Child Development, May/June 2005, Volume 76, Number 3, Pages 598–613

# Infants' Use of Synchronized Visual Information to Separate Streams of Speech

George Hollich
*Purdue University*

Rochelle S. Newman
*University of Maryland*

Peter W. Jusczyk
*Johns Hopkins University*

In 4 studies, 7.5-month-olds used synchronized visual–auditory correlations to separate a target speech stream when a distractor passage was presented at equal loudness. Infants succeeded in a segmentation task (using the head-turn preference procedure with video familiarization) when a video of the talker's face was synchronized with the target passage (Experiment 1, $N = 30$). Infants did not succeed in this task when an unsynchronized (Experiment 2, $N = 30$) or static (Experiment 3, $N = 30$) face was presented during familiarization. Infants also succeeded when viewing a synchronized oscilloscope pattern (Experiment 4, $N = 26$), suggesting that their ability to use visual information is related to domain-general sensitivities to any synchronized auditory–visual correspondence.

One of the earliest hurdles for infants acquiring a language is learning to follow one speech stream over others. 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. To understand her mother, the infant must be able to separate her mother's speech from that of the other voices in her environment. Similar situations play out the world over for infants attending day care, infants in communal childrearing situations, infants in muti-generational households—in short, for almost any infant learning a language somewhere other than the confines of an acoustic isolation chamber.

The scientific study of these types of events is related to a wide range of psychological disciplines. Those who study adult speech perception might look at the situation as a classic example of streaming, or what is colloquially called the cocktail party problem (Cherry, 1953). For an adult at a cocktail party, the confusion of voices makes it difficult to attend to a given speech stream and rapidly degrades intelligibility of individual talkers. In contrast, those who study infant speech perception might focus on the difficulty of pulling out words from the fluent stream of speech in the first place, or what is colloquially called the segmentation problem. Words in fluent speech are not marked by spaces, commas, or periods. That infants can succeed in this task at all is remarkable; that they could segment speech when other voices are speaking at the same time would seem all but impossible. Finally, those who study visual speech might emphasize how children could use visual information contained in the mother's face to overcome both of these apparent problems. This study attempts to straddle these disciplines by demonstrating that infants use synchronized visual information to aid them in streaming and thereby help them segment the speech stream at levels of distraction otherwise impenetrable.

## *Streaming*

Over the past 50 years, research has demonstrated that when faced with the buzzing confusion of multiple voices, adults use many kinds of cues to separate one voice from another. These cues include: location in space (Broadbent, 1954; Cherry, 1953; Hirsh, 1950; Pollack & Pickett, 1958), frequency range (Bregman & Pinker, 1978), voice pitch and gender (Broadbent, 1952), and onset times and am-

This research was supported by National Institute of Child Health and Human Development Research Grant 15795 and National Institute of Mental Health Senior Scientist Award 01490 to Peter W. Jusczyk. Much of the data were presented at the fall 2001 meeting of the Acoustical Society of America in Ft. Lauderdale, Florida, and we thank the attendees for their insightful questions and comments. We also thank Ann Marie Jusczyk and Natasha Scheitlin for their assistance in recruiting and testing the children. Although this article was written after the death of Professor Jusczyk, the work was completed with his contribution and support, and he is therefore listed as a full author. This article is dedicated to Professor Jusczyk, our collaborator, mentor, and friend.

Correspondence concerning this article should be addressed to George Hollich, Department of Psychological Sciences, 703 Third St, Purdue University, West Lafayette, IN 47907-2004. Electronic mail may be sent to ghollich@purdue.edu.

plitude modulation (Bregman, Abramson, Doehring, & Darwin, 1985; Dannenbring & Bregman, 1978). This list of cues also includes metacognitive strategies that appear to fill in missing speech sounds (Warren, 1970).

Although virtually no similar studies of infant streaming have been done, most of these cues are likely to be much less useful to infants. First, infants' ability to localize sound is poor. For example, 7-month-olds do not appear to separate sounds that are less than 19 degrees apart (Ashmead, Clifton, & Perris, 1987). This poor localization would likely restrict infants' ability to use spatial location as a cue for separating voices. Second, a variety of studies indicate that infants' skill at auditory discrimination in general is worse than adults. Infants have poorer auditory thresholds for speech (Trehub, Bull, & Schneider, 1981). They require greater intensity levels to discriminate speech sounds in quiet (Nozza, Rossman, & Bond, 1991), and they require even greater intensity levels to discriminate speech sounds in white noise (Nozza, Rossman, Bond, & Miller, 1990). Although white noise is no doubt different from fluent speech as a distractor, it seems plausible that infants would have comparable difficulty discriminating speech sounds when other voices are used as the distractor. These auditory difficulties are problematic because even subtle hearing difficulties can severely blunt the force of acoustic cues to stream segregation. Elderly adults who have normal hearing on pure tone tests have difficulty understanding speech in white noise, suggesting that the ability to separate voices and parse speech may be affected by even very subtle differences in hearing acuity (Bergman, 1971). Finally, infants' lack of experience with the language means that they cannot rely on extensive linguistic knowledge to help them compensate for difficulty in auditory discrimination. Second-language learners, who likewise have less linguistic knowledge, experience considerable difficulties comprehending speech in white noise (Mayo, Florentine, & Buus, 1997; Takata & Nábelek, 1990). Given such acoustic and metacognitive limitations, it is reasonable to suppose that attention to one speech stream over another presents a special problem for infants.

### *Segmentation*

In addition to following one speech stream over another, infants must also be able to recognize and segment meaningful units from that stream. Less than 7% of the speech directed at children is in the form of isolated words (van de Weijer, 1998). Furthermore, this relatively low percentage occurs even when mothers are explicitly directed to teach their children individual words (Aslin, Woodward, LaMendola, & Bever, 1996) and even when they are in noisy listening environments (Newman, 2003). Nonetheless, by the time they are 8 months of age, infants are capable of extracting meaningful units from the acoustic stream (Jusczyk & Aslin, 1995; Echols, Crowhurst, & Childers, 1997). For example, Jusczyk and Aslin (1995) familiarized 7.5-month-old infants to fluent speech passages written around two target words. Infants were later tested on their ability to recognize those words when played in isolation. Infants listened longer to the words that had occurred in the familiarized stories than to words that had not occurred in those stories, suggesting that they had segmented and remembered the target words. Further studies have found that infants are able to use a variety of cues to segment words from fluent speech. For example, statistical probability (Saffran, Aslin, & Newport, 1996), metrical stress (Cutler, 1990), phonotactic cues (Jusczyk, Luce, & Charles Luce, 1994), and many other acoustic speech cues (Johnson & Jusczyk, 2001) have been implicated in 7.5-month-olds' abilities to segment words from the flowing stream of speech.

What happens to these segmentation abilities when other distracting voices are talking? Again, acoustic cues are blunted by the situation described in the opening paragraph. It is likely that the streaming problem makes segmentation all but impossible. Infants cannot segment what they cannot separate from background voices. Newman and Jusczyk (1996) explored the issue of segmentation while segregating voices, using the same stimuli as Jusczyk and Aslin (1995). However, in the Newman and Jusczyk study, the familiarization passages were blended with speech from a distractor voice. The authors found that infants were able to segment words only when these familiarization passages, played at 72 dB, were 10 dB more intense than the simultaneous distractor voice. This is roughly equivalent to the noise level one might experience while having a one-on-one conversation at a reasonably quiet restaurant. It is worth noting that infants in this study could process isolated words at a 5-dB signal-to-noise ratio, although they failed to do so even in that task when the target and distractor passages were of equal loudness. These results are also consistent with other research on infants' ability to discriminate consonants in white noise, which demonstrated that infants could recognize a/ba/–/ga/distinction only at an 8-dB signal-to-noise ratio or higher (Nozza et al., 1990; Trehub et al., 1981).

Together, such studies suggest that infants find it nearly impossible to discriminate speech sounds, let alone segment fluent speech, at signal-to-noise ratios less than or equal to 5 dB.

Unfortunately, this level of interference is less than that typically found in elementary schools (Picard & Bradley, 2001). Although noise-level evaluations have not been done for day care facilities (which would be more relevant to infants), these settings are likely equally loud, or louder, than elementary school classrooms. Likewise, it is a virtual certainty that noise levels in some homes are equally loud. Given such levels of background interference, segmentation of speech by purely auditory means seems unlikely for most infants.

### *Visual Speech*

This research on stream segregation and word segmentation in infants has focused on cues in the auditory speech stream itself. Another potentially valuable source of information appears in the talker's face. Both prosodic and phonetic information are available to children if they are looking at the person who is talking (Kuhl & Meltzoff, 1982; see also Green & Kuhl, 1989, 1991). Deaf individuals can visually perceive some aspects of spoken speech (Bernstein & Demorest, 1993). And although children are notoriously bad at lipreading (Massaro, 1987), by a very early age infants are adept at noticing and acting on synchronous correlations between sight and sound (Bahrick, 1987, 1988, 1992; Lewkowicz, 1986; Lewkowicz & Lickliter, 1994; Meltzoff & Borton, 1979). In this manner, synchrony could be a cue to pay attention to a particular sound stimulus and could thereby aid infants in focusing on one stream of speech. Bahrick and colleagues (Bahrick, 2001; Bahrick & Lickliter, 2000) have suggested that temporal synchrony is one of the most consistent relations to which infants are sensitive. Thus, even 4-week-olds can direct their attention based on synchrony between sound and sight (Bahrick, 2001). Furthermore, 10- to 16-week-old infants direct their attention to speech presented in synchrony with a dynamic video of that speaker's face compared with a video that is not synchronized (Dodd, 1979; Pickens et al., 1994).

This increase in attention to synchronized audiovisual information may confer a special advantage when segregating different streams of speech or when auditory cues are less salient or missing altogether. Even static visual information can enhance adults' attention to an auditory stimulus (see Reisberg, 1978), and several studies have demonstrated that adult listeners are better able to identify speech or speech sounds when they have access to both the visual (e.g., face or face-like) and auditory components of the signal than to either one alone (McLeod & Summerfield, 1990; Rosenblum, Johnson, & Saldaña, 1996; Sumby & Pollack, 1954). In addition, synchronous audiovisual information seems particularly useful to direct auditory attention in the presence of white noise (Grant & Seitz, 2000), and one might expect a similar advantage when faced with competing speech streams. Second-language learners gain particular benefit from the presence of visual information (Reisberg, McLean, & Goldfield, 1987), suggesting that this information may help compensate for weaker lexical knowledge. In addition, the presence of visual information can alter the apparent location of an auditory sound source (Driver, 1996; Massaro, 1998). When two sound streams are presented from the same loudspeaker, the presence of a visual face corresponding to one stream has the effect of pulling apart the two sound sources. If infants are likewise sensitive to such audiovisual (amodal) information, they might experience a similar advantage from the presence of visual information. They might be able to use visual information to help them in speech streaming, thereby helping them successfully segment speech.

### *Summary and Overview of the Current Studies*

In sum, separating streams must be an especially arduous task for infants because it requires both a sensitive auditory system and an ability to attend selectively to a given signal, both of which are still developing in infants (Bargones & Werner, 1994; Nozza et al., 1991; Nozza et al., 1990; Nozza & Wilson 1984; Sinnott, Pisoni, & Aslin, 1983; Trehub et al., 1981). Several studies have suggested that infants are sensitive to relations between auditory and visual information (Kuhl & Meltzoff, 1982, 1984; Kuhl, Williams, & Meltzoff, 1991). For adults, visual information can provide an additional means of directing attention to a speech stream in the presence of white noise (Sumby & Pollack, 1954). Although performance in white noise may not be directly comparable with speech streaming, it seems likely that the presence of visual information may be an equally or especially important factor in the ability of infants to attend to and segment words from speech in the context of a multitalker environment.

The present studies examined whether dynamic, synchronized visual information would improve 7.5-month-olds' abilities to reliably attend to and segment the speech stream when a distractor passage was presented at equal loudness (0 dB signal-to-

noise ratio). We focused on this level of distraction because no prior work has shown that infants could succeed at this level. We focused on this age because 7.5 months is the age when infants first demonstrate an ability to segment fluent streams of speech (Jusczyk & Aslin, 1995). All studies made use of video familiarization. Experiment 1 examined the situation in which infants saw a video display of a woman talking at the same time they heard both her voice and another voice speaking. Experiment 2 compared this situation (with a synchronized video display) with that of an unsynchronized display, whereas Experiment 3 compared these results with familiarization with a static picture. Experiment 4 examined the influence of a novel form of synchronized visual information, an oscilloscope display. Together, these studies examined the extent to which synchronized visual information, even unfamiliar synchrony, can aid in the segregation of speech in noise.

## Experiment 1

This study used a modified version of the head-turn preference procedure used successfully by Jusczyk and Aslin (1995) and Newman and Jusczyk (1996). Stimuli were modeled after those in the Newman and Jusczyk study, but with the addition of video information during the familiarization phase (the test phase was designed to be virtually identical to previous studies and had no video component). During familiarization, the target and distractor signals were presented at a 0-dB signal-to-noise ratio—10 dB poorer than that shown to be the limit of infant segmentation abilities in prior studies (Newman & Jusczyk, 1996). If infants were able to succeed in this difficult task, it would be strong evidence of the importance of visual information in infant stream segregation in noise.

### *Method*

*Participants*. Participants were 30 infants with mean age of 7 months 13 days (range = 7 months 2 days to 7 months 28 days) and an equal number of boys and girls. Two additional participants were excluded as a result of fussiness. All participants were recruited using mass mailings and were from monolingual, English-speaking homes. In this study (and all studies) the distribution of participants was predominantly Caucasian, with less than 10% participation by ethnic or racial minorities.

*Stimuli*. The blended familiarization stimulus was designed to be as close as possible to that in the original Newman and Jusczyk (1996) study with the added visual component. A video recording was created (using a Sony TRV9000 Digital8 Camcorder) displaying a close-up of the face of a Caucasian female speaker of American English as she read four passages in infant-directed speech (an exaggerated, excited manner of speaking that is known to attract infant attention). As in Newman and Jusczyk, each fluent passage was constructed around a target word (either *cup*, *dog*, *bike*, or *feet*; see the Appendix). At the same time, audio recordings of the female speaker's performance were made using a Shure microphone attached to a 12-bit analog-to-digital converter running at a 10-kHz sampling rate and low-pass filtered at 4.8 kHz. The resulting files (in AIFF format) were stored on a VAX station model 3176 computer. The female speaker also produced the four target recordings of the test items, each of which contained 15 repetitions of the target word (e.g., ''cup, cup, cup''), also in infant-directed speech, using the same equipment and settings.

Distractor passages consisted of a male speaker reading the Method section of the original Newman and Jusczyk (1996) study (see the Appendix). The speaker (also a native speaker of American English) read the paper in a monotone manner. This manipulation was done to maintain consistency with the original studies and to minimize the chance that infants would attend to the male voice during familiarization. That is, we wanted to make the male voice distracting but not intrinsically interesting. Prior studies (Fernald & Kuhl, 1987) have shown that infants prefer to listen to the fundamental frequency modifications of infant-directed speech over those of adult-directed speech, and infants may also prefer listening to female voices over male voices. By ensuring that the female voice used an infant-directed style, we increased the likelihood that infants would select that voice as the one to which they chose to attend. In effect, children were given every auditory opportunity to succeed in the task, so that when they failed (as they did in the second and third studies), it was not for lack of trying.

To create the familiarization stimuli, the average intensity levels of the audio recordings of the male and female passages were adjusted using a waveform program on the computer until they were of equal root mean square (RMS) amplitude. Because both recordings were of natural speech, they involve changes in amplitude from point to point in the sentence. Matching RMS amplitudes thus ensures that the average amplitudes for each passage were the same. RMS was chosen as a measure of loudness rather than peak-to-peak amplitude because RMS

provides a better metric of overall loudness (because of the variability of speech) and because we wished to maintain consistency with the original study. The audio files were then transferred to and digitally combined on an Apple Power Mac G4 computer running the EditDV program (by Digital Origin). This blending was done so that the male voice was always speaking whenever the female spoke a target word. In addition, the male passage was trimmed so that the onset and offset of the passage was simultaneous with the onset and offset of the female passage. Next, the video of the female was transferred from the camcorder into the computer using the FireWire interface and the EditDV program. The resultant file was a NTSC digital video file (in DV format) with a frame rate of 29.97 frames per second. Using the EditDV program, this video was then synchronized with the blended audio product and exported back to the camcorder to produce the final videos used in the study. When played, the familiarization passages were all approximately 22 s and were 72 dB in average amplitude. The final familiarization video consisted of two repetitions each of the audiovisual passages for the two familiarized words. The final video was thus 88 s (44-s exposure to each passage across the two 22-s blocks).

*Apparatus and procedure*. The procedure consisted of familiarization and test phases, each administered in a different room and using different setups: one for displaying the video and the other for administering the head-turn preference procedure. During the familiarization phase, the infant was seated on the caregiver's lap approximately 45 in. from a large Sony (56-in.) LCD presentation display attached to the Sony TRV9000 Digital8 camcorder. The audio was played using the built-in speakers on this display. A large plywood partition, painted white, stretching from wall to wall covered all but the screen of the display, the speakers, and the lens of a second identical camcorder used to record infant responding. Small metal grilles, also painted white, covered the speakers, causing them to blend in with the plywood covering. The effect of this setup was to make it seem as if the screen was built into a large white wall.

After the infant was seated comfortably and the parent was blindfolded, the video was played through to completion regardless of infant looking. It should be noted, however, that looking times during this phase were uniformly high across all studies, with an average attention per passage of 28.8 s ($SD = 9.81$). This is comparable to the 30-s familiarization used by Newman and Jusczyk (1996). Immediately after the 88-s video was finished, the infant and parent were escorted into the next room, where the testing phase was conducted. The delay in moving from room to room was usually less than 30 s (as the rooms were adjacent to each other), and the time between familiarization and test was less than 1 min as a result. Although the shift in context likely made the task more difficult, the rationale was that if the infants succeeded despite the change in context, this would be further evidence for the power of audiovisual synchrony.

During the test phase, the infant sat on the caregiver's lap in the center of a three-sided enclosure made from $4 \times 6$ ft pegboard panels. This enclosure was painted white and had a green light mounted at eye level on the wall facing the infant and two red lights mounted on the sides. A white curtain suspended around the top of the booth shielded the infant's view of the rest of the room. An experimenter hidden behind the booth initiated each trial by operating a response box linked to an Apple Power Mac G4 computer, which controlled the selection and randomization of the stimuli. Audio was fed through a Harmon Kardon audio amplifier (HK-3250) to one of two Cambridge Soundworks (Ensemble II) loudspeakers mounted on the opposite walls of this enclosure (hidden from the infants but immediately behind the lights). Both the experimenter and caregiver wore Peltor Aviation 7050 sound-insulated headphones that played masking music to prevent them from hearing the stimulus materials throughout the duration of the experiment.

A trial began with the flashing of the green light on the center panel. When the infant fixated on the green light, it was extinguished and a red light on one of the side panels began to flash. When the infant made a head turn of at least 30 degrees toward the flashing light, the experimenter initiated the speech sample from the loudspeaker under that light. When the infant turned away from the light for at least 2 s, the trial ended and the green center light began to flash, signaling the beginning of a new trial. Information about the direction and duration of head turns and the total trial duration were stored in a data file on the computer. Any time the infant spent looking away (whether it was 2 s or less) was not included when measuring the total listening time.

In this test phase, all infants were exposed to the four recording of the words, presented twice in randomized blocks across the trials. Two of the words thus served as targets and two served as nontargets. Only two test blocks were used because of the concern that after the long familiarization period, infants would become fussy with additional test blocks. It was predicted that if infants had seg-

mented the words during familiarization, they should look longer at the lights when the speech stimuli consisted of the words presented in the passages than when the speech stimuli consisted of the words that were not presented.

Frame-by-frame recoding from the videos was conducted on a subset (one fourth) of participants in this and the following studies for reliability purposes. Correlations between looking time coded from video and the original coding were above 95%.

*Results and Discussion*

Preliminary analysis indicated that participants' gender, as well as which items infants were familiarized with, had no effect on the overall results (in this or any subsequent study). We therefore collapsed across these factors in our final analyses. The mean looking times for familiar and unfamiliar words are presented in the first row of Table 1. An alpha level of .05 was used for all statistical tests. Infants looked significantly longer at the test words that had occurred in the familiarized passage than to the unfamiliar words, $t(29) = 4.39$, $p < .001$. These results indicate that infants successfully segmented the target words from the speech stream, even though the target passages had occurred in the context of noise.

This is the first demonstration that infants are capable of perceiving speech in situations in which the target and distractor passages have the same average amplitude. Comparing these results with those from previous research suggests that the presence of visual information had an important effect on infants' ability to segregate and segment the target speech stream. Furthermore, although one might assume that the presence of infant-directed speech or the salience of a female voice were driving this effect, in the original Newman and Jusczyk (1996) study these were also available, and infants never came close to succeeding at this signal-to-noise ratio. The only difference between our study and theirs was the addition of video to the familiarization phase. Therefore, some property of the video must be driving the effect.

Table 1
*Mean Looking Times in Seconds (SE in Parentheses) for All Experiments (S/N = 0 dB)*

| | Target | Nontarget | Difference |
|---|---|---|---|
| Experiment 1 (moving face) | 10.91 (.59) | 8.93 (.53) | 1.98* |
| Experiment 2 (unsynchronized) | 9.73 (.66) | 10.24 (.62) | −0.51 |
| Experiment 3 (static face) | 9.69 (.56) | 9.29 (.46) | 0.40 |
| Experiment 4 (oscilloscope) | 11.07 (.52) | 9.64 (.64) | 1.43* |

*$p < .05$.

There are two possible reasons for the increased performance as a result of the video in the current study. Synchronized video may have helped infants attend to the correct stream of speech (may have helped them stream)—as we would like to conclude. However, given that dynamic video information is intrinsically interesting, it is possible that it was merely the motion of the video that increased infants' attention. That is, it may be that any moving stimulus, not necessarily one correlated with the audio signal, would have had a similar effect. To test this possibility, Experiment 2 examined whether a situation in which a moving video was not synchronized to the audio stimulus had the same effect of allowing infants to succeed in this task.

## Experiment 2

This experiment familiarized infants with an unsynchronized video display of the female talker speaking the passages. Infants were expected to be unable to segment the speech in this condition, but this condition ensured that any effects seen could not be the result of increased attention due to moving video.

*Method*

*Participants*. Participants were 30 infants with a mean age of 7 months 7 days (range = 6 months 23 days to 7 months 15 days) and an equal number of boys and girls. Only 1 additional participant was excluded as a result of fussiness. All participants were recruited using mass mailings and were from monolingual, English-speaking homes.

*Design, stimuli, apparatus, and procedure*. The design, apparatus, and procedure were the same as in the previous experiment. However, in this experiment, a new unsynchronized display was created for the familiarization video. This display was constructed by switching the videos within the familiarization stimuli, such that the video for one target was played while the audio for the other was played. Thus, an infant might see the video for cup while hearing the audio for dog. In this manner, infants in this experiment saw and heard exactly the same videos as in the previous synchronized experiment, but the videos did not match the audio. This moving display could still attract attention and had exactly the same temporal and acoustic properties as in the

previous study, but the correlation between the video and audio was missing. If this correlation was responsible for infants' success in the previous study, infants should have more difficulty in the current experiment.

*Results and Discussion*

The mean looking times to the familiar and unfamiliar words are presented in the second row of Table 1. A paired *t* test conducted on these looking times was not significant, $t(29) = 0.94$, $p > .05$; infants showed no significant evidence of segmentation. Unsynchronized video during familiarization was not sufficient, by itself, to allow infants to succeed in this task. It is also worth pointing out that the average listening times to the target lists were not shorter in this experiment than in Experiment 1 (in fact, they were longer). Thus, infants had not simply become upset and unwilling to participate in the test trials as a result of having seen the unsynchronized video.

Nonetheless, before we can conclude that synchrony was driving the effect, one more explanation needs elimination. It is possible that in the current unsynchronized experiment infants were disturbed by the unsynchronized display—that the lack of synchronization pushed infants to search elsewhere for the target speech stream and actually made them pay less attention to the target stream during familiarization than they would have otherwise. This may have prevented them from having the opportunity to learn the words. Several studies have suggested that infants dislike attending to visual signals that mismatch what they are hearing. This is the case both for situations in which the auditory and visual information are out of temporal synchrony (Dodd, 1979; Pickens et al., 1994) and for situations in which the information is matched in time but mismatched in content (Kuhl & Meltzoff, 1984; Walton & Bower, 1993).

To examine this issue more fully, we looked at the amount of time infants spent looking at the video in the familiarization phase by examining a subset of infants for whom detailed familiarization data were available ($n = 16$; see Figure 1). Although infants did listen slightly less than in Experiment 1, an average of 33.5 s ($SD = 7.0$) versus 38.5 s ($SD = 6.4$) to each passage (or 67 s vs. 76 s overall), they still listened to each passage longer than did the infants in the Newman and Jusczyk (1996) study (who nonetheless succeeded in this task). Thus, infants appear to have paid attention to the familiarization stimuli long enough to have segmented the words, had they been able to separate the speech of the two talkers. This issue is examined further in Experiment 4.

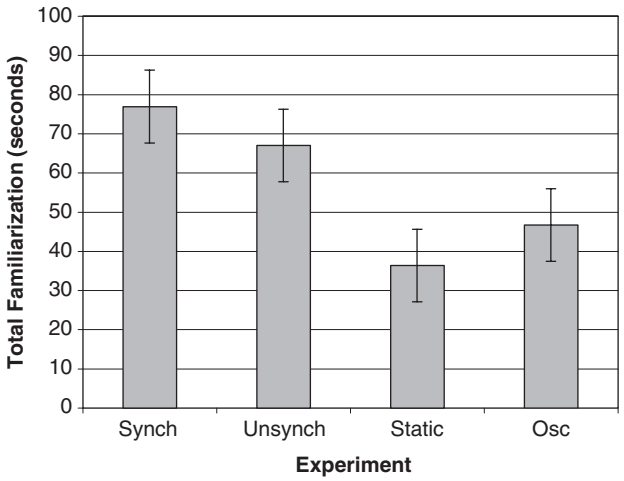


*Figure 1.* Mean looking in seconds during familiarization by experiment (error bars indicate standard deviations). Synch = synchronized; unsynch = unsynchronized; osc = oscilloscope.

However, regardless of familiarization times, it also could be that infants succeeded in Experiment 1 simply because the female face encouraged them to attend to that stream of speech, and the lack of synchrony in Experiment 2 encouraged them to actively attend elsewhere. If this is so, a face that was neither synchronous nor asynchronous (i.e., a static picture of a female face) might likewise encourage infant attention. Infants prefer to look at a female face when they hear a female voice (Walker, 1982), and it would not be surprising if the reverse were also true: Seeing a static female face could facilitate infant attention to the female speech stream. Work with adults has shown some facilitation of attention to a speech stream with the presence of a static face (Reisberg, 1978). For that matter, any nonconflicting visual stimuli could increase the motivation to attend. Although both of these explanations are a kind of visual facilitation, they are very different from the kind of dynamic contribution we would like to conclude. Were either of these explanations true, we would expect infants to succeed if the visual component of familiarization was limited to a static face.

## Experiment 3

Experiment 3 eliminated the dynamic aspect of the familiarization display. If infants' success in Experiment 1 was partially due to increased attention as a result of seeing a female face, infants should succeed in this task when static visual information alone is present. Alternatively, if the dynamic synchronization was the critical aspect of the displays, infants should fail in this task, as they did in Experiment 2.

This experiment was similar to the two previous studies except for the type of visual signal presented. In the present experiment, the video consisted of a static picture of the talker's face. This picture was created by selecting a frame at random from each of the original dynamic stimuli and then freezing that frame throughout the presentation. Given that faces are inherently interesting to infants, and that infants can match female faces with female voices (Walker, 1982), it was possible that infants could use this static visual information alone to succeed in segmentation.

### *Method*

*Participants*. Participants were 30 infants with a mean age of 7 months 10 days (range = 6 months 24 days to 7 months 27 days) and an equal number of boys and girls. Only 1 additional participant was excluded as a result of fussiness. All participants were recruited using mass mailings and were from monolingual, English-speaking homes.

*Design, stimuli, apparatus, and procedure*. The design, apparatus, and procedure were the same as in the previous experiment. However, in this experiment, a new display was created for the familiarization video. It displayed one static frame of the female while the blended audio stimulus played. This static display could still attract attention, without the potentially damaging effects of an unsynchronized display.

### *Results and Discussion*

The mean looking times to the familiar and unfamiliar words are presented in the third row of Table 1. A paired $t$ test conducted on these looking times was not significant, $t(29) = 0.16$, $p > .05$, indicating that infants had not segmented the words in the same way they had in Experiment 1. That is, infants showed no significant evidence of segmentation when familiarized with a static display. However, an examination of the familiarization times from a subset of infants (not all familiarization trials were available for analysis) suggests that infants were paying considerably less attention to the static display (total familiarization: $M = 36$ s, $SD = 13.32$; per word: $M = 18$ s, $SD = 6.66$) then to the unsynchronized display from Experiment 2 (total familiarization: $M = 67$ s, $SD = 14.02$; see Figure 1). Thus, it is possible that infants simply did not obtain enough familiarization time or that boredom was partially responsible for the lack of significant results. Experiment 4 explores the issue of familiarization and asks whether infants' success in the first task was merely due to length of familiarization or specific experience with faces, or whether their success may be the result of a more general process of audiovisual integration.

## Experiment 4

The previous studies apparently demonstrate that infants can use synchronized visual information to help them segregate different streams of speech. There are several possible accounts for how they accomplish this task. First, specific experience with faces may help infants connect what they see with what they hear at a very early age. Indeed, Spelke and Owsley (1979) found that by 3.5 months of age, infants associate the sound of their mother's voice with the sight of her face. Furthermore, Bahrick, Netto, and Hernandez-Reif (1998) have demonstrated that even 4-month-olds can match new voices and faces on the basis of the talker's age, distinguishing readily between children and adults. Similarly, by 5 months of age, infants will preferentially watch a face that matches the affect of the voice they are hearing (happy or sad; Walker, 1982). For speech perception, Kuhl and Meltzoff (1982, 1984) have found that 18- to 20-week-olds are already sensitive to the link between certain vowels and mouth position. Given such evidence, it is reasonable to expect that experience, and sensitivities to faces in particular, may help infants succeed in this task.

Alternatively, it is possible that infant sensitivities to temporal synchrony are so strong that any synchronized visual stimulus would be sufficient to produce the benefit in this task. Perhaps infants' successful performance with the synchronized face display 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. As evidence that such integration in adults may not be limited to feature-specific face information, Rosenblum and Saldaña (1996) found improvement in phoneme recognition using point-light faces (in which one can only see the kinematics of movement) over performance with auditory stimuli alone.

For infants, auditory–visual integration has also been shown for visual events other than faces. For example, 4-month-old infants recognize the correspondence between the sight of a bouncing object and a sound (Spelke, 1979), 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

multimodal information (also called amodal information) attracts significant infant attention. However, there has been no evidence that infants integrate an auditory speech signal with a visual signal other than a face. Speech is a much more complicated acoustic event than are most of the signals tested in studies of infants' auditory–visual integration. Thus, integrating a speech signal with a visual stimulus may be the result of particular experience with auditory–visual correspondences.

Experiment 4 attempts to disentangle whether the results from Experiment 1 are the result of face-specific processing and experience, or whether they are the result of domain-general sensitivities to amodal invariants. To address this issue, we changed the video familiarization to be a moving oscilloscope pattern. 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 residual face-specific effects. Despite this difference, however, oscilloscope patterns maintained a close correspondence to the auditory signal, allowing us to distinguish effects of audiovisual correspondence from effects of facial information per se. As an added benefit, we expected that the oscilloscope display would prove less interesting during familiarization than an unsynchronized face, thereby allowing us to ascertain whether differences in familiarization time were causing the effects observed.

### *Method*

*Participants*. Participants were 26 infants with a mean age of 7 months 10 days (range = 7 months 1 day to 7 months 28 days) and an equal number of boys and girls. Again, only 1 participant was excluded because of fussiness. All participants were recruited using mass mailings and were from monolingual, English-speaking homes.

*Design, stimuli, apparatus, and procedure*. The design, apparatus, and procedure were the same as in the previous experiment. However, in this experiment, a new display was created for the video familiarization. The waveform of the female passages across a 30-ms running window was played (using Harrier-Soft's AmadeusII software), video-recorded (via camcorder), and subsequently synchronized with the blended audio in the manner described in Experiment 1. This process 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, the correlated motion of the line would be expected to cue infants into the female talker's stream. If the effects in Experiment 1 were a result of infants' particular experience with faces, infants would be expected to fail on this task.

### *Results and Discussion*

The mean looking times to the familiar and unfamiliar words are presented in the fourth row of Table 1. Infants listened significantly longer to words that had occurred in the target passage than to words that had not, demonstrating successful segmentation of those words, $t(25) = 2.28$, $p < .05$. Furthermore, the familiarization times were low, much lower than those in Experiment 2 (an average of 23 s per word, $SD = 6.61$, or 46 s, $SD = 13.23$, for total familiarization; see Figure 1). Thus, infants showed evidence of segmentation despite minimal familiarization and minimal face-specific information.

This raises the question of what visual information infants were attending to in this task. The oscilloscope shares several features with faces that could have allowed infants to succeed. Many of the most salient visual cues to speech track the amplitude of the acoustic signal. For example, the lips open the widest on the vowel of each syllable (Grant & Seitz, 2000). Like a face, the oscilloscope showed the widest amplitude discursions synchronous with syllables. Infants may have been cuing in to this movement and using it in a similar manner to lip movement to hone in on the target speech stream. If this is true, infants should fail if changing colors or a blinking light signaled the synchrony instead. Although future studies will examine this issue, it is important to note that even if infants were using the cue of vertical movement, they succeeded in this task with unfamiliar visual circumstances and none of the standard cues researchers talk about when discussing face perception (Cassia, Turati, & Simion, 2004; Johnson & Morton, 1991; Nelson, 2001).

Thus, it appears that the presence of a synchronized oscilloscope pattern was sufficient to allow infants to succeed at this segmentation task at a signal-to-noise ratio approximately 10 dB lower than that at which they would be expected to succeed without this visual information. This suggests that it is 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 was enough

to allow them to segment the speech stream in a noisy and often ambiguous acoustic environment.

This information is unlikely to help infants directly segment the stream but rather likely allows them to hone in on the target speech stream and then segment it by specific acoustic cues. This makes good sense, ontogenetically. Given the number and abstract nature of possible audio and visual combinations, it seems advantageous to start with a system that first pays attention to any audiovisual synchrony and then notices and acts on the consistencies found there. Indeed, even songbirds show enhanced learning when audio and visual stimuli (a strobe light) are synchronized (Hultsch, Schleuss, & Todt, 1999). Like the oscilloscope display for our infants, birds rarely encounter a strobe light outside the confines of a laboratory, yet the synchrony between it and birdsong is apparently highly salient. We suggest that the mechanism for detection of synchrony is likely primitive and is present in a wide range of species (see Bahrick & Lickliter, 2000, for a more detailed discussion of this possibility). From this initial synchrony detection mechanism, infants (and other organisms) could then develop a more sophisticated understanding of the relationship (as in Turati, 2004) between sight and sound.

## General Discussion

Across the four experiments, 7.5-month-olds were shown to use auditory-visual correspondences to separate two different streams of speech at signal-to-noise ratios lower than those previously demonstrated. In Experiment 1, infants used a dynamic visual display of a talker to attend to and segment the fluent speech stream of a female speaker when presented at the same average loudness as a male distractor voice. Infants did not succeed in this task if familiarized with an unsynchronized (Experiment 2) or static video display of that speaker's face (Experiment 3), implying that it was the synchronization that produced the effect. In Experiment 4, infants succeeded in this stream segregation task when the familiarization display was one with which they were unlikely to have had any prior experience (a moving oscilloscope pattern). These results suggest that infants gained a significant advantage by having synchronized visual information complement the auditory stream in noise.

Taken together, the results of the four experiments suggest that it was the motion of the video stimulus and its correspondence to the auditory signal that enabled infants to succeed at a 0-dB signal-to-noise ratio. Specifically, only when the visual display was synchronized (either in the synchronized face condition or the oscilloscope condition) did infants significantly prefer the familiar words during the test phase. These results suggest that infants in Experiments 2 (unsynchronized face) and 3 (static face) could not attend to and segment the speech stream at a 0-dB signal-to-noise ratio.

Potentially more telling than the result of any single study, however, is a comparison across the results from all four experiments. After all, nonsignificance in Experiments 2 and 3 does not necessarily mean that infants gained a significant advantage in the synchronized experiments (1 and 4). For example, one could imagine a case in which the results of two unsynchronized experiments just failed to reach significance whereas the ''successful'' experiments just barely reached significance. In such a case, the actual difference in performance between these experiments could be negligible. That was not the case in these studies, however. An analysis of variance (ANOVA) on the looking times from the four experiments showed a significant interaction between the factors of familiarization and display type, $F(3, 116) = 4.82$, $p < .01$. Fischer's post hoc results indicated that this difference was in the synchronized face condition differing significantly from the static condition ($p = .02$) and unsynchronized condition ($p < .001$), and in the oscilloscope condition differing significantly from the unsynchronized condition ($p < .01$). The oscilloscope condition did not differ significantly from the static condition although there was a trend in that direction ($p = .14$). Infants thus gained a significant advantage by having synchronized face information complement the audio over unsynchronized or static displays. By 7.5 months, infants appear capable of using the correspondences found in the talker's face to disambiguate the speech streams and to segment speech in situations that otherwise would have been beyond their ability. They also can use any synchronized visual information to aid them over unsynchronized or conflicting information.

### *Amount of Advantage*

The exact gain infants incurred by having access to dynamic visual information is not clear. Newman and Jusczyk (1996) found that infants were at threshold performance for auditory-only stimuli at a signal-to-noise ratio of+10 dB. Although direct comparisons with previous studies are difficult, the fact that this experiment used participants of the same age, the same test passages, and the same testing procedure as Newman and Jusczyk (1996)

makes us fairly confident that comparison is possible. Because infants in the present study performed the identical task at signal-to-noise ratios of 0 dB, it appears that the presence of visual information led to an effective improvement in the signal-to-noise ratio of 10 dB. It should be noted that infants did not succeed in pilot studies ($N = 16$) at a-5-dB signal-to-noise ratio. This makes us think that 0 dB is the limit at which infants could succeed in this task.

Adults have been shown to gain an approximately 15-dB advantage for the presence of dynamic visual information (Sumby & Pollack, 1954) in white noise. However, it is important to note that other researchers have reported wide variations depending on the particular passages and the participant's experience at lipreading. For example, Macleod and Summerfield (1990) found that adults showed from a 3- to 22-dB improvement depending on the sentence and from a 6- to 15-dB improvement depending on the participant's skill at lipreading. Because infants are less experienced at visual phoneme distinctions and have poorer auditory perception than adults, we would expect them to fall on the lower side of such estimates. Nonetheless, with a 10-dB improvement, it seems that the advantage infants show for visual information compares favorably with that demonstrated by adult listeners. It is also worth noting that the adult studies were conducted using white noise. Separating speech streams is in some sense a very different task from recognizing speech in noise and may even involve separate perceptual mechanisms (Hygge & Ronnberg, 1992; Jones, Alford, Bridges, Tremblay, & Macken, 1999).

### *Source of the Advantage*

Although it is clear that infants benefited from synchronous audiovisual information during familiarization, the source of this effect is less clear. It is possible that infants used audiovisual information to aid purely in segmentation or purely in streaming. Infants could have used audiovisual information purely as an attentional aid, or they could have used some combination of these methods.

With regard to use of visual information for segmentation, there is evidence that visual information and auditory information are integrated during perception (Gibson, 1969) and that visual information can alter the perceived phonemic representation of a signal (Green et al., 1991; McGurk & McDonald, 1976). For example, when presented with a visual display of a person saying ''ga'' and the auditory stimulus of that person saying ''ba,'' college-age participants typically report hearing a fusion of the two syllables, ''da'' (McGurk & McDonald, 1976). Furthermore, Kuhl and Meltzoff (1982, 1984; see also Patterson & Werker, 1999) demonstrated that 18- to 20-week-olds do something similar when they recognize the correspondence between the shape of a speaker's mouth and the resultant vowel. When shown video presentations of two different speakers, one producing the vowel/i/(with a spread mouth) and the other the vowel/a/(with an open mouth), infants tended to watch the video that matched the sound they heard, although there is some evidence that this matching ability is specific to the infants' native language (Dodd & Burnham, 1988). Similarly, 5-month-olds have been shown to integrate visual and auditory information in this manner (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983; Rosenblum et al., 1996; Rosenblum, Schmuckler, & Johnson, 1997). Moreover, infants prefer novel, but possible, face–voice pairings over novel, but impossible, pairings (Walton & Bower, 1993). All this suggests that infants not only have some knowledge of the relationship between sound and articulation but that speech representations are not necessarily limited to the auditory modality. Thus, infants could have used visual information to disambiguate confusable phonemes in speech processing, which in turn directly aided segmentation in this task.

However, there are several reasons to believe that this was not the source of advantage in the present results. First, infants in the present study demonstrated nearly the same advantage of visual information for an oscilloscope pattern as for a human face (although there was a trend toward greater improvement in the synchronized face display). Given their lack of experience with oscilloscopic displays, it is unlikely that infants were using specific visual representations of speech to disambiguate potentially confusable items. Although we admit the possibility that some improvement specific to segmentation could show up in a more fine-grained analysis or with greater power added to the studies, it is clear that the large differences seen between the synchronized experiments and the static and unsynchronized experiments could not have been due to segmentation-specific effects.

Second, previous research has suggested that children are far less adept at lipreading than are adult listeners (Massaro, 1987). This, too, implies that infants (who are even younger) would be less able than adults to use visual information to identify specific phonemes, much less segment words from speech. In this manner, although it is technically possible that dynamic visual information helped infants identify particular phonemes within the

speech stream (and thereby aided segmentation directly), it is more likely that this information helped infants segregate the different streams of sound or to attend selectively to the target speech stream, and this increase in segregation ability or attention allowed for better segmentation.

We next consider the possibility that some of the effects in the current study were driven by attentional factors during familiarization. Dodd (1979) demonstrated that infants listen longer to synchronized passages than to unsynchronized passages, and it has long been known that dynamic visual information attracts more attention than static information. Together this leads to the possibility that the synchronized displays increased looking times during familiarization, which then led to subsequent successful segmentation. Although detailed familiarization data are unavailable for all children, the data available (see Figure 1) suggest that infants did listen, on average, 5 s longer per passage during the familiarization phase of the synchronized face experiment ($M = 38.5$ s, $SD = 6.4$) than during the unsynchronized face experiment ($M = 33.5$ s, $SD = 7.0$); however, in the static task the infants listened on average only 18.8 s ($SD = 6.6$), and in the oscilloscope task the familiarization times averaged only 23.4 s ($SD = 6.6$). Although these data confirm that static displays are less interesting than moving displays and that moving faces are considerably more interesting to children than wiggly lines on a screen or static faces, they also confirm that added attention during familiarization is not enough to explain the results. Infants paid more attention during the unsynchronized task than during the oscilloscope task, yet they failed to segment the stream despite the additional time.

In essence, given the dramatic differences in test trial performance between the oscilloscope experiment and the unsynchronized experiment, it is highly unlikely that a few additional seconds of familiarization were the sole cause of successful performance in the synchronized tasks. Indeed, in the initial (synchronized face) study, except for infants who fussed out, there was no correlation between familiarization times and successful segmentation. Consistent with this viewpoint, looking only at infants who had familiarization times greater than 38 s per passage during the unsynchronized study (the average amount of time infants in the first study attended to the familiarization stimuli), we found the worst average performance ($M = 10.38$, $SD = 5.06$, to the target vs. $M = 14.58$, $SD = 6.54$, to the nontarget). Thus, although one could do a study that forces maximum familiarization times for the unsynchronized and static experiments, this has little likelihood of succeeding.

Instead, we believe it was infant sensitivity to temporal synchronies between visual and auditory displays that helped them segregate the different speech streams. That is, the results of Experiment 4 (using an oscilloscope display) argue against the notion that the audiovisual processing in this task was experience dependent or that faces and speech were a special combination because of their biological relevance. Over time, infants surely develop domain-specific knowledge (e.g., between faces and sound). However, in the current studies, given low familiarization times in the oscilloscope display and the lack of any correlation between familiarization times and successful segmentation, it is more likely that it was the audiovisual synchrony alone that allowed infants to segregate successfully the speech streams. More specifically, the movement of the visual display likely cued the infants to attend to certain aspects of the auditory signal, aspects that were most strongly correlated with the visual. In short, synchrony drove attention.

This use of synchrony to attend to a masked stimulus is similar to the phenomenon of comodulation masking release. In this phenomenon, an audio signal outside a narrow band of noise is used to cue participants to signals hidden within that noise (Nelken, Rotmani, & Yosef, 1999). Again, the synchrony (or comodulation) between the two signals is the critical cue. Bahrick (2004) has found a similar importance of redundancy within the visual domain (for infants). Although both of these involve different neural pathways, we suggest that the neural computations used to link synchronous activity are similar (if not identical). Thus, we suggest that infants begin life sensitive to any synchronous multimodal information, be it sight and sound, touch and sight, and so forth, and from these domain-general sensitivities, domain-specific specializations develop. Regardless of the ultimate viability of this hypothesis, from these results it is clear that synchronized audiovisual information alone can be used to aid segregation and segmentation of speech.

### *Areas for Future Study*

There are several areas for future study in addition to the theoretical areas outlined earlier. One question is whether infants would still use visual information without the distraction of background noise. If the benefit incurred by visual information is solely in allowing infants to segregate the target stream from other streams, then visual information

would not influence infant performance in easier listening environments. However, visual information could be useful in other ways. Although infants are adept at segmenting consonant–vowel–consonant words from fluent speech at 8 months, they are not as adept at segmenting other types of words. In particular, words with a weak–strong stress pattern and words beginning with vowels are particularly problematic for infants to segment (Jusczyk, 1998; Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999). Future work will examine whether infants are capable of succeeding in these more difficult segmentation tasks when they are familiarized with audiovisual speech rather than with auditory speech alone, or if younger infants might succeed in the segmentation task if audiovisual information is presented.

Another potential area for exploration concerns infants' ability to use visual information as an aid to auditory localization. Localization is an important skill, as it aids in separating streams of speech. Infants' ability to localize auditory information is poor relative to adult performance (Ashmead et al., 1987). Whereas adults can detect differences in location in the range of 1 to 2 degrees of angle, Ashmead et al. (1987) found that infants (ages 26 to 30 weeks) were only able to discriminate sound displacements of approximately 19 degrees. This suggests that infants likely experience situations in which speakers' voices appear to come from locations that are indistinguishable from each other. In such cases, visual information may help infants calibrate their own auditory localization apparatus (see also Aronson & Rosenbloom, 1971).

Even for adult listeners, for whom auditory localization skills are precise, visual information regarding the location of a sound source can lead to an illusory mislocation of the apparent source. This is known alternately as the ventriloquism effect, when used specifically for audiovisual relations, or the visual capture effect more generally (Hay, Pick, & Ikeda, 1965; Jack & Thurlow, 1973; Mateeff, Hohnsbein, & Noack, 1985). The effect is such that adults perceive the source of the sound as coming from the visual stimulus, even though the actual locations are discrepant. For speech segregation, this illusion can enhance listeners' selective spatial attention to speech sounds (Driver, 1996). That is, when two auditory signals come from the same location, the presence of a visual signal elsewhere pulls the matching auditory signal away from the distractor, aiding in segregation abilities. If infants are likewise susceptible to this ventriloquism effect, visual information may be a particularly potent cue to aid in segregation abilities.

Similarly, it is also unclear whether infants would be capable of succeeding in the present stream segregation task if the video displayed the male (distractor) voice rather than the target voice. If visual information serves primarily as an aid to segregating the two speech streams, the presence of a video display of the distractor voice would likely improve infants' performance on words in the target voice over their performance for an auditory signal alone. On the other hand, the presence of such a video might direct infants' attention to the incorrect stream, thus impairing their later recognition of words from the target stream (while improving their later recognition of words from the distractor stream). Similarly, it is not known what infants would do if a male face was synchronized with the female stream. Would the gender mismatch override any beneficial effects of synchrony?

One additional possibility is that this task may provide a means of comparing the strength of different infant listening preferences. That is, infants tend to prefer listening to female voices over male voices and to infant-directed speech over adult-directed speech. We suspect that they also prefer attending to audiovisual stimuli than to auditory-only stimuli. The relative strength of such preferences is not known, however. Given an auditory signal of a female voice speaking in an infant-directed speaking style, and a monotonic male speaker for whom audiovisual stimuli are present, to which signal would infants attend? Presenting infants with multiple voices simultaneously could provide a means of testing trade-offs among these preferences.

Finally, the present task provides a means of exploring the temporal limits of auditory and visual integration by young infants. Recent work suggests that adults integrate information from auditory and visual streams across a wide range of temporal asynchronies (ranging from 100 ms auditory lead to 233 ms auditory lag). Moreover, the best integration takes place when the visual signal leads the auditory signal by a small amount (van Wassenhove, Grant, & Poeppel, 2001, 2002). Infants are also sensitive to temporal correspondence between auditory and visual signals. For example, Dodd (1979; see also Pickens et al., 1994) found that 3- to 4-month-olds attend longer to audiovisual speech that is in synchrony than to asynchronous situations greater than 400 ms. However, if integration skills are still developing in the young infant, we might expect that infants would tolerate larger temporal asynchronies than would adults. By presenting the auditory and visual stimuli from Experiment 1 at different onset asynchronies,

we could examine this issue in more depth. Given the results from Experiment 2, we predict that infants would only succeed at the present segmentation task when the auditory information and visual information were integrated during perceptual processing. When the temporal asynchronies became too great, infants would no longer experience the signal-to-noise ratio advantage caused by the presence of visual information.

Regardless of the outcome of such future studies, the current results make it clear that synchronized video helps infants separate streams of speech and thereby segment the speech stream. In addition, given that infants often find themselves in situations more noisy and complex than the acoustic isolation chambers of traditional infant testing, this work adds to our understanding of how infants segment speech and learn a language in more ecologically realistic situations.

## References


Aronson, E., & Rosenbloom, S. (1971). Spatial co-ordination of auditory and visual information in early infant perception. *Science*, *1*, 1161–1163.

Ashmead, D. H., Clifton, R. K., & Perris, E. E. (1987). Precision of auditory localization in human infants. *Developmental Psychology*, *23*, 641–647.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development*, *10*, 387–416.

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. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Child Psychology*, *53*, 197–209.

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, *79*, 253–270.

Bahrick, L. E. (2004). The development of perception in a multimodal environment. In G. Bremner & A. Slater (Eds.), *Theories of infant development* (pp. 90–120). Malden, MA: Blackwell.

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

Bahrick, L. E., Netto, D., & Hernandez-Reif, M. (1998). Intermodal perception of adult and child faces and voices by infants. *Child Development*, *69*, 1263–1275.

Bargones, J. Y., & Werner, L. A. (1994). Adults listen selectively; infants do not. *Psychological Science*, *5*, 170–174.

Bergman, M. (1971). Hearing and aging. *Audiology*, *10*, 164–171.

Bernstein, L. E., & Demorest, M. E. (1993). Speech perception without audition. *Journal of the Acoustical Society of America*, *94*, 1887.

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception and Psychophysics*, *37*, 483–493.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*, 19–31.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*, 51–55.

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*, 191–196.

Cassia, V. M., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science*, *15*, 379–383.

Cherry, E. C. (1953). Some experiments in the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In Altmann, G. T. M. (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 105–121). Cambridge, MA: MIT Press.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones. *Perception and Psychophysics*, *24*, 369–376.

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

Dodd, B., & Burnham, D. (1988). Processing speechread information. *Volta Review: New Reflections on Speechreading*, *90*, 45–60.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lipreading. *Nature*, *381*, 66–68.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, *36*, 202–225.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*, 279–293.

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

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197–1208.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45*, 34–42.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 278–288.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*, 524–536.

Hay, J. C., Pick, H.L, & Ikeda, K. (1965). Visual capture produced by prism spectacles. *Psychonomic Science*, *2*, 215–216.

Hirsh, I. J. (1950). The relation between localization and intelligibility. *Journal of the Acoustic Society of America*, *22*, 196–200.

Hultsch, H., Schleuss, F., & Todt, D. (1999). Auditory-visual stimulus pairing enhances perceptual learning in a songbird. *Animal Behavior, 58*, 143–149.

Hygge, S., & Ronnberg, J. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech & Hearing Research*, *35*, 208–215.

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the ''ventriloquism'' effect. *Perceptual and Motor Skills*, *37*, 967–969.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 1–20.

Johnson, M. H., & Morton, J (1991). *Biology and cognitive development: The case of face recognition*. Oxford, England: Blackwell.

Jones, D., Alford, D., Bridges, A., Tremblay, S., & Macken, B. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 464–473.

Jusczyk, P. (1998). Constraining the search for structure in the input. *Lingua*, *106*, 197–218.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1–23.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Preference for the predominant stress patterns of English words. *Child Development*, *64*, 675–687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Jusczyk, P. W., Luce, P. A., & Charles Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

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

Kuhl, P. K., & Meltzoff, A. N. (1984). Intermodal speech perception. *Infant Behavior and Development*, *7*, 361–381.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using non-speech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 829–840.

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

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

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, *219*, 1347–1349.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Mateeff, S., Hohnsbein, J., & Noack, T. (1985). Dynamic visual capture: Apparent auditory motion induced by a moving visual target. *Perception*, *14*, 721–727.

Mayo, L., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language & Hearing Research*, *40*, 686–693.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices: A new illusion. *Nature*, *264*, 746–748.

McLeod, A., & Summerfield, A. Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*, 29–43.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature*, *282*, 403–404.

Nelken, I., Rotmani, Y., & Yosef, O. B. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, *397*, 154–157.

Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development*, *10*, 3–18.

Newman, R. S. (2003). Prosodic differences in mothers' speech to toddlers in quiet and noisy environments. *Applied Psycholinguistics*, *24*, 539–560.

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

Nozza, R. J., Rossman, R. N. F., & Bond, L. C. (1991). Infant-adult differences in unmasking thresholds for the discrimination of consonant-vowel syllable pairs. *Audiology, 30*, 102–112.

Nozza, R. J., Rossman, R. N. F., Bond, L. C., & Miller, S. L. (1990). Infant speech-sound discrimination in noise. *Journal of the Acoustical Society of America*, *87*, 339–350.

Nozza, R. J., & Wilson, W. R. (1984). Masked and unmasked puretone thresholds of infants and adults: Development of auditory frequency selectivity and sensitivity. *Journal of Speech and Hearing Research*, *27*, 613–622.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5 -month-old infants. *Infant Behavior and Development*, *22*, 237–247.
Picard, M., & Bradley, J. (2001). Revisiting speech interference in classrooms. *Audiology*, *40*, 221–244.
Pickens, J., Field, T., Nawrocki, T., Martinez, A., Soutullo, D., & Gonzalez, K. (1994). Full-term and preterm infants' perception of face-voice synchrony. *Infant Behavior and Development*, *17*, 447–455.
Pollack, I., & Pickett, J. M. (1958). Stereophonic listening and speech intelligibility against voice babble. *Journal of the Acoustic Society of America*, *30*, 131–133.
Reisberg, D. (1978). Looking where you listen: Visual cues and auditory attention. *Acta Psychologica*, *42*, 331–341.
Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–114). Hillsdale, NJ: Erlbaum.
Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech & Hearing Research*, *39*, 1159–1170.
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.
Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*, 347–357.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
Sinnott, J. M., Pisoni, D. B., & Aslin, R. N. (1983). A comparison of pure tone auditory thresholds in human infants and adults. *Infant Behavior & Development*, *6*, 3–17.
Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15*, 626–636.
Spelke, E. S., & Owsley, C. J. (1979). Intermodal exploration and knowledge in infancy. *Infant Behavior and Development*, *2*, 13–27.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.
Takata, Y., & Nábelek, A. K. (1990). English consonant recognition in noise and in reverberation by Japanese and American listeners. *Journal of the Acoustical Society of America*, *88*, 663–666.
Trehub, S. E., Bull, D., & Schneider, B. A. (1981). Infants' detection of speech in noise. *Journal of Speech and Hearing Research*, *24*, 202–206.
Turati, C. (2004). Why faces are not special to newborns: An alternative account of the face preference. *Current Directions in Psychological Science*, *13*, 5–8.
van de Weijer, J (1998). *Language input for word discovery.* Nijmegen, Netherlands: Wageningen, Ponsen & Loijen.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2001, November). *Timing of auditory-visual integration in the McGurk effect*. Paper presented at the Society of Neuroscience annual meeting, San Diego, CA.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2002, April). *Temporal integration in the McGurk effect*. Paper presented at the Cognitive Neuroscience annual meeting, San Francisco, CA.
Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology, 33*, 514–535.
Walton, G. E., & Bower, T. G. R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development*, *16*, 223–243.
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.

## Appendix

### *Six-Sentence Passages*

*Cup*

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

*Dog*

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

*Feet*

The feet were all different sizes. This girl has very big feet. Even the toes on her feet are large. The shoes gave the man red feet. His feet get sore from standing all day. The doctor wants your feet to be clean.

*Bike*

His bike had big black wheels. The girl rode her big bike. Her bike could go very fast. The bell on the bike was really loud. The boy had a new red bike. Your bike always stays in the garage.

### *Distractor Passage*

Although the order of the four passages within each block was randomized, each infant was tested on four blocks, for a total of 16 test trials. Both familiarization and test trials began with the blinking of the green light in the center of the front panel. Once the infant had oriented in that direction, the light was turned off and one . . .