information provide the observer with synchronous and overlapping patterns of stimulation that are correlated in multiple ways (Munhall & Vatikiotis-Bateson, 2004; Yehia, Rubin, & Vatikiotis-Bateson, 1998). For example, in addition to always having synchronous onsets and offsets, the auditory and visual streams of a speech utterance are usually invariant in terms of their duration, tempo, rhythmical patterning, intensity variations, and even affective tone. Studies comparing comprehension of auditory-only speech versus audiovisual speech have shown that comprehension of the latter is significantly better (Sumby & Pollack, 1954; Summerfield, 1979). Comparable data are currently not available for infants, but evidence from studies of prelingually deaf children wearing cochlear implants indicates that they exhibit multisensory enhancement effects in tasks that measure their ability to comprehend speech utterances (Bergeson, Pisoni, & Davis, 2005).

The audio-visual (A-V) temporal synchrony that is always available in audiovisual speech is an especially important source of multisensory redundancy for an immature and perceptually inexperienced organism. There are several reasons for this. First, the detection of A-V temporal synchrony requires only the perception of the onset and offset of stimulus energy across different modalities; it does not require the perception of any of the complex dynamic stimulus features. Second, because the patterns of auditory and visual information that specify everyday objects and events are always temporally synchronized (J. J. Gibson, 1966), temporally synchronous patterns of auditory and visual stimulation are available to infants from birth on (E. J. Gibson, 1969; Lewkowicz, 2000a; Thelen & Smith, 1994) and thus provide a ready-made basis for intersensory integration. Third, the detection of temporal A-V synchrony is mediated primarily by low-level, subcortical tecto-thalamo-insular pathways (Bushara, Grafman, & Hallett, 2001). This means that the neural mechanisms necessary for the detection of temporal A-V temporal synchrony are likely to be present from an early age. Finally, A-V temporal synchrony provides a simple but powerful way of bootstrapping infants' discovery of the coherent nature of their multisensory world. As long as they can detect the onsets and offsets of the auditory and visual attributes of a speech utterance, they can begin to perceive a talking face as a coherent and unified event. Once they do that, through perceptual learning and differentiation, they can go on to discover the finer and higher level invariant perceptual attributes that also specify faces and voices (e.g., identity, affect, and gender).

This work was supported by National Institute of Child Health and Human Development Grant R01 HD35849. I thank Erinn Frances Beck for her assistance in data collection. I also thank Lorraine Bahrick and Robert Lickliter for comments on an early draft of this article.

Correspondence concerning this article should be addressed to David J. Lewkowicz, Department of Psychology, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431. E-mail: lewkowic@fau.edu

Evidence indicates that intersensory perception emerges early in life and that it improves and broadens in scope rapidly during the first year of life (Lewkowicz, 2000a, 2002, 2003; Lewkowicz & Lickliter, 1994; Lickliter & Bahrick, 2000; Neil, Chee-Ruiter, Scheier, Lewkowicz, & Shimojo, 2006; Walker-Andrews, 1997; Wallace, 2004). This evidence shows that infants are sensitive to various types of intersensory cues—temporal synchrony, intensity, duration, tempo, rhythm, and affect—and that they take advantage of them in their response to multisensory objects and events. Furthermore, consistent with the likelihood that detection of A-V temporal synchrony is relatively easy and that the neural mechanisms necessary for its detection are relatively low level, the ability to perceive A-V temporal synchrony emerges very early in life and is so general in scope that infants can detect it across a wide range of object–event complexity. For example, studies have shown that infants can perceive the temporal synchrony between the audible and the visible attributes of multisensory objects regardless of whether these are static or sounding objects (Lewkowicz, 1986), moving or sounding objects (Bahrick, 1988; Lewkowicz, 1992a, 1992b, 1996b; Scheier, Lewkowicz, & Shimojo, 2003), talking faces (Dodd, 1979; Lewkowicz, 1996a, 2000b, 2003), or singing faces (Lewkowicz, 1998). Moreover, studies have shown that infants can (a) choose one of two superimposed moving objects on the basis of a sound that is synchronous with one of them (Bahrick, Walker, & Neisser, 1981), (b) use the specific spatiotemporal synchrony between a sound and an ambiguous visual event to disambiguate the event (Scheier et al., 2003), (c) segregate competing streams of auditory speech on the basis of a face whose movements are synchronous with one of the streams (Hollich, Newman, & Jusczyk, 2005), (d) take advantage of A-V synchrony to achieve more robust learning and discrimination (Bahrick & Lickliter, 2000; Lickliter, Bahrick, & Honeycutt, 2002, 2004; Walker-Andrews, 1986), and (e) associate synchronous faces and voices (Bahrick, Hernandez-Reif, & Flom, 2005; Brookes et al., 2001). Finally, young infants can even integrate nonhuman (i.e., monkey) faces and vocalizations as long as they are temporally synchronous (Lewkowicz & Ghazanfar, 2006; Lewkowicz, Sowinski, & Place, 2008).

Despite the general importance of A-V temporal synchrony in infant perception, and despite the fact that infants are sensitive to it from an early age, it is interesting to note that relative to adults, infants are less sensitive to A-V synchrony relations. For example, adults can detect a desynchronization of the audible and visible attributes of a bouncing, sounding object when the audible bounce precedes the visible one by as little as 80 ms (Lewkowicz, 1996b). In contrast, infants can detect such a desynchronization only when the audible bounce precedes the visible bounce by at least 350 ms (Lewkowicz, 1996b). Likewise, adults can perceive a desynchronization of the audible and visible streams of an audiovisual speech utterance when the former precedes the latter by anywhere between 41 ms (Grant, van Wassenhove, & Poeppel, 2004) and 130 ms (Dixon & Spitz, 1980). In contrast, infants can detect such a desynchronization only when the audible stream precedes the visible one by 633–666 ms (Lewkowicz, 2000b, 2003).

Currently, the precise magnitude of the infant threshold for the perception of A-V speech synchrony is not known. Knowing its magnitude has important implications because it would provide baseline information on a critical perceptual capacity in early development and would yield important insights into the development of temporal processing skills in infancy and their role in perceptual and cognitive development (Lewkowicz, 1989). It is already known that infants possess impressive timekeeping abilities. For example, they are able to anticipate specific time intervals (Colombo & Richman, 2002) and can predict the future occurrence of visual events on the basis of short-term experience (Canfield & Haith, 1991; Wentworth & Haith, 1998). These types of timekeeping abilities are consistent with findings that infants are sensitive to A-V synchrony relations and, together, are a testament to the early emergence of timekeeping skills. Nonetheless, infant timekeeping abilities are inferior to those in adults, as indicated by findings that time estimation abilities improve during development (Brannon, Suanda, & Libertus, 2007) and that it takes many years before children can understand the various aspects of the time dimension (Nelson, 1986, 2007).

Determining the size of the A-V speech synchrony threshold not only would shed additional light on the development of temporal processing skills in infancy but, more specifically, would also help determine whether the size of the infant intersensory temporal contiguity window (ITCW) is affected by the nature of the input (i.e., speech vs. nonspeech) and whether experience affects its magnitude. The concept of the ITCW was first introduced by Lewkowicz (1996b) to account for infant response to A-V temporal synchrony relations and captures the fact that integration of concurrent auditory and visual sensory inputs does not require such inputs to be in perfect temporal register. That is, even when such inputs are physically separate in time, they are perceived as part and parcel of the same event as long as they fall into what has been called the "psychological present" (Fraisse, 1982). The length of the physical separation that determines the psychological present is what determines the size of the ITCW. As indicated earlier, the available evidence suggests that the size of the ITCW is affected by the nature of the input, but in the absence of more direct empirical evidence, it is still not clear to what extent it is affected. Moreover, evidence suggests that the size of the ITCW is affected by short-term experience. For example, when adults are first tested with audiovisually asynchronous events, they perceive them as such, but after short-term exposure to them, they begin to respond to them as if they were synchronous (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Navarra et al., 2005; Vroomen, Keetels, de Gelder, & Bertelson, 2004). In other words, short-term adaptation modifies the ITCW in adults. The question is whether it does so in infants as well.

Given the aforementioned importance of determining the precise threshold for the detection of the A-V temporal synchrony threshold for audiovisual speech, the current study had three specific goals. The first, addressed in Experiments 1 and 2, was to obtain an estimate of the size of the infant ITCW for audiovisual speech. The second, also addressed in Experiments 1 and 2, was to investigate the possible differential effects of short-term experience on the size of the ITCW. Finally, the third goal, addressed in Experiment 3, was to determine what specific stimulus attributes mediate infant response to A-V speech synchrony relations.

To address the first goal, I habituated infants either to an audiovisually synchronous syllable and then tested for detection of increasing degrees of A-V asynchrony (Experiment 1) or to an audiovisually asynchronous syllable and then tested for detection of decreasing degrees of A-V asynchrony (Experiment 2). Importantly, in both experiments, the different levels of temporal A-V synchrony spanned a large enough range so that it would include perceptually synchronous as well as asynchronous syllables. It was expected that as long as the ability to perceive A-V temporal synchrony relations in audiovisual speech is robust, then infants should be able to detect synchrony relation changes regardless of whether they first learn a synchronous or an asynchronous audiovisual syllable. To address the second goal, I compared the results from Experiments 1 and 2 to determine whether initial exposure to an asynchronous audiovisual event produces the kind of adaptation effect observed in adults. Finally, to address the third goal, I habituated infants with the visible /ba/ and a synchronous tone rather than the audible /ba/. Replacing the audible /ba/ with the tone had two consequences. First, the tone effectively eliminated the energy and spectral variations that are typical of an acoustic speech signal and thus would make it possible to evaluate the contribution of these two sources of information to the perception of A-V temporal synchrony in audiovisual speech. Second, replacing the audible /ba/ with the tone also eliminated the normally available correlation between the dynamic changes in acoustic and visual information that normally specify an audiovisual utterance. If this correlation mediates infant response to A-V temporal synchrony relations, then its absence should result in a failure to detect A-V desynchronization. If, however, neither the variations in acoustic energy and spectral information nor their correlation with the dynamic facial information mediates detection of A-V synchrony, then infants should still detect A-V desynchronization. The latter result would suggest that infant perception of A-V temporal synchrony relations is based on a relatively low-level but functionally powerful sensitivity to stimulus-energy onsets and offsets. This, in turn, would suggest that perception of A-V temporal synchrony relations in infancy is governed by a domain-general mechanism.

Experiment 1

To determine the size of the ITCW, I habituated infants first to an audiovisually synchronous /ba/ syllable. Then they were given a set of test trials in which the audible and visible syllable attributes were desynchronized by 366, 500, and 666 ms, and the question was whether infants would detect any of these asynchronies as novel. Prior findings (Lewkowicz, 2000b, 2003) have indicated that infants can detect an asynchrony of 633–666 ms, but it is currently not known whether they also might be able to detect lower temporal asynchronies.

Method

Participants. Thirty-seven infants completed the experiment. Of these, 5 were subsequently eliminated from data analyses because they either took too many trials to habituate, exhibited spontaneous regression effects, or exhibited fatigue effects (see the *Results and Discussion* section for more details). The remaining 32 infants (18 boys and 14 girls) contributed data in this experiment and consisted of separate groups of 8 infants at each of the following ages: 4 months (mean age = 17 weeks, *SD* = 1.0 weeks), 6 months (mean age = 26.4 weeks, *SD* = 1.5 weeks), 8 months (mean age = 34.6 weeks, *SD* = 0.4 weeks), and 10 months (mean age = 42.9 weeks, *SD* = 0.5 weeks). An additional 15 infants were tested but did not complete the session because 14 of them fussed and because 1 was distracted by parental interference. The infants in all experiments were full-term at the time of birth, had birth weights above 2,500 g and Apgar scores of 7 or higher, had no prenatal or perinatal complications, and were healthy at the time of testing. The ethnic proportions of the infants in all the experiments reported here were approximately 5% Hispanic, 77% non-Hispanic, and 18% unreported. The corresponding racial proportions were 4% Black or African American, 0% Asian, 3% more than one race, 73% White or Caucasian, and 20% unreported. Participants came from Palm Beach and Broward counties in South Florida, and their names were obtained from birth records provided by the Florida Department of Health.

Apparatus and stimuli. All the stimuli presented in this experiment were multimedia movies. One, the attention-getter movie, showed a continuously expanding and contracting green disk in the middle of the screen. A second movie, used to test for possible fatigue effects, showed a 1-min segment of a Winnie the Pooh cartoon (presented at 73.1–76.5 dB, A scale, against an ambient sound pressure level of 50 dB). The remaining four 1-min movies showed the face of a female actor, looking directly at the camera, repeatedly uttering the syllable /ba/. Each of these movies was constructed with the aid of Premiere 6.0 (Adobe Systems, San Jose, CA) and consisted of concatenated copies of a 4-s video clip of a /ba/ syllable (the following settings were used to construct the movies: 720 × 480-pixel image, Cinepak codec video compression, 1,024 kbps audio sample rate, and 29.97 frames/s).

Each 4-s video clip began with the actor holding her lips in the closed position until 1.66 s had elapsed. At this point, the actor began to make the preparatory mouth movements for producing the syllable by constricting her lips. She then opened her mouth, phonated, and then closed her mouth. The phonation began at 2.033 s and ended at 2.231 s. Mouth motion ended at 2.693 s. As soon as she stopped moving her lips, she resumed a still face with lips closed until the end of the clip. Figure 1A shows still images of the actor's face corresponding to each of these phases; Figures 1B and 1C show spectrograms and waveforms of the accompanying sound. The actor held her head still throughout the entire clip and had a neutral expression on her face. Because the actor held her head still throughout the clip, the transition between each utterance was nearly imperceptible and resembled a person repeatedly uttering the /ba/ syllable in a natural fashion.

One of the four syllable movies depicted the actor uttering an audiovisually synchronous /ba/. This movie was presented during the habituation phase. This movie was also used as the familiar test trial and thus was dubbed the FAM 0 test trial movie (*FAM* designates the fact that this was the familiar test trial, and the number after it designates the temporal separation between the visual and the auditory signals in milliseconds). The other three movies served as the novel test trial movies. In these movies, the auditory signal was moved ahead of the visual one by 11 (NOV 366), 15 (NOV 500), and 20 (NOV 666) video frames, respectively.

All movies were presented on a 17-in. (43.18-cm) computer monitor at a distance of 50 cm from the infant, who was seated either in an infant seat or in the parent's lap in a dimly lit, sound-attenuated, single-wall chamber (iModules Software, Overland Park, KS). Most infants were seated alone in the infant seat. In those few cases in which infants were seated in the parent's lap, the parent was blind with respect to the hypothesis under test, wore

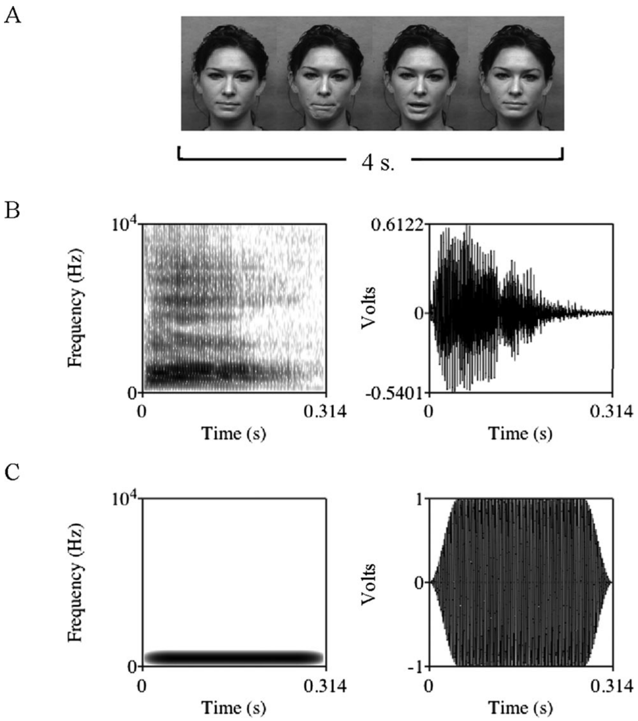

Figure 1. Figure 1A shows images of the actor's face with the different gestures performed during the production of a single instance of the /ba/ syllable. The first and last gestures show the closed position of the lips prior to a phonation and after its termination, respectively. The second gesture from the left shows lip position in the frame preceding phonation, and the third gesture from the left shows lip position during maximal phonation. Figure 1B shows the spectrogram and waveform of the /ba/ syllable presented in Experiments 1 and 2. Figure 1C shows the spectrogram and waveform of the tone presented in Experiment 3. The individual whose face appears here gave signed consent for her likeness to be published in this article.

headphones and listened to white noise during the test, and was asked to sit as still as possible and not interact with the baby. The audio part of the movie was presented through two speakers placed on each side of the monitor. The mean sound pressure level of the audio portion of the syllable was 87.9 dB (A scale). A video camera was located on top of the stimulus-presentation monitor and was used to transmit an image of the infant's face to a monitor outside the chamber.

Procedure. An experimenter, who was located outside the testing chamber, observed the infant on a video monitor and could neither see nor hear the stimuli being presented. An infant-controlled habituation–test procedure was used, and thus the infant's looking behavior controlled movie presentation. Whenever the infant looked at the stimulus-presentation monitor, the movie began to play and continued to do so until the infant either looked away from the monitor for more than 1 s or accumulated a total of 60 s looking time. Once either of these two conditions was met, the attention-getter appeared again and remained there until the infant looked back at it. At this point, the attention-getter was turned off, and the next trial began.

The experiment began with a single pretest trial during which the Winnie the Pooh cartoon was played. The purpose of this trial was to measure the infant's initial level of attention. Once this trial ended, the habituation phase began. This was followed by the habituation phase in which infants were habituated to the synchronous movie. Habituation continued until an infant reached a predefined habituation criterion. The criterion required that the infant's total duration of looking during the last three habituation trials decline to 50% of the total duration of looking during the infant's first three habituation trials. Once the infant reached the habituation criterion, the habituation phase ended, and the next trial initiated the test phase. The four test trials (FAM 0, NOV 366, NOV 500, and NOV 666) were presented during the test phase and were administered in counterbalanced order across infants according to a Latin square design. This resulted in four test-trial-order groups. Once all the test trials were presented, a single posttest trial was administered. Here the Winnie the Pooh cartoon was played again, and once it ended, the experiment ended. The duration of looking during the posttest trial provided a measure of the infant's terminal level of attention and served as a check on fatigue effects.

Results and Discussion

As a first step, I conducted an analysis to determine whether any infants (a) failed to pay sufficient attention during the habituation phase, (b) exhibited spontaneous regression to the mean in the test phase, and/or (c) exhibited fatigue by the end of the experimental session. Failure to pay sufficient attention during the habituation phase was defined as taking more than 25 trials to reach the habituation criterion. Two infants took more than 25 trials to reach the habituation criterion, and thus their data were eliminated from further analyses. Spontaneous regression to the mean in the test phase was defined as a looking duration score in the FAM 0 test trial that exceeded the mean duration of looking in that trial by more than two standard deviations. On the basis of this criterion, 2 infants exhibited spontaneous regression to the mean, and as a result their data were eliminated from further analyses. Finally, excessive fatigue by the end of the experimental session was defined as a looking duration score during the posttest trial that was less than 150% of the looking duration score obtained during the FAM 0 test trial. According to this criterion, 1 infant exhibited fatigue effects, and this infant's data were eliminated from further analyses.

Figures 2A and 2B show the results from the habituation and test trials, respectively, for the remaining 32 infants. Figure 2A shows only the data from the first three and last three habituation trials because the infant-controlled procedure permits a variable number of habituation trials across infants, and thus this is the only way to effectively present the habituation results for the entire group. As can be seen, attention declined during the habituation phase, and a 4 × 6 (Age × Trials) repeated measures analysis of variance (ANOVA), with Age as the between-subjects factor and Trials as the within-subjects factor, showed that the response decline observed during the habituation trials was statistically significant, $F(5, 140) = 31.3$, $p < .001$, $\eta_p^2 = .53$. The ANOVA also indicated that there was a significant Age effect, $F(3, 28) = 3.35$, $p = .033$, $\eta_p^2 = .26$, but because this was due primarily to longer overall looking in the 4-month-old infants than in the other three groups, this effect has no bearing on the interpretation of the data. More important, the Age × Trials interaction was not sig-

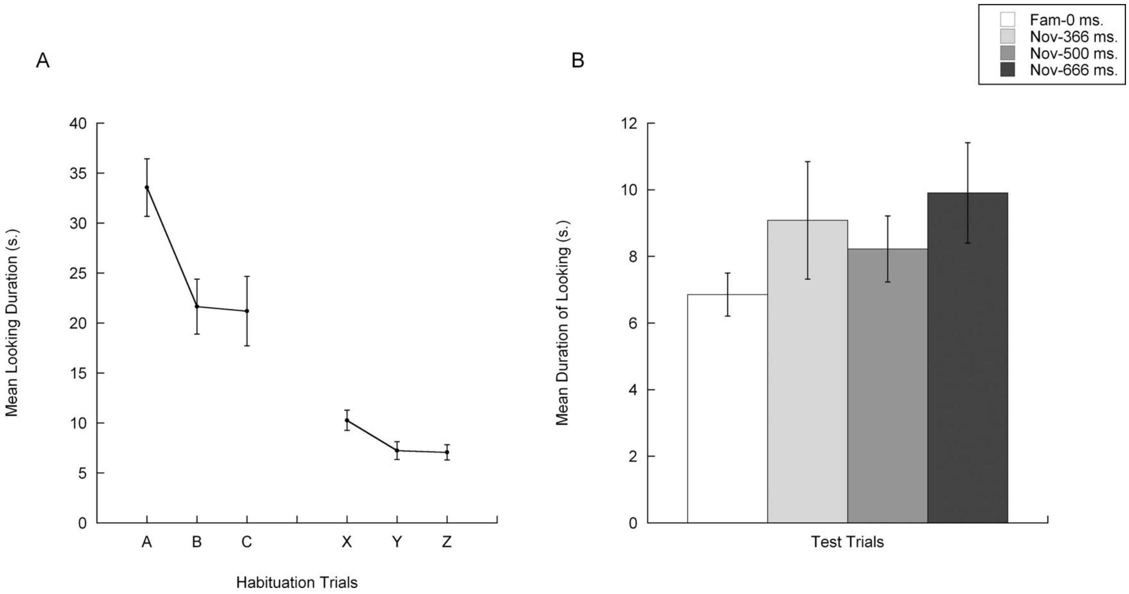

Figure 2. Results from Experiment 1. Figure 2A shows the mean duration of looking during the habituation phase. Because different infants took different numbers of trials to reach the habituation criterion, Figure 2A depicts the average duration of looking in the first three (A, B, and C) and the last three (X, Y, and Z) habituation trials. Figure 2B shows the mean duration of looking during the test trials. Error bars indicate the standard error of the mean.

nificant. As a result, these data indicated that, regardless of age, infants exhibited statistically reliable response habituation.

As indicated earlier, prior findings (Lewkowicz, 2000b, 2003) have shown that infants can detect an A-V speech asynchrony of 633–666 ms. Consequently, I performed a set of a priori planned contrast analyses to determine which of the three asynchronies presented here infants detected. I conducted an ANOVA first to determine whether infants' age or test trial order affected responsiveness. Thus, I submitted the data from the four test trials to a $4 \times 4 \times 4$ (Age $\times$ Test Trial Order $\times$ Test Trial Type) mixed, repeated measures ANOVA, with Age and Test Trial Order as the between-subjects factors and Test Trial Type as a within-subjects factor. This analysis yielded no significant meaningful main effects or interactions, indicating that responsiveness was not affected by the infants' age or the order of the test trials. The planned contrast analyses compared the duration of looking in each of the asynchrony test trials with the duration of looking in the FAM 0 test trial. These analyses indicated that infants did not exhibit significant response recovery in the NOV 366 or the NOV 500 test trials but that they did exhibit significant response recovery in the NOV 666 test trial, $F(1, 16) = 5.22$, $p = .036$, $\eta_p^2 = .066$. The finding that infants detected the 666-ms asynchrony replicates prior results indicating that infants can detect an A-V speech asynchrony of 633 ms (Lewkowicz, 2003) and 666 ms (Lewkowicz, 2000b). The finding that infants did not exhibit significant response recovery in the NOV 500 and NOV 366 test trials suggests that the threshold for the detection of A-V asynchrony is located somewhere between 500 ms and 666 ms.

Experiment 2

Using a convergent-operations approach, Experiment 2 was designed to provide an independent test of infant sensitivity to A-V speech synchrony relations by adopting the opposite testing approach. That is, here infants were first habituated to what was established by Experiment 1 to be a perceptually detectable A-V asynchrony of 666 ms and then given test trials during which their ability to detect smaller A-V asynchronies was investigated. From a theoretical perspective, this procedure can produce one of two possible outcomes. If sensitivity to A-V synchrony differences is absolute, then infants should discriminate an A-V asynchrony only between 666 ms and 0 ms. If, however, sensitivity to A-V synchrony depends on initial short-term experience, and if experience with temporally asynchronous audible and visible syllable attributes serves to highlight their specific temporal relation, then infants might exhibit better discriminative performance.

Some clues as to which is the more likely outcome can be gleaned from work with adults showing that A-V synchrony thresholds are not absolute but, rather, susceptible to the effects of prior short-term experience (Fujisaki et al., 2004; Navarra et al., 2005; Vroomen et al., 2004). Specifically, this work has demonstrated that when adults are first adapted to an asynchronous audiovisual event, they actually begin to respond to what they normally perceive as asynchronous audiovisual events as synchronous ones. In other words, their sensitivity to A-V temporal synchrony relations is reduced by short-term adaptation to asynchronous events. The most reasonable explanation for this effect is that it reflects the massive perceptual experience that adults have accumulated during a lifetime of exposure to mostly synchronous multisensory events. The result of this experience is that adults approach the task with a perceptual bias for A-V synchrony. The adult findings suggest that infants also might exhibit adaptation effects. In contrast to those of adults, however, the adaptation effects may be opposite in infants. Infants have had far less perceptual experience and, as a result, are not likely to possess the same bias for audiovisual synchrony. Thus, initial exposure to an asynchronous event may act to highlight the temporal discrepancy of its audible and visible attributes and, in the process, may increase infant sensitivity to A-V temporal relations. The current experiment tested this possibility.

Method

Participants. Forty-three infants completed the experiment. Of these, 6 were eliminated from data analysis because they either took too many trials to habituate ($n = 1$), exhibited spontaneous regression effects ($n = 2$), or exhibited fatigue effects ($n = 3$). The remaining 37 infants (18 boys and 19 girls) contributed data and consisted of separate groups of eight 4-month-olds (mean age = 17.2 weeks, $SD = 1.5$ weeks), ten 6-month-olds (mean age = 26.2 weeks, $SD = 1.2$ weeks), eleven 8-month-olds (mean age = 34.7 weeks, $SD = 1.1$ weeks), and eight 10-month-olds (mean age = 43 weeks, $SD = 0.8$ weeks). An additional 12 infants were tested but did not complete the experiment because 11 of them fussed and 1 was distracted.

Apparatus and stimuli. All stimulus materials and apparatus were identical to those employed in Experiment 1.

Procedure. The procedure was identical to that used in Experiment 1 except that here infants were habituated to an A-V asynchrony of 666 ms. After habituation, infants were given the following test trials: FAM 666, NOV 500, NOV 366, and NOV 0. Once the last of these test trials was presented, the posttest trial was administered and the experiment ended.

Results and Discussion

Figures 3A and 3B show the results from the habituation and test trials, respectively. Figure 3A shows that looking declined during the habituation phase. A 4 × 6 (Age × Trials) repeated measures ANOVA, with Age as the between-subjects factor and Trials as the within-subjects factor, yielded a significant Trials effect, $F(5, 165) = 32.03$, $p < .001$, $\eta_p^2 = .49$, and no other significant effects. This finding shows that, as in Experiment 1, infants exhibited a statistically reliable response decrement during the habituation phase and that this did not depend on their age.

The initial analysis of the test trial data, consisting of a 4 × 4 × 4 (Age × Test Trial Order × Test Trial Type) mixed, repeated measures ANOVA, with Age and Test Trial Order as the between-subjects factors and Test Trial Type as a within-subjects factor, indicated that there were no significant meaningful main effects or interactions. The planned contrast analyses, comparing the duration of looking in each novel test trial with the duration of looking in the FAM 666 test trial, indicated that infants did not exhibit significant response recovery in the NOV 500 test trial but that they did in the NOV 366, $F(1, 21) = 7.09$, $p = .014$, $\eta_p^2 = .039$, and the NOV 0, $F(1, 21) = 6.13$, $p = .022$, $\eta_p^2 = .097$, test trials. The finding that infants detected the difference in the A-V synchrony relation between the FAM 666 test trial and the two maximally different test trials shows that infants successfully encoded the asynchronous A-V relation and that they then detected changes in that relation. These results provide converging evidence that infants are sensitive to A-V speech synchrony relations by showing that in addition to being able to detect a broadening of the temporal relation between the audible and the visible syllable attributes specifying a speech utterance (Experiment 1), they are able to detect its narrowing.

The finding that infants detected an A-V temporal change of 666 ms (i.e., 666 ms vs. 0 ms) as well as a change of 300 ms (i.e., 666 ms vs. 366 ms) indicates that they exhibited greater sensitivity to A-V temporal relations following initial experience with a temporally discordant A-V speech syllable. In other words, highlighting the A-V temporal relations during the initial learning phase in-

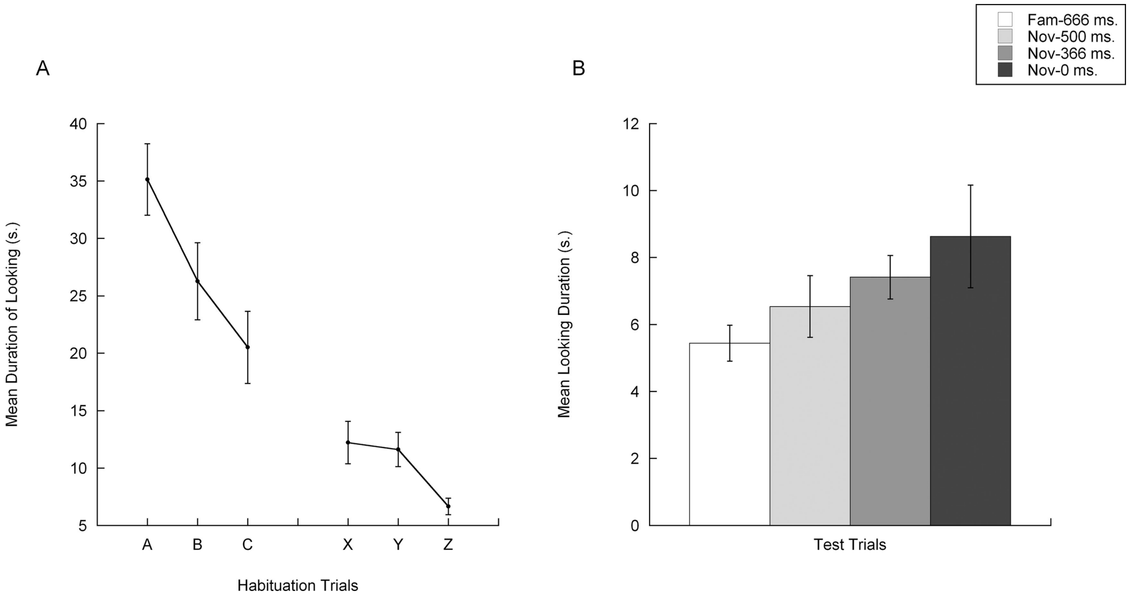

Figure 3. Results from Experiment 2. Figure 3A shows the mean duration of looking during the habituation phase, and Figure 3B shows the mean duration of looking during the test trials. Error bars indicate the standard error of the mean.

creased infants' sensitivity to A-V temporal relations. This is consistent with the theoretical assumption put forth earlier that infants may respond in this way because they, unlike adults who have had a lifetime of massive experience with A-V temporal synchrony, have not as yet acquired the same degree of perceptual bias or preference for synchronous audiovisual events.

Although this interpretation is a reasonable one, there is an alternative explanation that might account for the better performance obtained in this experiment. It is possible that the better overall discriminative performance obtained in this experiment was due to the fact that infants took more time to encode the asynchronous syllable and thus had more exposure to it. To test this possibility, I compared the number of trials needed to reach the habituation criterion and the overall amount of time that it took the infants to habituate across Experiments 1 and 2. This comparison indicated that the mean number of trials needed to reach habituation in Experiment 1 (7.4) and in Experiment 2 (8.5) did not differ, $t(67) = 1.2$, $p = .22$. The comparison of the duration of looking during the first and last three habituation trials across the two experiments (see Figures 2A and 3A) also did not yield any differences. Specifically, a $2 \times 4 \times 4 \times 6$ (Experiment × Age × Test Trial Order × Trials) mixed, repeated measures ANOVA on the duration of looking during the habituation phase, with Experiment, Age, and Test Trial Order as the between-subjects factors and Trials as a within-subjects factor, indicated that there was an overall significant Trials effect, $F(5, 185) = 46.7$, $p < .001$, $\eta_p^2 = .558$, but that there were no other significant effects; the *F* ratio for the Experiment × Trials interaction was $F(5, 185) = .82$, *ns*. Overall, these comparisons indicate that infants had similar exposure to the two types of syllables in the two experiments and thus rule out the possibility that the different results from Experiments 1 and 2 were due to differential learning rates.

Experiment 3

Experiments 1 and 2 demonstrated that regardless of whether infants were initially exposed to a synchronous or an asynchronous audiovisual speech syllable, they successfully detected changes in the temporal relations between the syllable's audible and visible attributes. These results confirm and further extend prior reports of successful learning and discrimination of A-V speech synchrony relations (Lewkowicz, 2000b, 2003). In addition, they are consistent with prior findings showing that infants are less sensitive to such relations in audiovisual speech events than in audiovisual nonspeech events (Lewkowicz, 1996b).

As indicated earlier, it has been shown in multiple studies that infants are sensitive to A-V synchrony relations in nonspeech events. The current results add to this body of evidence and demonstrate once again not only that infants are sensitive to such relations in speech events but that the degree of their sensitivity to such relations depends on specific prior short-term experience. Together this body of findings raises an interesting question that has so far not been answered in the literature: Do the energy and spectral variations that are typical of the acoustic part of audiovisual speech play a role in infant response to A-V synchrony relations? The earlier cited findings showing that infants can detect A-V synchrony relations inherent in events consisting of moving objects and nonspeech impact sounds suggest that energy and spectral variations are not necessary for infant detection of A-V synchrony relations. Adding to this possibility is evidence reported by Hollich et al. (2005) showing that infants do not need the kinetic information that normally specifies facial speech gestures to segregate auditory speech. That is, infants are able to segregate simultaneous speech streams even when the accompanying visual information is a dynamic synchronous oscilloscope trace. In other words, infants need only the global envelope that defines the dynamic visual information specifying a talking face to segregate two concurrent speech streams.

The fact that infants can detect A-V synchrony relations when simple, nonspeech sounds are presented and that they can segregate simultaneous speech streams with the aid of a simple oscilloscope trace suggests that they may detect A-V synchrony relations primarily on the basis of low-level attributes (e.g., energy onsets and offsets). If that is the case, then the energy and spectral variations that are typical of an acoustic speech signal may not play an important role in infant response to A-V speech synchrony relations. This, in turn, would suggest that the correlation between the energy and spectral variations of the acoustic signal on the one hand and the kinetic energy information inherent in the facial gestures of a speaker on the other do not contribute to infant response to audiovisual speech either. To test these possibilities, in Experiment 3 I habituated and tested infants in the same way as in Experiment 1 except that this time I presented a tone in place of the audible syllable. As a result, here infants heard a sound that had a constant intensity and a linear spectral profile.

Method

Participants. Thirty-five infants completed the experiment. Of these, 5 infants were eliminated from data analysis because 2 exhibited spontaneous regression effects and 3 exhibited fatigue effects. The remaining 30 infants (13 boys and 17 girls) contributed usable data and consisted of separate groups of nine 4-month-olds (mean age = 17.5 weeks, *SD* = 0.9 weeks), eight 6-month-olds (mean age = 26.5 weeks, *SD* = 1.1 weeks), five 8-month-olds (mean age = 34.4 weeks, *SD* = 0.6 weeks), and eight 10-month-olds (mean age = 43.3 weeks, *SD* = 1.1 weeks). Fifteen additional infants were tested but did not contribute data because they fussed.

Apparatus and stimuli. All stimulus materials and apparatus were identical to those employed in Experiment 1. The only exception was that the acoustic part of the syllable was replaced with a 440-Hz tone. The tone's duration was .314 s, its rise and fall times were set at 50 ms (see Figure 1C), and it was presented at 87 dB (A scale).

Procedure. The procedure was identical to that used in Experiment 1. In brief, infants were habituated to the synchronous syllable and then tested with the following asynchrony test trials: FAM 0, NOV 366, NOV 500, and NOV 666. In all cases, the onset of the tone preceded the onset of lip motion.

Results and Discussion

Figures 4A and 4B show the results from the habituation and test trials, respectively. Figure 4A shows that looking declined during the habituation phase. A two-way 4×6 (Age × Trials) repeated measures ANOVA, with Age as the between-subjects factor and Trials as the within-subjects factor, indicated that the Trials effect was significant, $F(5, 130) = 22.2$, $p < .001$, $\eta_p^2 =$

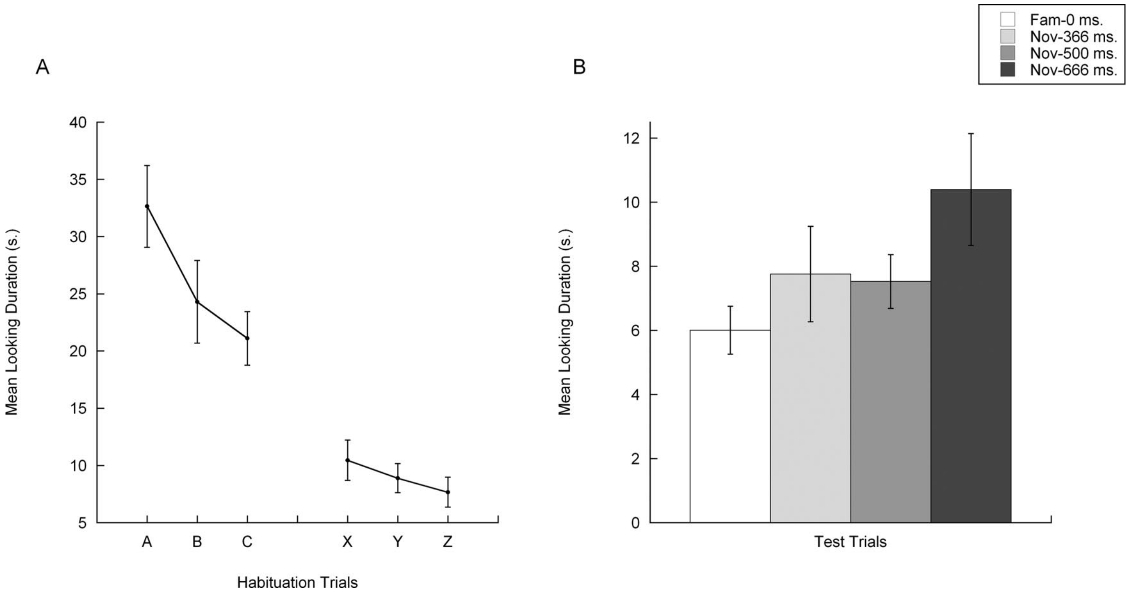

Figure 4. Results from Experiment 3. Figure 4A shows the mean duration of looking during the habituation phase, and Figure 4B shows the mean duration of looking during the test trials. Error bars indicate the standard error of the mean.

.461, and that no other effects were significant. As in the first two experiments, infants decreased their looking during the habituation phase, and this was not affected by their age.

To determine whether infants detected any of the A-V asynchronies, I performed an overall 4 × 4 × 4 (Age × Test Trial Order × Test Trial Type) mixed, repeated measures ANOVA, with Age and Test Trial Order as the between-subjects factors and Test Trial Type as a within-subjects factor. This analysis yielded no significant meaningful main effects or interactions. The planned contrast analyses, comparing the duration of looking in each novel test trial with the duration of looking in the FAM 0 test trial, indicated that infants did not exhibit significant response recovery in the NOV 366 or the NOV 500 test trials but that they did in the NOV 666 test trial, $F(1, 16) = 6.05$, $p = .026$, $\eta_p^2 = .135$. These findings replicate the results from Experiment 1. In addition, they demonstrate that the energy and spectral variations that are typical of an acoustic speech signal and the correlation of these variations with the kinetic information inherent in the facial gesture specifying a speech syllable are not necessary for infant detection of A-V synchrony relations.

General Discussion

The overall aim of the current study was to investigate infant perception of audiovisual speech synchrony. This was accomplished by investigating (a) the size of the ITCW for audiovisual speech, (b) whether short-term experience affects the size of the ITCW, and (c) the mechanism underlying infant response to A-V speech synchrony relations. Through a convergent-operations approach, the size of the ITCW was obtained by habituating infants to an audiovisually synchronous syllable in Experiment 1 or to an audiovisually asynchronous syllable in Experiment 2 and then testing to determine whether infants could detect A-V temporal changes. The changes consisted of increasing levels of A-V asynchrony in Experiment 1 and decreasing levels of A-V asynchrony in Experiment 2. Although infants detected the A-V temporal changes in both experiments, they exhibited greater sensitivity following habituation to the asynchronous syllable, demonstrating that short-term experience has differential effects on infant response to A-V temporal synchrony relations. Finally, Experiment 3 showed that infant perception of audiovisual speech synchrony relations is governed by a domain-general mechanism that is sensitive to the onset and offset of stimulus energy rather than to the variations in acoustic and gestural information.

The Role of Experience in Infant Response to Audio-Visual Speech Synchrony

The results from Experiments 1 and 2 provided different estimates of the ITCW. On the one hand, Experiment 1 suggested that the ITCW for audiovisual speech is relatively wide (666 ms), and on the other hand, Experiment 2 suggested that the ITCW is narrower (i.e., 300 ms). The 666-ms asynchrony threshold found in Experiment 1 is consistent with previous findings from studies in which infants were habituated to a synchronous audiovisual speech utterance and tested with an asynchronous utterance (Lewkowicz, 2000b, 2003). At first blush, the 300-ms asynchrony threshold obtained in Experiment 2 does not appear to be consistent with previous findings. It should be noted, however, that infants do not usually experience asynchronous audiovisual speech, nor for that matter, asynchronous audiovisual events in general. This suggests that the results from Experiment 2 most likely reflect infants' normal perceptual experience, which is biased in favor of synchronous audiovisual events. If that is the case, then it is reasonable to posit that initial exposure to an asynchronous, and thus unusual, event elicits greater attention that is specifically directed at the event's temporal features. This, in turn, leads to better learning of the A-V temporal relations and facilitates subsequent detection of changes in those relations.

The finding of increased sensitivity to A-V temporal relations in infancy following initial exposure to an asynchronous audiovisual event is novel and interesting. Equally interesting, however, is that the effects of such short-term exposure are opposite to what has been reported in adult experiments. In adults, initial adaptation to an asynchronous audiovisual event induces lower sensitivity to A-V temporal relations and causes adults to bind the auditory and visual attributes representing perceptually asynchronous audiovisual events and leads them to perceive such events as synchronous (Fujisaki et al., 2004; Navarra et al., 2005; Vroomen et al., 2004). Expressed in terms of the ITCW, the adult findings demonstrate that short-term experience with A-V asynchrony leads to an increase in the size of the ITCW. In contrast, the current findings demonstrate that short-term experience with A-V asynchrony in a developmentally immature and inexperienced perceptual system leads to a decrease in the size of the ITCW.

Why might short-term experience with an audiovisual event have differential effects at different points in development? The most probable answer is that the effects of short-term experience with audiovisual events are most likely modulated by the effects of long-term experience with such events. As indicated earlier, our everyday perceptual experiences consist of virtually exclusive exposure to synchronous audiovisual events. The one major difference, however, is that infants and adults bring vastly different histories to the A-V synchrony detection task. Whereas infants bring relatively little experience with audiovisual events to an A-V synchrony perception task, adults bring a lifetime of experience to a task such as this. In other words, adults' massive experience provides them with many and varied opportunities to form what Welch and Warren (1980) refer to as the "unity assumption." This assumption is essentially a strong perceptual bias for unified audiovisual events. When such events are not audiovisually unified, adults attempt to unify them anyway. Moreover, this bias is strongest for those audiovisual events that are most frequently experienced. For example, adults are usually exposed to same-gender faces and voices. As a result, they require a greater desynchronization of the face and an accompanying vocalization if the face and vocalization represent a person of the same gender than if the face and vocalization represent a person of a different gender (Vatakis & Spence, 2007). Similarly, because adults are exposed more to biologically meaningful motion and concurrent sounds, their judgments of the correspondence between a periodically moving visual object and concurrent sounds are better when the visual object represents biological motion than when it does not (Saygin, Driver, & de Sa, 2008).

In view of the fact that infants bring far less experience with unified audiovisual events to the A-V synchrony perception task, they do not have a strong unity assumption. As a result, it is not surprising that they exhibit different effects of short-term exposure to an asynchronous audiovisual event than adults. It is also not surprising that they exhibit such different effects despite their relatively impressive intersensory perceptual abilities. Given that infants do not have a strong unity assumption, they tend to focus their attention on the only event feature that is salient—the specific temporal relation between the event's auditory and visual attributes—and encode that relation rather than ignore it. By doing so, they become more sensitive to changes in the magnitude of that temporal relation.

From a developmental standpoint, it could be argued that it is fortuitous that infants do not possess the strong type of unity assumption found in adults because this would be quite maladaptive. The infant's main task is to differentiate his or her multisensory world into distinct but intermodally unified multisensory objects and events, and this occurs through a process of increasing specificity (E. J. Gibson, 1969, 1984). This, in turn, depends on the infant achieving greater, not less, temporal precision and requires developmental narrowing, rather than broadening, of the ITCW. Currently, it is not known how rapidly the ITCW narrows in development except that it does not narrow throughout the first months of life (Lewkowicz, 2000a). Regardless of how rapidly the ITCW narrows in development, there is no doubt that it is more adaptive to possess a perceptual system that is tuned to the specific temporal alignment of multisensory inputs than one that can be overridden by a perceptual bias. In this way, as perceptual learning and differentiation proceed, and as greater perceptual specificity is achieved, infants can avoid the possibility of forming adventitious associations between, say, the face of one of two concurrently talking people and the wrong person's voice.

Mechanism Mediating Infant Perception of Audio-Visual Speech Synchrony

The results from Experiment 3 showed that neither the energy and spectral variations that are typical of an acoustic speech signal nor the correlation between these variations and dynamic gestural cues mediated infant detection of A-V synchrony. Although these cues eventually come to play a role in adult response to audiovisual speech (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Munhall & Buchan, 2004), the current results suggest that they are not necessary for infant response to audiovisual speech synchrony. This, in turn, suggests that what matters more is the energy transitions found at the start and the end of the syllable. That would explain why infants can detect A-V synchrony relations across many types of audiovisual events. As a result, when the current findings are considered together with all the findings cited earlier showing that infants can perceive A-V synchrony across a broad range of audiovisual events, the most reasonable conclusion is that detection of A-V temporal synchrony relations in infancy is mediated by a domain-general mechanism that responds to overall energy onsets and offsets.

Relation Between Intersensory Synchrony Cues and Dynamic Pattern Cues

Several prior studies have shown that infants are able to match the auditory and visual attributes of speech even when A-V temporal synchrony cues are not available (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2002, 2003). These studies have shown that when infants can see two faces uttering different syllables and at the same time hear one of the syllables, they look longer at the syllable that matches the heard syllable even though the onset and offset of the heard syllable corresponds to the onset and offset of both visible syllables. At first blush, these findings might be seen as inconsistent with the current findings. The fact is, however, that audiovisual speech is a dynamic and multidimensional event that can be specified by the synchronous onsets of acoustic and gestural energy as well as by correlated dynamic patterns of acoustic and gestural energy (Munhall & Vatikiotis-Bateson, 1998, 2004; Yehia et al., 1998). As a result, it is entirely possible to respond to audiovisual speech in terms of one but not the other perceptual cue. The studies showing that infants can match faces and voices in the absence of intersensory temporal cues show that young infants can take advantage of one perceptual cue in the absence of the other. In contrast, the findings from the current study show that infants can perceive A-V synchrony cues in the absence of higher level intersensory correlations. Thus, together these findings suggest that infants can respond to both of these cues independently of each other. Under normal ecological conditions, however, intersensory synchrony and dynamic pattern cues are closely linked and probably support one another, and their relationship is probably affected by experience.

To illustrate how the two cues can complement each other but how this can differ as a function of prior experience, I pose two examples: one in which the facial and vocal information is familiar (i.e., the infant's mother, father, or sibling) and one in which the information specifies a stranger. In the case of familiar information, the infant does not need to rely a great deal on temporal synchrony information to determine that the particular voice that he or she hears coming from another room corresponds to a particular person who happens to appear a second or two later. Recognition in this case is based on prior experience of the frequent association of the face and the voice (Bahrick et al., 2005; Brookes et al., 2001) and probably also on the ability to remember the dynamic acoustic pattern heard in the absence of the face, the matching of this pattern to the acoustic pattern heard in the presence of the face, and finally on the ability to match the currently heard acoustic pattern with the visible gestural pattern. If this scenario is correct, then it suggests that even young infants are able, in the absence of A-V synchrony cues, to integrate the audible and visible attributes of speech in a manner similar to adults. That is, like adults, infants appear to be able to extract some sort of global dynamic information from the acoustic and gestural aspects of a familiar speech utterance and then link the two (Kamachi et al., 2003; Munhall & Buchan, 2004). In contrast, in the case of unfamiliar audiovisual information, infants are likely to require a closer temporal contiguity between audible and visible attributes of utterances to perceive them as unitary events. The greater reliance on A-V synchrony cues in the case of unfamiliar talking faces illustrates the critical importance of the effects of early experience in infant perception of audiovisual speech. Generally, infants have greater exposure to multisensory speech than to other types of multisensory events, and audiovisual speech has far greater value and meaning to them than other types of multisensory events. This would suggest that infants should learn and be able to perceive the intersensory unity of talking faces earlier in life than the unity of other types of multisensory events. Indeed, studies support this conclusion. For example, 3-month-old infants can match the visual and audible emotional expressions produced by their mother but not by a stranger, and as in the case of intersensory matching of speech syllables, they can do so even in the absence of A-V synchrony (Kahana-Kalman & Walker-Andrews, 2001).

It thus appears that A-V synchrony cues and other types of intersensory invariance cues (e.g., the correlation between dynamic acoustic and gestural cues in talking faces) play different roles and at different times in development. This is evident in a number of empirical studies. For example, when audiovisual pattern cues (i.e., rhythm) compete with intersensory synchrony cues for infant attention, the pattern cues dominate responsiveness in young infants, but as perceptual differentiation proceeds and as perceptual processing abilities improve, infants become capable of responding to A-V synchrony cues as well as to pattern cues (Lewkowicz, 2003). Similarly, when young infants are presented with an unfamiliar audiovisual event, they respond to it primarily on the basis of A-V temporal synchrony as they attempt to integrate its auditory and visual attributes, whereas older infants no longer do. The most dramatic example of this is the finding that when infants are confronted with nonnative faces and vocalizations (i.e., monkey calls), 4- and 6-month-old infants match them, whereas 8- to 18-month-old infants no longer do (Lewkowicz & Ghazanfar, 2006; Lewkowicz et al., 2008). Presumably, the younger infants integrate on the basis of synchrony because all other perceptual features are unfamiliar to them. In contrast, the older infants do not integrate on the basis of synchrony because their greater perceptual skills enable them to process the higher level perceptual features of the faces and vocalizations, but because these features are unfamiliar, they are unable to extract the relevant intersensory cues. This interpretation is consistent with findings that when faces and voices are more familiar (i.e., human), infants can make the shift away from a reliance on A-V synchrony cues. For example, 5-month-old infants can match the auditory and visual expressions of affect produced by a stranger only when they are presented in synchrony, whereas 7-month-old infants can do so even when they are not presented in synchrony (Walker-Andrews, 1986).

Conclusion

In sum, there is little doubt that intersensory temporal synchrony provides ready-made opportunities for the perception of a unified multisensory world. This is clear from research indicating that human (Bahrick et al., 2004; Lewkowicz & Kraebel, 2004) and avian (Lickliter et al., 2002, 2004) infants are better at learning temporally synchronized multisensory events, that infants respond more rapidly to spatially concordant than discordant auditory and visual stimulation when it is also temporally concordant (Neil et al., 2006), and that adults exhibit greater behavioral and neural responses to temporally and spatially concordant multisensory inputs (Meredith, Nemitz, & Stein, 1987; Stein & Meredith, 1993). The current findings add to this body of evidence by showing that infants' sensitivity to A-V temporal synchrony cues depends on the short- and long-term effects of perceptual experience as well as on the complex interplay between synchrony cues and other, higher level, intersensory cues. In fact, an interesting developmental picture emerges when these findings are combined with the finding that the ability to perform spatial intersensory integration emerges later than the ability to respond to temporal intersensory cues (Neil et al., 2006; Wallace, 2004). This picture suggests that the heterochronous developmental emergence of responsiveness to different types of intersensory relations reflects the operation of heterogeneous intersensory perceptual mechanisms and that this, in part, reflects the effects of perceptual experience (Lewkowicz, 2002). If this conclusion is correct, then it is important that future research investigate how various intersensory cues interact with one another in early development and how the combined effects of perceptual experience and neural development influence responsiveness to various constellations of such cues.

References

Bahrick, L. E. (1988). Intermodal learning in infancy: Learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59,* 197–209.

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face–voice relations. *Developmental Psychology, 41*(3), 541–552.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190–201.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3), 99–102.

Bahrick, L. E., Walker, A., & Neisser, U. (1981). Selective looking by infants. *Cognitive Psychology, 13*(3), 377–390.

Bergeson, T. R., Pisoni, D. B., & Davis, R. A. O. (2005). Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants. *Ear and Hearing, 26*(2), 149–164.

Brannon, E. M., Suanda, S., & Libertus, K. (2007). Temporal discrimination increases in precision over development and parallels the development of numerosity discrimination. *Developmental Science, 10*(6), 770–777.

Brookes, H., Slater, A., Quinn, P. C., Lewkowicz, D. J., Hayes, R., & Brown, E. (2001). Three-month-old infants learn arbitrary auditory-visual pairings between voices and faces. *Infant and Child Development, 10*(1–2), 75–82.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21*(1), 300–304.

Canfield, R. L., & Haith, M. M. (1991). Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology, 27*(2), 198–208.

Colombo, J., & Richman, W. A. (2002). Infant timekeeping: Attention and temporal estimation in 4-month-olds. *Psychological Science, 13*(5), 475–479.

Dixon, N. F., & Spitz, L. T. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719–721.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11*(4), 478–484.

Fraisse, P. (1982). The adaptation of the child to time. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 113–140). New York: Academic Press.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773–778.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Appleton.

Gibson, E. J. (1984). Perceptual development from the ecological approach. In M. E. Lamb, A. L. Brown, & B. Rogoff (Eds.), *Advances in developmental psychology* (pp. 243–286). Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton-Mifflin.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication, 44*(1–4), 43–53.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*(3), 598–613.

Kahana-Kalman, R., & Walker-Andrews, A. S. (2001). The role of person familiarity in young infants' perception of emotional expressions. *Child Development, 72*(2), 352–369.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13*(19), 1709–1714.

Kuhl, P. K., & Meltzoff, A. N. (1982, December 10). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development, 7*(3), 361–381.

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior & Development, 9*(3), 335–353.

Lewkowicz, D. J. (1989). The role of temporal factors in infant behavior and development. In I. Levin & D. Zakay (Eds.), *Time and human cognition: A life-span perspective* (pp. 9–62). Amsterdam: North-Holland.

Lewkowicz, D. J. (1992a). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior & Development, 15*(3), 297–324.

Lewkowicz, D. J. (1992b). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Perception & Psychophysics, 52*(5), 519–528.

Lewkowicz, D. J. (1996a). Infants' response to the audible and visible properties of the human face: I. Role of lexical-syntactic content, temporal synchrony, gender, and manner of speech. *Developmental Psychology, 32*(2), 347–366.

Lewkowicz, D. J. (1996b). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance, 22*(5), 1094–1106.

Lewkowicz, D. J. (1998). Infants' response to the audible and visible properties of the human face: II. Discrimination of differences between singing and adult-directed speech. *Developmental Psychobiology, 32,* 261–274.

Lewkowicz, D. J. (2000a). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281–308.

Lewkowicz, D. J. (2000b). Infants' perception of the audible, visible and bimodal attributes of multimodal syllables. *Child Development, 71*(5), 1241–1257.

Lewkowicz, D. J. (2002). Heterogeneity and heterochrony in the development of intersensory perception. *Cognitive Brain Research, 14,* 41–63.

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology, 39*(5), 795–804.

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences, USA, 103*(17), 6771–6774.

Lewkowicz, D. J., & Kraebel, K. (2004). The value of multimodal redundancy in the development of intersensory perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 655–678). Cambridge, MA: MIT Press.

Lewkowicz, D. J., & Lickliter, R. (Eds.). (1994). *The development of intersensory perception: Comparative perspectives.* Hillsdale, NJ: Erlbaum.

Lewkowicz, D. J., Sowinski, R., & Place, S. (2008). The decline of cross-species intersensory perception in human infants: Underlying mechanisms and its developmental persistence. *Brain Research, 1242,* 291–302.

Lickliter, R., & Bahrick, L. E. (2000). The development of infant intersensory perception: Advantages of a comparative convergent-operations approach. *Psychological Bulletin, 126*(2), 260–280.

Lickliter, R., Bahrick, L. E., & Honeycutt, H. (2002). Intersensory redundancy facilitates prenatal perceptual learning in bobwhite quail (*Colinus virginianus*) embryos. *Developmental Psychology, 38*(1), 15–23.

Lickliter, R., Bahrick, L. E., & Honeycutt, H. (2004). Intersensory redundancy enhances memory in bobwhite quail embryos. *Infancy, 5*(3), 253–269.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. *Journal of Neuroscience, 7*(10), 3215–3229.

Munhall, K. G., & Buchan, J. N. (2004). Something in the way she moves. *Trends in Cognitive Sciences, 8*(2), 51–53.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123–139). Hove, England: Psychology Press.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 177–188). Cambridge, MA: MIT Press.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499–507.

Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science, 9*(5), 454–464.

Nelson, K. (1986). *Event knowledge: Structure and function in development.* Hillsdale, NJ: Erlbaum.

Nelson, K. (2007). *Young minds in social worlds.* Cambridge, MA: Harvard University Press.

Partan, S., & Marler, P. (1999, February 26). Communication goes multimodal. *Science, 283*(5406), 1272–1273.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior & Development, 22*(2), 237–247.

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*(1), 93–115.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour, 58,* 921–931.

Saygin, A. P., Driver, J., & de Sa, V. R. (2008). In the footsteps of biological motion and multisensory perception: Judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychological Science, 19*(5), 469–475.

Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2003). Sound induces perceptual reorganization of an ambiguous motion display in human infants. *Developmental Science, 6*(3), 233–241.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* Cambridge, MA: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255–266.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36,* 314–331.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744–756.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22*(1), 32–35.

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology, 22*(3), 373–377.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: Differentiation of multimodal information. *Psychological Bulletin, 121*(3), 437–456.

Wallace, M. T. (2004). The development of multisensory integration. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 625–642). Cambridge, MA: MIT Press.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Wentworth, N., & Haith, M. M. (1998). Infants' acquisition of spatiotemporal expectations. *Developmental Psychology, 34*(2), 247–257.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1–2), 23–43.

Received June 14, 2008
Revision received February 4, 2009
Accepted February 11, 2009 ■