Research Article

Infants and Adults Use Visual Cues to Improve Detection and Discrimination of Speech in Noise

Kaylah Lalonde^a and Lynne A. Werner^a

Purpose: This study assessed the extent to which 6- to 8.5-month-old infants and 18- to 30-year-old adults detect and discriminate auditory syllables in noise better in the presence of visual speech than in auditory-only conditions. In addition, we examined whether visual cues to the onset and offset of the auditory signal account for this benefit.

Method: Sixty infants and 24 adults were randomly assigned to speech detection or discrimination tasks and were tested using a modified observer-based psychoacoustic procedure. Each participant completed 1–3 conditions: auditory-only, with visual speech, and with a visual signal that only cued the onset and offset of the auditory syllable.

Results: Mixed linear modeling indicated that infants and adults benefited from visual speech on both tasks. Adults

relied on the onset–offset cue for detection, but the same cue did not improve their discrimination. The onset–offset cue benefited infants for both detection and discrimination. Whereas the onset–offset cue improved detection similarly for infants and adults, the full visual speech signal benefited infants to a lesser extent than adults on the discrimination task.

Conclusions: These results suggest that infants' use of visual onset–offset cues is mature, but their ability to use more complex visual speech cues is still developing. Additional research is needed to explore differences in audiovisual enhancement (a) of speech discrimination across speech targets and (b) with increasingly complex tasks and stimuli.

Visual speech is one of the most robust cues that adults use when listening to speech in noisy environments. Adults can detect, discriminate, and recognize speech with greater accuracy and at less favorable signal-to-noise ratios (SNRs) in audiovisual (AV) conditions than in auditory-only conditions (e.g., Grant & Seitz, 2000; Lalonde & Holt, 2016; Sumbly & Pollack, 1954).

Infants spend much of their time in background noise (Erickson & Newman, 2017; Lapierre, Piotrowski, & Linebarger, 2012; Manlove, Frank, & Vernon-Feagans, 2001; Picard, 2004; Voss, 2005). Nevertheless, typically developing infants rapidly learn the ambient language (Saffran, Werker, & Werner, 2006). This is an impressive feat, given that infants are much poorer than adults at detecting, discriminating, and recognizing auditory speech in noise (Leibold,

Bonino, & Buss, 2016; Nozza, Rossman, Bond, & Miller, 1990; Oster & Werner, 2017; Werner, 2013). Given infants' poor auditory-only speech-in-noise perception, visual speech could be one of the most important cues that help infants learn speech and language in noisy environments. However, few studies have directly examined the impact of providing infants with visual cues in the context of speech perception in noise. This study focuses on whether and how infants and adults use visual speech cues to detect and discriminate auditory speech in noise.

Although there is no research directly addressing the topic, previous research led us to hypothesize that infants would be able to use visual speech cues to detect and discriminate speech in noise. Infants show more robust dishabituation and enhanced neural responses to synchronous AV cues than to their auditory-only and visual-only components (Flom & Bahrnick, 2007; Lewkowicz, 1988a, 1988b, 1992b, 1996, 1998, 2000b; Reynolds, Bahrnick, Lickliter, & Guy, 2014; Valliant-Molina & Bahrnick, 2012). Within the first month of life, infants are also attuned to temporal and intensity relations across auditory and visual stimulation (Bahrnick, 2001; Lewkowicz, 2000a; Lewkowicz & Turkewitz, 1980). They rely on the spatial and temporal coincidence of information presented across sensory modalities to parse

^aDepartment of Speech & Hearing Sciences, University of Washington, Seattle

Correspondence to Kaylah Lalonde, who is now at the Center for Hearing Research, Boys Town National Research Hospital, Omaha, NE: kaylah.lalonde@boystown.org

Editor-in-Chief: Frederick (Erick) Gallun

Editor: Alexander Francis

Received March 6, 2019

Revision received May 30, 2019

Accepted July 8, 2019

https://doi.org/10.1044/2019_JSLHR-H-19-0106

Disclosure: The authors have declared that no competing interests existed at the time of publication.

sensory information into objects and events (Bahrick & Lickliter, 2012; Bahrick, Lickliter, & Flom, 2004; Lewkowicz & Kraebel, 2004; Morrongiello, Fenwick, & Chance, 1998). Finally, newborn to 6-month-old infants appear to have rather sophisticated knowledge of the common properties of visual and acoustic vowels. Infants preferentially look at a face that matches the speech they are hearing over a face articulating different speech (Aldridge, Braga, Walton, & Bower, 1999; Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003). In 5- to 15-month-old infants, this effect extends to visually distinct trisyllabic nonwords (Baart, Vroomen, Shaw, & Bortfeld, 2014).

Infants show a preference for facial configurations that match the speech they are hearing, thereby demonstrating sensitivity to the correspondence between auditory and visual speech cues. However, it is unclear whether this preference generalizes to the ability to use this correspondence to benefit from visual speech cues in noisy environments. In fact, previous studies measuring young children's ability to use AV correspondences to enhance speech recognition in noise show limited benefit of visual speech (e.g., Ross et al., 2011; Wightman, Kistler, & Brungart, 2006), bringing into question whether or not infants benefit from visual speech. One goal of the current investigation was to directly assess the extent to which infants use the correspondence between auditory and visual speech to improve detection and discrimination of auditory speech in noise.

Even if infants show an AV benefit, the factors responsible for this benefit may differ between infants and adults. Auditory and visual speech streams are correlated in multiple ways (Chandrasekaran, Trubanova, Stillitano, Calpier, & Ghazanfar, 2009; Munhall & Vaikiotis-Bateson, 2004; Yehia, Rubin, & Vatikotis-Bateson, 1998), and behavioral and neurophysiological evidence indicates that AV speech benefit can result from multiple mechanisms in adults (Eskelund, Tuomainen, & Anderson, 2011; Klucharev, Möttönen, & Sams, 2003; Miller & D'Esposito, 2005). First, adults use the synchronous onsets and offsets of auditory and visual speech to reduce uncertainty as to when auditory speech will occur. By marking when in time to listen/attend, the visual signal effectively improves the SNR of the auditory signal (Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000; Kim & Davis, 2004; Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011). In fluent speech, adults can also use correlations between the amplitude envelopes of auditory and visual speech signals (Grant, 2001). Through extensive experience with speech and language, adults also learn which salient acoustic and visual cues are associated with particular phonemes, syllables, and words. Adults can use this visual/multimodal phonetic knowledge to help discriminate between visually distinct speech sounds (Owens & Blazek, 1985) and to help recognize speech in noisy conditions (Lalonde & Holt, 2016; Tye-Murray, Sommers, & Spehar, 2007).

Although no studies have directly addressed mechanisms underlying AV speech perception benefit in infants, several findings led us to hypothesize that infants may rely on visual temporal cues to mark when to attend, rather

than relying on visual cues to provide phonetic information. First, 4- to 10-month-old infants' binding of auditory and visual speech and other stimuli is determined by the relative onsets and offsets—rather than the correlation between the envelopes—of auditory and visual signals (Lewkowicz, 1996, 2010). Second, infants can rely solely on the correlation between the amplitude envelopes of auditory and visual signals—in the absence of any phonetic information carried by the visual speech—to help segregate competing auditory speech streams; they derive the same benefit when visual speech is replaced by an oscilloscope tracing of the auditory envelope (Hollich, Newman, & Jusczyk, 2005). Third, whereas the cortical structures needed to access experience-based visual/multimodal speech representations show limited maturation during the first year of life, the subcortical pathways that underlie sensitivity to synchronous auditory and visual onsets are highly developed by 6 months of age (Bushara, Grafman, & Hallett, 2001; Eggermont & Moore, 2012).

Additional evidence suggesting that infants rely on temporal cues for AV speech enhancement comes from studies using sine-wave replicas of auditory speech (SWS). In SWS, the basic temporal characteristics of speech are preserved, but phonetic information is removed. Naïve adults often do not perceive SWS as speech. Therefore, differences in performance between SWS and unprocessed conditions (or between naïve and informed observers) have been interpreted as reflecting the use of temporal and phonetic cues, respectively. Adults match unprocessed auditory speech to visual speech much better than SWS, suggesting that they use phonetic information to make this match. However, 5- to 15-month-old infants match SWS to visual speech just as well as they match unprocessed auditory speech to visual speech, suggesting that they rely solely on temporal characteristics to make the match (Baart, Vroomen, et al., 2014).

The purpose of this study was to determine whether and how infants and adults use visual speech cues to help detect and discriminate speech in noise. Using an observer-based procedure, we measured infants' and adults' speech-in-noise detection and discrimination sensitivity under three cue conditions: auditory-only, AV, and with a visual stimulus that only provided cues about the onset and offset of speech. We hypothesized that infants and adults would detect and discriminate auditory speech in noise better when visual speech was present (AV) than when it was absent (auditory-only). If the AV benefits resulted exclusively from reducing uncertainty as to when auditory speech will occur, we expected to observe similar sensitivity with the AV cue as with the visual stimulus that only provided onset and offset cues.

We expected the pattern of results to vary across tasks and age groups. Whereas detection requires only basic awareness of the presence of speech, discrimination requires perception of more fine-grained differences between the spectrotemporal properties of speech sounds. In the detection task, the visual signal may primarily provide information about when to listen. Therefore, we hypothesized that the visual onset cue would account for a large portion

of infants' and adults' AV benefit on the detection task. In discrimination, phonetic information in the auditory and visual signals may also help distinguish between different consonants. We hypothesized that adults would use phonetic information in visual speech to aid speech discrimination. Therefore, we expected the visual onset cue to account for less of adults' AV benefit on the discrimination task than on the detection task. We further hypothesize that the use of phonetic information in visual speech to distinguish between different consonants requires cortical processing and experience with speech and language. Therefore, we expected visual onset cues to account for some—if not all—benefit in infants. Finally, because detection benefits are largely based on temporal cues, we expected infants' AV speech enhancement to be more similar to that of adults for the detection task than for the discrimination task.

Method

This study was approved by the institutional review board at the University of Washington, where participants were tested. Subjects were recruited through the Communication Studies Participant Pool, a facility that maintains a database of contact information for individuals who have expressed an interest in research participation. Some adults were recruited using flyers and newspaper advertisements. Written participant consent (adults) and parent assent (infants) were obtained for all participants. Participants (or their parents) received a \$15 gift card for each hour-long test session. Infants also received an "Infant Hearing Lab" t-shirt.

Infants were tested in two to four sessions, lasting less than 1 hr, using a modified observer-based psychophysical procedure (Werner, 1995). The number of visits depended on the family's availability and the time required to train and test the infant. Adults were tested in a single session lasting up to 1 hr.

Design

Adults and infants were randomly assigned to complete either a speech detection or a speech discrimination task. Participants who completed the detection task were trained to respond when they heard a spoken syllable presented in continuous speech-spectrum noise. Participants who completed the discrimination task heard the syllable /mu/ repeat continuously in speech-spectrum noise at an average rate of one per 1.65 s. They were trained to respond when they heard a different syllable (/gu/ or /lu/). Testing was completed in three conditions: auditory-only, AV, and onset–offset cue. Order of testing across the three visual conditions was counterbalanced across participants, with the caveat that infants always completed the auditory-only condition either first or second. This ensured that infants who only completed two conditions would have an auditory-only baseline for comparison with the other condition.

Participants

Participants in this study were adults between 18 and 30 years of age and infants between 6 and 8.5 months of age. Infants between 6 and 8.5 months of age were included because all indications are that that infants' sensory development is adequate for basic sensory reception and representation of the speech stimuli. Visual acuity is nearly or completely developed by 6 months of age (Sokol, 1978), and auditory sensitivity, spectral resolution, and temporal resolution are adultlike at 6 months of age (Buss, Hall, & Grose, 2012). In addition, previous studies have demonstrated that infants in this age range can detect and discriminate auditory speech in noise (Leibold et al., 2016; Nozza et al., 1990; Oster & Werner, 2017; Werner, 2013). Participants had normal hearing as indicated by parent report or self-report, passing a newborn hearing screening, reporting a negative family history of hearing loss, and reporting no risk factors for hearing loss. Participants or their parents also reported no treatment for otitis media in the 2 weeks preceding the study and no more than two prior episodes of otitis media. Participants passed a tympanometric screening in the test ear on each day of testing (peak admittance of at least 0.2 mmho at a pressure between –200 and 50 daPa). Adults were native English speakers. All but three infants were from homes where English is the primary language spoken. These three infants were included in the analyses below, because the pattern of results did not change when they were excluded.

In total, 30 adults and 88 infants were tested. Table 1 shows the number of subjects included/excluded in each task and age group. Data from 24 adults (12 per task) and 60 infants (25 for detection, 35 for discrimination) are reported. Whereas all adults completed testing in all three conditions, some infants completed only one or two conditions (see Table 1 for a breakdown).¹

Stimuli

The stimuli used in this experiment consisted of professional AV recordings of three syllables (/mu/, /gu/, and /lu/) spoken by a 46-year-old native English speaker. Visual stimuli included the talker's full face. Auditory stimuli were recorded at a resolution of 32 bits and a sampling rate of 24414 Hz. The syllable durations were 0.560, 0.533, and 0.609 ms, respectively. Differences between these consonants are highly salient, both auditorily and visually. Pilot testing indicated that adults recognize these stimuli perfectly in auditory-only and AV conditions in quiet and do not confuse these consonants visually (Lalonde & Werner, 2019). These recordings are part of a larger set of stimuli available at <https://osf.io/6gk7p>. The stimuli were edited

¹More infants were recruited for the discrimination task than the detection task, because our initial goal was to obtain repeated measures from 20 infants on each task. As Table 2 shows, it was more difficult to get infants to complete multiple conditions of the discrimination task.

Table 1. Number of subjects included and excluded from data analysis.

Target	Infants			Adults		
	Detection	Discrimination		Detection	Discrimination	
		/gu/	/lu/		/gu/	/lu/
Total included in analysis	25	19	16	12	6	6
Completed all three modalities	10	8	2	12	6	6
Completed AO and AV	4	2	2			
Completed AO and onset–offset cue	6	4	1			
Completed one modality	5	5	11			
Total excluded	10	7	11	2	1	3
Failed tympanometric screening	4	3	5			
Did not pass training/finish testing	6	4	6			
Computer error				1	1	2
Did not follow instructions				1		
Total recruited	35	26	27	14	7	9

Note. AO = auditory-only; AV = audiovisual.

in Final Cut Pro (Version 10.0.6; Apple) and Adobe Audition (Version 6, Adobe Systems).

Speech stimuli were presented in three visual conditions: auditory-only, AV, and onset–offset cue. In the auditory-only condition, a still image of the talker remained on the screen throughout testing. In the AV condition, auditory and visual stimuli were presented synchronously. In the onset–offset cue condition, the visual speech signal was replaced with two images of the talker: an open-mouthed picture presented from onset to offset of the associated auditory syllable and a closed-mouth picture presented between syllables. Visually, the talker did not say a specific syllable in the onset–offset condition; her mouth only opened and closed. This is similar to the stimulus used in a previous study to control for the fact that the visual stimulus clues subjects as to when to attend to the auditory stimulus (Ma, Zhou, Ross, Foxe, & Parra, 2009). The duration of the mouth opening varied only slightly across syllables, that is, 17, 16, and 18 frames for /mu/, /gu/, and /lu/, respectively, at a rate of 30 frames/s.

Speech stimuli were presented in a continuous 65 dB SPL speech-spectrum noise. The noise was created by passing a white noise through a filter matching the long-term average spectrum of 18 concatenated syllables spoken by the target talker (two each of /ga/, /gi/, /gu/, /la/, /li/, /lu/, /ma/, /mi/, and /mu/), including the three target syllables.

We measured six adults' detection threshold for each of the auditory syllables in the speech-spectrum noise using an adaptive three-interval forced-choice paradigm. These thresholds served as estimates of the auditory speech level corresponding to 0 dB nHL for each syllable. We adjusted the root-mean-square level of each syllable based on its mean threshold. Thresholds for the three stimuli fell within a 0.8-dB range, so the overall root-mean-square power differs by ≤ 0.8 dB across syllables. Stimuli were calibrated using a Zwislocki coupler (Knowles Electronics) and flat weighting. At the time of testing, levels were checked in the subject's ear canal using an Etymotic Research ER-7C probe microphone system.

Experimental Setup

Infants

The test setup for infants is shown in Figure 1. The infant sat in their caregiver's lap inside a double-walled sound-treated booth, facing a 27-in. widescreen monitor. To the left of the monitor, an assistant manipulated small, quiet toys to keep the infant facing midline and draw the infant's gaze toward the screen before the start of each trial. To the right of the stimulus monitor, two mechanical toys

Figure 1. Test setup for infants.

with lights in plexiglass boxes and a small monitor served as visual reinforcement. The experimenter sat outside the booth and observed through two adjustable cameras. One camera was focused on the infant's face; the other camera provided a full view of the infant and the assistant's toys. To ensure that the adults inside the booth did not bias the infant's response, both the caregiver and the assistant wore circumaural headphones. The assistant faced away from the screen and listened to the experimenter's instructions. The caregiver listened to music and wore a blindfold. The blindfold was placed such that the caregiver could see the infant but not the screen. In addition to these precautions, it would be difficult for either adult to hear the stimuli, because they were presented to the infant via an Etymotic Research ER-2 insert earphone in the right ear. As described below, experiments were also designed so that the visual stimulus did not independently indicate a signal.

Adults

The test setup was the same for adults as for infants, except that there was no assistant or caregiver inside the booth.

Experimental Procedure

Most of the procedural details are common across both age groups, both tasks, and all three visual conditions. Across all experiments, participants were trained to respond to an auditory "signal." In detection, the signal was the presentation of a speech sound. In discrimination, it was a change from one speech sound to another. No-signal trials were randomly interspersed among the signal trials. On no-signal trials, no sound (detection) or no speech sound change (discrimination) occurred. In other words, the auditory and visual stimuli did not change from the background. We proceed by describing the training and assessment procedures for the detection task in infants and then describing modifications for testing adults and for testing discrimination.

Detection Task

Participants who completed the detection task were trained to respond when they heard a spoken syllable presented in continuous speech-spectrum noise. No-signal trials were randomly interspersed among the signal trials. On no-signal trials, no speech sound was presented. In other words, the auditory and visual stimuli did not change from the background.

Participants were tested in three visual conditions: auditory-only, AV, and onset–offset cue. In the auditory condition, a still image of the talker remained on the screen throughout testing (see Figure 2a). In the AV and onset–offset conditions of the detection task, the video of the talker saying /mu/ or opening and closing her mouth played repeatedly, so that participants could not respond solely based on visual information (see Figure 2b).

Infants. The experimenter initiated a trial when the infant was attentive and looking at the screen and then observed the infant's behavior to determine whether or not a

signal was presented. Infants often responded by turning their heads to the assistant or visual reinforcement, opening their eyes wider, or momentarily freezing. When the observer believed a signal trial occurred, she pressed a button during the 4-s trial. The only information that the observer could use to make a decision was the infant's behavior; the observer, assistant, and parent were all blind to whether the signal was being presented. The observer received feedback on the computer screen at the end of each trial. Infants received visual and social reinforcement when the observer correctly identified a signal trial. This visual and social reinforcement served to train/condition infants' responses.

In each condition, the experiment proceeded in three phases: familiarization, training, and test. The purpose of the familiarization phase was to demonstrate the association between the signal and reinforcement. The familiarization phase included four signal trials and one no-signal trial. Speech stimuli were presented at a clearly audible SNR (14 dB), and infants were reinforced for every signal trial, regardless of their response. Reinforcement began 590–667 ms after the target stimulus, before the start of the next background stimulus. In the training phase, speech stimuli were presented at 10 dB SNR. The purpose of the training phase was to establish that the observer could reliably identify signal trials. Signal and no-signal trials were presented in a random order, and reinforcement was dependent on a correct signal response. Training continued until the observer responded correctly to four of the last five signal trials and four of the last five no-signal trials. If participants failed to reach the criterion within 40 trials, the experimenter gave the participant a break or ended testing for the day. On average, infants required 19.9 ($SD = 8.3$) training trials to reach the criterion on the detection task. Once infants trained to criterion, they completed the test phase at 2 dB SNR. The test phase included 10 signal trials and 10 no-signal trials, in a pseudorandom order.

Adults. Adults were tested using the same procedures as children, except that they were instructed to watch the screen and raise their hands when they heard the change that activated the reinforcer toys. As for infants, the familiarization phase demonstrated the association between the signal and reinforcement, the training phase was used to establish that the adult could reliably identify signal trials, and the test phase was used to assess how well the adult detected or discriminated speech in noise. On average, adults required 11.4 ($SD = 3.5$) training trials to reach the criterion on the detection task. Adults were tested alone in the booth at less favorable SNRs: 4 dB SNR for familiarization, –6 dB SNR for training, and –8 dB SNR for testing because pilot testing indicated that this would lead to similar performance across the adult and infant groups. Adults receive visual reinforcement from the mechanical toys but did not receive social reinforcement.

Discrimination Task

Participants who completed the discrimination task heard the syllable /mu/ repeat continuously in speech-spectrum

Figure 2. Example background, signal, no-signal, and foil trial stimuli in the detection and discrimination tasks. (a) Example of the visual signal for the auditory-only condition of both tasks. This single image remained on the screen throughout auditory-only testing. The auditory speech signal was identical to the audiovisual (AV) conditions. (b) Example of the AV detection condition. The white portions of the timeline represent the background and no-signal trials. The gray portion of the timeline represents a signal trial. The visual speech repeated continuously, but auditory speech only occurred on signal trials. (c) Example of the AV discrimination condition. The white portions of the timeline represent the background and no-signal trials. The gray portion of the timeline represents a signal trial, and the striped portion of the timeline represents a foil trial. The auditory and visual speech was /mu/ repeating in the background. On signal trials, both the auditory and visual speech changed. On foil trials, only the visual speech changed.

noise at an average rate of one per 1.65 s. They were trained to respond when they heard a different syllable (/gu/ or /lu/). No-signal trials were randomly interspersed among the signal trials. On no-signal trials, no speech sound change occurred.

In addition to signal and no-signal trials, the AV and onset–offset cue conditions of the discrimination task also included foil trials. During foil trials, the visual stimulus indicated a target signal, but the auditory signal continued to play the background syllable (see Figure 2c). The purpose of the foil trials was to reinforce that participants should respond to the auditory signal and to be certain that participants were not responding solely based on visual information.^{2,3} The foil trials proved unnecessary for the onset–offset cue condition of the discrimination task: A group of 24 adults was unable to discriminate the one-frame (approximately 30 ms) duration difference between the background and

target syllable, even when informed of the short duration difference between them.

To eliminate data from subjects who responded primarily to the visual signal, we initially excluded participants based on their response to AV foil trials. Specifically, we calculated d' (Green & Swets, 1966) using the proportion of correct signal trials for hit rate and the proportion of incorrect foil trials for false-alarm rate and eliminated all subjects for whom d' with foil trials was less than 0.8. However, these observations from six infants and one adult were included in the analyses below, because the pattern of results did not differ appreciably when they were excluded.

The discrimination experiment proceeded in the same three phases as the detection experiment, with two exceptions.⁴ In the AV and onset–offset cue conditions of the discrimination task, foil trials were used in lieu of no-signal trials. This way, we trained infants to respond to the auditory change rather than the visual change. Second, the test phase in the AV and onset–offset cue conditions included 30 trials (10 signal trials, 10 no-signal trials, and 10 no-signal trials, in a pseudorandom order). SNRs for discrimination testing were the same as for detection testing in each age group and experimental phase.

Infants. On average, infants required 19.5 ($SD = 8.7$) training trials to reach the criterion on the discrimination task.

²Note that, in a previous study, adults who were presented these incongruent pairs in quiet (auditory /mu/–visual /gu/ and auditory /mu/–visual /lu/) never reported the visual consonant (which would result in a false alarm on foil trials in this experiment). Instead, they typically reported being aware that the auditory and visual signals did not match (Lalonde & Werner, 2019). Therefore, we do not expect McGurk-like effects (McGurk & MacDonald, 1976) caused participants to perceive the foils as signals.

³Foil trials were not necessary in the auditory-only conditions and the detection task, because the signal and no-signal visual stimuli were the same as the background visual stimuli.

⁴During familiarization training

Adults. On average, adults required 16.2 ($SD = 10.6$) training trials to reach the criterion on the discrimination task.

Analysis and Results

Sensitivity (d') was calculated based on signal hit rates and no-signal false-alarm rates (Green & Swets, 1966; Hautus, 1995). Hit rate is defined as the portion of correct signal trials. False-alarm rate is defined as the portion of incorrect no-signal trials. A value of 0.025 was added to hit and false-alarm rates of 0 and subtracted from hit and false-alarm rates of 1, because d' takes on infinite values when hit rate or false-alarm rate is 0 or 1. With this common correction (Hautus, 1995; MacMillan & Kaplan, 1985), d' could range from -3.92 to $+3.92$.

Figure 3 shows box plots of d' scores for the two tasks and age groups. White, gray, and black circles indicate individual performance in the auditory-only, onset–offset cue, and AV conditions, respectively. Individual data for participants who completed multiple conditions are shown in Figure 4. Overall, adults and infants seemed to benefit from both cues in detection. In discrimination, adults benefited more from the AV cue than from the onset–offset cue. In infants, this AV discrimination benefit was smaller.

These observations were statistically evaluated by fitting a linear mixed model using the lmer function in the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). Linear mixed modeling was chosen for this analysis, because it did not require exclusion of data from subjects who could not complete multiple conditions. Statistical results are shown in Table 2 and described below.

Detection

For the detection task, the predictors that were examined included age (coded categorically as infants and

adults) and modality (auditory-only, AV, and onset–offset cue). The reference condition was infants' performance in the auditory-only condition. To explore differences in benefit between the onset–offset cue and the AV cue, analyses were repeated with the onset–offset cue as reference.

The maximal random-effects structure supported by the data included a random intercept for subject. The random intercept had a variance of 0.164 and an SD of 0.405. The final model included a main effect of modality, with significant differences between the auditory-only condition and the onset–offset cue condition, $b = 0.409$, $t = 2.567$, $p = .0127$, and between the auditory-only condition and the AV cue condition, $b = 0.374$, $t = 2.214$, $p = .0305$. There was also a marginal interaction of modality (auditory-only vs. AV cue) and age, $b = 0.491$, $t = 1.891$, $p = .0637$. This marginal interaction likely reflects the somewhat larger mean AV cue benefit in adults (mean d' difference = 0.865) relative to infants (mean d' difference = 0.322). These results suggest that infants and adults benefited from both visual cues and that they benefited as much from the onset–offset cue as from the AV cue.

Discrimination

For the discrimination task, the predictors that were examined included age (coded categorically as infants and adults), modality (auditory-only, AV, and onset–offset cue), and speech target (/gu/ and /lu/). The reference condition was infants' performance for the speech target /gu/ in the auditory-only condition. To explore differences in benefit between the onset–offset cue and the AV cue, analyses were repeated with the onset–offset cue as reference.

The maximal random-effects structure supported by the data included a random intercept for subject. The random intercept had a variance of 0.057 and an SD of 0.239. The final model for discrimination included a main effect of modality, with a significant difference between the auditory-only and AV cue conditions, $b = 0.599$, $t = 3.058$, $p = .0031$,

Figure 3. Box and whisker plots of adult detection, infant detection, adult discrimination, and infant discrimination data, overlaid with circles representing individual data. Open circles indicate auditory-only (A-only) conditions, gray circles indicate onset–offset cue conditions, and black circles indicate audiovisual (AV) conditions.

Figure 4. Individual repeated-measures results for adults (left) and infants (right) on the detection and discrimination tasks. (a) Detection results for participants tested in auditory-only (A-only) and audiovisual (AV) conditions. (b) Detection results for participants tested in A-only and onset–offset cue conditions. (c) Discrimination results for participants tested in A-only and AV conditions. (d) Discrimination results for participants tested in A-only and onset–offset cue conditions. Note that the A-only data repeat in the same relative location in the two detection graphs and two discrimination graphs, respectively, to allow within-subject comparisons of participants tested in all three modalities.

and a marginal difference between the auditory-only and onset–offset cue conditions, $b = 0.350$, $t = 1.764$, $p = .0824$. There was a significant interaction between age and modality for the auditory-only versus AV conditions, $b = 1.181$, $t = 4.164$, $p < .0001$, and onset–offset versus AV conditions,

$b = 1.313$, $t = 4.385$, $p < .0001$. The effect of age was not significant in the auditory-only reference condition, $b = 0.162$, $t = 0.775$, $p = .4401$, suggesting that we successfully equated group performance in the auditory-only condition by using different SNRs across groups. The interaction likely reflects the larger AV cue benefit in adults (mean d' difference = 1.743) relative to infants (mean d' difference = 0.553). Finally, there was an interaction between target and modality for the auditory-only versus onset–offset cue conditions, $b = -0.639$, $t = -2.227$, $p = .0290$, and a marginal interaction of target and modality for the onset–offset cue versus AV cue conditions, $b = 0.565$, $t = 1.868$, $p = .0656$.

To explore the two-way interactions, we generated separate mixed models for the infant and adult data, each with predictors of modality (auditory-only, AV, onset–offset cue) and speech target (/gu/, /lu/). For adults, the random intercept had a variance of 0.125 and an SD of 0.354. The final model only included an effect of modality. Adults' sensitivity was higher with the AV cue than with the onset–offset cue, $b = 1.845$, $t = 7.172$, $p < .0001$, or in the auditory-only condition, $b = 1.743$, $t = 6.776$, $p < .0001$. These results suggest that adults benefited from the AV cue but not from the onset–offset cue and that this effect was independent of speech target.

For infants, the random intercept had a variance of 0.016 and an SD of 0.125. The final model included a main effect of modality, with significant differences between the auditory-only and AV cue conditions, $Beta = 0.636$, $t = 3.358$, $p = .0017$, and the auditory-only and onset–offset cue conditions, $Beta = 0.444$, $t = 2.305$, $p = .0270$. Overall, infants demonstrated higher sensitivity with both visual cues than in the auditory-only condition. There was also an interaction of modality and target. The interaction was significant for the onset–offset cue versus auditory-only conditions, $b = -0.954$, $t = -2.834$, $p = .0066$, and the onset–offset cue versus AV cue conditions, $b = 0.785$, $t = 2.167$, $p = .0347$.

To explore the interaction of speech target and modality, we examined the groups of infants that were tested using each speech target separately (see Figure 5). For each target, we generated a mixed model with an effect of modality (auditory-only, AV, onset–offset cue). For target /gu/, the random intercept had a variance of 0.009 and an SD of 0.097. There was a main effect of modality, with differences between the auditory-only and AV cue conditions, $b = 0.632$, $t = 3.340$, $p = .0022$, and the auditory-only and onset–offset cue conditions, $b = 0.442$, $t = 2.289$, $p = .0298$. There was no significant difference between performance with the AV cue and that with the onset–offset cue. These results suggest that infants tested with the /gu/ target benefited from both visual cues and benefited equally from the AV and onset–offset cues.

For target /lu/, the random intercept had a variance of 0.249 and an SD of 0.499. The final model for the /lu/ target included only a significant difference between the AV cue and onset–offset cue conditions, $b = 0.614$, $t = 2.696$, $p = .0351$. Neither cue differed significantly from the auditory-only condition. The lack of visual cue benefits

Table 2. Results of linear mixed models.

Task	Predictors	<i>b</i>	<i>t</i>	<i>p</i>
Detection	Modality (auditory-only/AV)	0.374	2.214	.0305
	Modality (auditory-only/onset–offset cue)	0.409	2.567	.0127
	Modality (onset–offset cue/AV)	–0.036	0.197	.8440
	Age	–0.206	–0.915	.3633
	Age: modality (auditory-only/AV)	0.491	1.891	.0637
	Age: modality (auditory-only/onset–offset cue)	0.243	0.958	.3423
	Age: modality (onset–offset cue/AV)	0.248	0.927	.3577
Discrimination	Modality (auditory-only/AV)	0.599	3.058	.0031
	Modality (auditory-only/onset–offset cue)	0.350	1.764	.0824
	Modality (onset–offset cue/AV)	0.249	1.185	.2398
	Age	0.162	0.775	.4401
	Target	0.171	0.882	.3801
	Age: modality (auditory-only/AV)	1.181	4.164	< .0001
	Age: modality (auditory-only/onset–offset cue)	–0.135	–0.462	.6455
	Age: modality (onset–offset cue/AV)	1.313	4.385	< .0001
	Target: modality (auditory-only/AV)	–0.074	–0.268	.7896
	Target: modality (auditory-only/onset–offset cue)	–0.639	–2.227	.0290
	Target: modality (onset–offset cue/AV)	0.565	1.868	.0656
Discrimination in adults	Modality (auditory-only/AV)	1.743	6.776	< .0001
	Modality (auditory-only/onset–offset cue)	–0.102	–0.395	.696
	Modality (onset–offset cue/AV)	1.845	7.172	< .0001
Discrimination in infants	Modality (auditory-only/AV)	0.636	3.358	.0017
	Modality (auditory-only/onset–offset cue)	0.444	2.305	.0270
	Modality (onset–offset cue/AV)	0.192	0.942	.3511
	Target	0.352	1.736	.0875
	Target: modality (auditory-only/AV)	–0.169	–0.543	.5896
	Target: modality (auditory-only/onset–offset cue)	–0.954	–2.834	.0066
	Target: modality (onset–offset cue/AV)	0.785	2.167	.0347
	Modality (auditory-only/AV)	0.632	3.340	.0022
Discrimination in infants, /gu/	Modality (auditory-only/onset–offset cue)	0.442	2.289	.0298
	Modality (onset–offset cue/AV)	0.190	0.937	.3556
	Modality (auditory-only/AV)	0.357	1.941	.101
Discrimination in infants, /lu/	Modality (auditory-only/onset–offset cue)	–0.257	–1.241	.261
	Modality (onset–offset cue/AV)	0.614	2.696	.0351

Note. The default reference condition is auditory-only in infants (with /gu/ as the target in discrimination). For comparisons between onset–offset and audiovisual (AV) cue conditions, onset–offset cue is the reference. Bold text indicates a significant effect or interaction.

for infants tested with the /lu/ target is difficult to interpret and may reflect the small sample size. The onset–offset cue and AV cue conditions with target /lu/ include only seven and five observations, respectively. This occurred primarily because few infants tested with the /lu/ target were able to complete more than one condition (see Table 1). Whereas infants tested with target /gu/ completed testing in 82% of attempted conditions, infants tested with target /lu/ completed testing in 65.7% of attempted conditions.

Discussion

The purposes of this study were twofold. First, we aimed to determine whether infants use visual speech cues to better detect and discriminate auditory speech in noise. Second, we examined whether infants rely on the same cues as adults to do so. More specifically, we examined how much of infants' and adults' benefit can be ascribed to visual onset–offset cues.

Our findings indicated that infants—like adults—benefited from the AV cue on both the detection and discrimination tasks. Adults relied heavily on the visual onset–

offset cue for detection benefit, but the same cue did not improve their discrimination performance. This pattern of results suggests that adults relied on more fine-grained and/or phonetic visual cues for discrimination benefit. In contrast, infants relied on the visual onset–offset cue for both detection and discrimination. Infants benefited from the onset–offset cue to a similar degree as adults on the detection task. However, they were less successful than adults at using phonetic and/or more fine-grained spectrotemporal cues to improve discrimination. These results suggest that 6- to 8.5-month-old infants are relatively mature in their ability to use visual onset–offset cues to better detect speech in noise but are still developing in their ability to use phonetic and/or more fine-grained spectrotemporal cues.

Mature AV Speech Enhancement

Adults relied on temporal cues for speech detection benefit but relied on phonetic and/or more fine-grained spectrotemporal cues for speech discrimination benefit. These findings are consistent with multiple mechanism accounts of AV speech enhancement (Eskelund et al., 2011; Miller &

Figure 5. Box and whisker plots of infant discrimination data for targets /gu/ and /lu/, overlaid with circles representing individual data. Open circles indicate the auditory-only (A-only) conditions, gray circles indicate the onset–offset cue conditions, and black circles indicate the audiovisual (AV) conditions.

D'Esposito, 2005; Peelle & Sommers, 2015). Researchers have categorized visual speech cues as providing information about phonetic form and about timing (Kim & Davis, 2014; Paris, Kim, & Davis, 2013). These form and timing cues are processed in analogous speech-specific versus not specific (Eskelund et al., 2011) or prediction versus constraint (Peelle & Sommers, 2015) mechanisms of AV enhancement/integration. Additionally, distinct neural circuits have been characterized for perceptual binding versus detecting spatiotemporal correspondence (Baart, Stekelenburg, & Vroomen, 2014; Miller & D'Esposito, 2005).

The results of the current investigation are consistent with a growing body of literature showing that adults use different mechanisms for different AV speech perception tasks. Evidence from studies using SWS suggests that adults rely on basic temporal cues for some AV speech tasks and on phonetic information for others. For example, McGurk effects only occur when participants perceive auditory stimuli as speech (Eskelund et al., 2011; Stekelenburg & Vroomen, 2012; Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Stekelenburg, 2011). The effects are seen when adults are making decisions based on phonetic information (when aware of the speechlike nature of SWS) but not when they are making decisions based on acoustic information (when they are naïve to the speechlike nature of SWS). Likewise, naïve adults are much poorer at matching SWS to visual speech than they are at matching unprocessed auditory speech to visual speech (Baart, Vroomen, et al., 2014), suggesting that they rely on phonetic information for AV speech matching. On the other hand, temporal order judgments, synchrony judgments, and AV detection benefits are similar for naïve and informed listeners, suggesting that adults rely on basic temporal cues for these tasks (Eskelund et al., 2011; Vroomen & Stekelenburg, 2011).

In both the current study and the existing literature, visual temporal cues have proved highly beneficial for speech detection in noise. Such findings are consistent with the idea

that visual cues provide information about when onsets and peaks of the auditory signal will occur (e.g., Grant & Seitz, 2000). In the current investigation, these temporally based benefits did not generalize to the discrimination task. We expected some generalization, given the hierarchical nature of speech (Erber, 1982): One must detect a signal in order to discriminate salient differences between background and target stimuli. Therefore, cues that help detect speech should theoretically enhance intelligibility and, thus, improve AV discrimination. One previous study demonstrated such downstream effects of visual temporal cues on speech sound discrimination/identification (Schwartz, Berthommier, & Savariaux, 2004). An ambiguous visual speech signal helped French-speaking adults detect an auditory prevoicing cue, resulting in better discrimination between syllables with initial voiced plosives (/dy/, /du/, /gy/, /gu/) and other syllables with either voiceless plosives (/ty/, /tu/, /ky/, /ku/) or no consonant (/y/, /u/). The advantage was observed even when there was no phonetic information in the visual signal (when the same visual speech signal was paired with all auditory consonants). However, the advantage disappeared when a nonspeech visual shape was used to provide the temporal information (a red bar that increased and decreased in size with the same time course as the visual speech).

Several methodological differences between the Schwartz et al. (2004) study and the discrimination task in the current study may account for these discrepant results. Whereas participants in the current study were only required to respond when they heard a change in the consonant, participants in the Schwartz et al. study chose a response from 10 alternatives. Perhaps more importantly, there are large differences between the stimuli used in these two studies. The onset–offset visual stimulus in the current investigation provided less detailed timing information than the visual speech stimulus in the Schwartz et al. study. It is possible that the gradual change from closed to open was

necessary for the temporal benefits. Additionally, a visual cue that helps detect the prevoicing bar may not be as useful with the set of stimuli in the current investigation as it was in the Schwartz et al. study, because the target and background auditory syllables in the current investigation all began with voiced consonants. The latter explanation would suggest that the application of visual temporal cues to syllable discrimination and identification is limited to specific speech contrasts.

Infants' Use of Visual Onset Cues

Infants and adults benefited from the visual onset–offset cue to a similar degree. Infants' adultlike use of simple visual temporal cues appears to persist into childhood. Previous studies have shown adultlike AV benefit for tone detection in a broadband masker for 5- to 13-year-olds (Bonino, Leibold, & Buss, 2013) and adultlike AV speech detection benefit among 6- to 8-year-olds (Lalonde & Holt, 2016).

Maturity in the ability to use onset information is consistent with the intersensory redundancy hypothesis (Bahrick, Flom, & Lickliter, 2002; Bahrick & Lickliter, 2012). The premise of the intersensory redundancy hypothesis is that infants initially attend to and process the amodal properties of AV stimuli (those specified across multiple modalities), beginning with basic attributes such as synchrony and intensity relations before advancing to higher order AV relations, such as duration, rhythm, and prosody. As infants gain experience with AV stimuli, they begin to process information more flexibly and can attend to multiple properties of the event (including modality-specific information).

Synchronous onsets are the foundation of early AV integration (Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009) and are arguably the most fundamental type of amodal information (Bahrick & Lickliter, 2012). Sensitivity to these cues only requires that infants perceive stimulus energy onsets and offsets, rather than extracting any complex relation between the auditory and visual stimuli (Lewkowicz, 2010). Therefore, synchronous onsets and offsets are likely the first visual cues that infants can use to aid in auditory speech-in-noise perception.

In fact, decades of research have demonstrated that sensitivity to simultaneous auditory and visual onsets emerges early in life (Bahrick, 1988; Dodd, 1979; Lewkowicz, 1986, 1996; Spelke, 1979). For example, 4- to 10-month-old and 2- to 8-month-old infants detect asynchrony between simple speech and nonspeech auditory and visual events (Lewkowicz, 1996, 2010). Three- and 4-month-old infants can match auditory and visual nonspeech events based on temporal synchrony (Bahrick, 1988; Spelke, 1979), and 3-month-olds fail to make the same matches if stimuli are asynchronous (Bahrick, 1988). Finally, in the speech domain, 10- to 16-week-old infants attend more to synchronous AV speech than to AV speech presented with a 400-ms asynchrony (visual lead; Dodd, 1979).

Sensitivity to onset synchrony emerges earlier than sensitivity to other amodal properties and cross-modal

associations and facilitates sensitivity to higher level relations between auditory and visual signals (Bahrick, 1992, 1994, 2001; Bahrick & Lickliter, 2012; Lewkowicz, 2000a). Without synchrony, infants often fail to respond to changes in duration-, rate-, and tempo-based AV intersensory information until later in development (Lewkowicz, 1986, 1988a, 1988b, 1992a, 2000a). In the speech domain, stimulus energy onsets are also the basis on which 4- to 10-month-old infants detect synchrony of a consonant–vowel speech syllable (Lewkowicz, 2010) and, on which newborn and 4- to 6-month-old infants match nonhuman auditory and visual vocalizations (Lewkowicz, Sowinski, & Place, 2008; Lewkowicz et al., 2010).

Quite a bit of research has demonstrated that infants are sensitive to AV onset synchrony relations. However, it was unclear whether infants' preference for synchronous speech, their ability to detect asynchrony, and their ability to match based on synchrony would facilitate infants' auditory detection of speech in noise. If anything—based on 4- to 10-month-old infants' relative insensitivity to asynchrony—we would have predicted decreased benefit from synchronous visual onsets in infants relative to adults (Lewkowicz, 2010). Nevertheless, infants exhibited similar AV enhancement of speech detection as adults. These results suggest that tests of sensitivity to asynchrony and tests of AV enhancement of speech detection measure different things.

We know of no other research that measured infants' use of visual onset–offset cues to enhance speech detection and discrimination. However, infants' early ability to use visual onset–offset cues is consistent with previous animal studies. Prenatal quails exposed to a flash of light at the beginning of a five-note bobwhite maternal call learn the call faster than those who receive auditory-only or sequential auditory and visual prenatal exposure (Jaime, Bahrick, & Lickliter, 2010). The onset cue is particularly important; having the flash at the beginning of the call is more effective than having it at any other time before, during, or after the call.

It seems worth noting that infants benefited from the visual onset–offset cue in conditions in which adults did not. The visual onset–offset cue helped infants—but not adults—discriminate the target /gu/ from the background /mu/. This result suggests that infants extract and/or attend to different visual speech cues than adults do, an interpretation consistent with both the infant AV cue matching literature and the infant auditory speech perception literature. At 3 weeks of age, infants' cardiac responses to auditory and visual stimuli suggest that they respond to auditory stimuli in relation to their similarity to the intensity of previously presented visual stimuli. Thus, infants—unlike adults—match intensity across modalities (Lewkowicz & Turkewitz, 1980). Infants also seem to use different auditory cues than adults to discriminate auditory syllables in noise (Cabrera & Werner, 2017); whereas infants seem to depend on fast amplitude modulation cues more than adults to discriminate syllables in quiet, adults seem to depend on frequency modulation cues more than infants do.

Infants' Use of Fine-Grained Spectrotemporal and/or Phonetic Cues

The full visual speech signal did not improve infants' speech detection or discrimination of target /gu/ beyond that observed with only the onset–offset cue. Although infants who discriminated target /lu/ performed better with the full AV signal than with the onset–offset cue, neither visual cue resulted in better performance than the auditory-only condition. The latter result is difficult to interpret and may reflect the small sample size—and an even smaller sample of repeated measures—for the /lu/ target. The results of this study therefore do not provide clear evidence that infants used fine-grained spectrotemporal and/or phonetic cues from the visual signal. However, it is possible that infants used different cues when they had access to the full visual speech signal (in the AV condition) than they used when they only had access to visual onset–offset cues. Therefore, we are unable to make conclusions about whether or not infants use fine-grained spectrotemporal and/or phonetic cues in AV speech perception.

Although we cannot resolve whether infants used fine-grained spectrotemporal and/or phonetic cues from the visual speech signal, it is clear from the results that infants were not as good as adults at using these cues. Infants benefited less than adults from the AV cue on the discrimination task. This result was expected given that the cortical mechanisms needed to access experience-based visual/multimodal speech representations show limited maturation during the first year of life (Bushara et al., 2001; Eggermont & Moore, 2012). Moreover, studies with preschool- and school-age children show gradual increases across age in the use of visual speech cues to enhance speech perception in noise (e.g., Ross et al., 2011; Wightman et al., 2006; but see Knowland, Evans, Snell, & Rosen, 2016). Nevertheless, there were some reasons to suspect that infants use phonetic information. Previous studies have shown that newborn, 2-month-old, and 4-month-old infants preferentially look at a face that matches the speech they are hearing over a face articulating different speech, even when both visual signals have the same temporal characteristics (Aldridge et al., 1999; Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003). This suggests that infants are sensitive to something other than the temporal correspondence between the signals and that this sensitivity is not something they must learn through experience with speech and language.

Results from AV matching studies with infants are consistent with the immature use of visual phonetic cues observed in the current study. Prior studies demonstrated early sensitivity to some nontemporal AV cues, but infants are not as good as adults at matching phonetic information in auditory and visual speech. Across studies, approximately 60%–75% of infants demonstrate a preference for visual speech that matches the auditory signal over visual speech that does not match, whereas all adults match at ceiling (Aldridge et al., 1999; Baart, Vroomen, et al., 2014; Kuhl & Meltzoff, 1982, 1984; Lalonde & Holt, 2015; Patterson & Werker, 1999, 2003).

As noted in the introduction, the fact that infants are sensitive to the correspondence between auditory and visual speech does not mean that they use this correspondence to benefit from visual speech cues in noisy environments. As Shaw and Bortfeld (2015) recently pointed out, infants' ability to match auditory and visual speech is frequently mischaracterized as AV speech "integration," which is one reason the field has not achieved a more thorough characterization of the mechanisms underlying AV speech perception/enhancement. To the best of our knowledge, only one previous study demonstrated that infants use visual speech in AV speech perception. Hollich et al. (2005) assessed whether 7.5-month-old infants could use synchronous, congruent visual speech to segregate two competing auditory speech streams. Infants were presented auditory-only or AV passages spoken by a female talker, with a male distractor reading the methods section of a research paper in the background. The passages from the target talker contained certain key words (*cup, dog, bike, or feet*). During the test phase, infants familiarized with the AV passage demonstrated preference for auditory target words that were included in the passage over nontarget words that were not included in the passage, but infants familiarized with the auditory-only passage did not. The authors concluded that infants used the visual cues to segregate the competing speech streams and segment the target words from the fluent passage.

In two additional experiments, infants familiarized with asynchronous AV speech failed to preferentially look to target words, but infants familiarized with an oscilloscope tracing that showed the auditory envelope demonstrated the same preference as infants familiarized with synchronous AV speech (Hollich et al., 2005). These results suggest that infants relied solely on temporal cues in visual speech—in the absence of any phonetic information—to help segregate the competing streams and segment the target words. The temporal cues they used were more complex than the visual onset–offset cues, which would not have been very useful for segregating words from the middle of fluent passages. The oscilloscope tracings provided information about the correlation between the amplitude envelopes of the auditory and visual signals, a cue that helps adults better detect sentences in noise (Grant & Seitz, 2000) and that helps adults' auditory cortex to track the temporal amplitude envelope of auditory speech (Schroeder & Foxe, 2005; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008; ten Oever, Schroeder, Poeppel, van Atteveldt, & Zion-Golumbic, 2014; Zion-Golumbic, Cogan, Schroeder, & Poeppel, 2013). In the current investigation, infants did not use such complex visual temporal cues. However, the short duration (consonant–vowel) stimuli in the current investigation were not optimal for observing infants' use of longer duration temporal cues, such as amplitude envelope information.

Differences in Discrimination Difficulty for Target /gu/ and Target /lu/

Infants tested with the /gu/ target were more likely to pass training and finish testing than infants tested with the

/lu/ target. We were surprised to find that infants in our study had so much more difficulty discriminating target /lu/ from background /mu/ than discriminating target /gu/ from background /mu/, given that both targets differ from the background in both place and manner of articulation. These results are consistent with Nam and Polka's (2016) suggestion that, in early infancy, some phonetic units—particularly stops—have higher perceptual salience than others. This “stop bias” in early infancy appears to be grounded in the acoustic–phonetic properties of stops. At 5–6 months of age, English- and French-learning infants show a general perceptual bias favoring stops over fricatives, despite the fact that /v/ is more common than /b/ in French (Nam & Polka, 2016). The same infants take longer to habituate to the syllable /bas/ than the syllable /vas/, show a more robust novelty preference when the new syllable is /bas/ than when it is /vas/, and show a preference for /bas/ over /vas/ in sequential preferential looking. Nam and Polka proposed that stops serve as natural referents in infants' perception of consonant manner of articulation, due to the fast-amplitude rise time. In the current study, the abrupt shift in signal amplitude for the /gu/ consonant likely enhanced the auditory contrast and made the change from /mu/ to /gu/ easier to detect than the change from /mu/ to /lu/ (Delgutte & Kiang, 1984). Additional research is needed to explore differences in AV enhancement of speech discrimination across speech targets.

Limitations and Future Directions

In order to create an AV speech discrimination task that required participants to rely on the auditory signal (rather than simply responding to visual changes), we had to include incongruent AV speech stimuli (foil trials). These foil trials did not appear to affect infants' ability to learn the task; for the /gu/ target, infants passed training at a similar rate for AV discrimination (0.824) as for AV detection (0.789) and onset–offset cue discrimination (0.866). However, some of the AV discrimination benefit in adults may reflect a bias to respond when there was a change in the visual signal. During the test phase, adults' average false-alarm rates for no-signal and foil trials were 4% and 31%, respectively. In infants, the false-alarm rates were also somewhat higher for foil trials ($M = 37\%$) than for no-signal trials ($M = 22\%$), but the difference was less marked. More research is needed to understand the interaction of bias and benefit in this experimental task.

The visual cues that infants rely on when processing AV speech may depend on the nature of the task. Therefore, it is possible that other tasks would have encouraged infants to use phonetic information for AV speech enhancement in noise more than the tasks used in this study. Although adults in the current study seemed to use phonetic mechanisms of AV enhancement for the discrimination task, listeners can rely on any salient difference between two syllables to discriminate. Thus, discrimination requires an adequate representation of the stimuli but does not require access to phonetic representations of speech (Lalonde

& Holt, 2016). Given their immature speech representations, infants may have defaulted to the temporal cues. In future studies, we will continue to probe infants' use of phonetic information for AV enhancement by testing speech identification. Specifically, we will train infants to identify exemplars that belong to a particular phonetic category, despite variability in production across vowel contexts and repetitions. The identification task will require the listener to evaluate the signal with respect to internal perceptual representations of multimodal speech and, therefore, is more likely to invoke phonetic mechanisms of AV enhancement.

The visual cues that infants rely on when processing AV speech may depend on the stimuli. Therefore, it is possible that other stimuli would have encouraged infants to use phonetic information for AV speech enhancement in noise more than the stimuli used in this study. It may be the case that infants can use their sensitivity to the phonetic information for AV speech enhancement, but only do so when it is the only cue available. In fact, studies examining infants' sensitivity to visual phonetic information have used stimuli without temporal cues (i.e., continuous vowels in AV vowel matching studies; Aldridge et al., 1999; Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003) or have simultaneously provided temporal and phonetic cues (Baart, Vroomen, et al., 2014). On the same note, much of the research on infants' perception of AV speech has focused on their perception of vowels. Future studies will examine infants' use of temporal and phonetic cues to discriminate vowels.

Although detection and discrimination are necessary for speech recognition, the simple tasks and stimuli used in this study are quite removed from recognition and comprehension of fluent speech in natural social interactions. As Lewkowicz (2000b) pointed out, infants' perception of fluent speech may differ from detection of isolated, simple phonemic contrasts. Furthermore, this study focused only on visual articulatory cues. In their interactions with infants, adults provide many nonspeech cues, such as hand gestures, head movements, postural changes, and facial expressions that are temporally related to the auditory speech signal (e.g., Nomikou & Rohlfsing, 2011). Future studies will examine infants' AV speech enhancement using increasingly complex tasks and stimuli. Future studies should also examine the combined benefit of visible articulatory movements and body movements for infant AV speech perception and language learning.

Implications

The results of this study suggest that visual speech can help infants compensate for the acoustically noisy environments in which they learn language. In addition, it provides audiologists and deaf educators with the evidence base for recommending that parents promote and make use of AV skills in support of the language acquisition of an infant or a child with hearing impairment (e.g., Harrison, 2011).

Acknowledgments

This research was funded by National Institute on Deafness and Other Communication Disorders Grants R01 DC000396 (awarded to Werner), F32 DC015387 (awarded to Lalonde), and T32 DC005361 (awarded to Perkel). The authors are grateful for the support of the University of Washington Infant Hearing Lab, particularly Kimberly Gonzalez. Lori Leibold provided feedback on a previous draft of this article.

References

- Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, 2(1), 42–46.
- Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, 53, 115–121.
- Baart, M., Vroomen, J., Shaw, K. E., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, 130(1), 31–43.
- Bahrnick, 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.
- Bahrnick, L. E. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Child Psychology*, 53(2), 180–199.
- Bahrnick, L. E. (1994). The development of infants' sensitivity to arbitrary intermodal relations. *Ecological Psychology*, 6(2), 111–123.
- Bahrnick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, 79(3), 253–270.
- Bahrnick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology*, 41, 352–363.
- Bahrnick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 183–206). Oxford, United Kingdom: Oxford University Press.
- Bahrnick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, 13, 99–102.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed effects models using Eigen and S4 (R package Version 1.1-8) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=lme4>
- Bernstein, L. E., Auer, E. T., Jr., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44, 5–18.
- Bonino, A. Y., Leibold, L. J., & Buss, E. (2013). Effect of signal-temporal uncertainty in children and adults: Tone detection in noise or a random-frequency masker. *The Journal of the Acoustical Society of America*, 134, 4446–4457.
- Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience*, 21, 300–304.
- Buss, E., Hall, J. W., III, & Grose, J. H. (2012). Development of auditory coding as reflected in psychophysical performance. In L. A. Werner, R. R. Fay, & A. N. Popper (Eds.), *Human auditory development* (pp. 107–136). New York, NY: Springer.
- Cabrera, L., & Werner, L. A. (2017). Infants' and adults' use of temporal cues in consonant discrimination. *Ear and Hearing*, 38, 497–506.
- Chandrasekaran, C., Trubanova, A., Stillitano, S., Calpier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLOS Computational Biology*, 5(7), 1–18.
- Delgutte, B., & Kiang, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *The Journal of the Acoustical Society of America*, 75, 897–907.
- Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, 11, 478–484.
- Eggermont, J. J., & Moore, J. K. (2012). Morphological and functional development of the auditory nervous system. In L. A. Werner, R. R. Fay, & A. N. Popper (Eds.), *Human auditory development* (pp. 61–105). New York, NY: Springer.
- Erber, N. P. (1982). *Auditory training*. Washington, DC: Alexander Graham Bell Association for the Deaf and Hard of Hearing.
- Erickson, L. C., & Newman, R. S. (2017). Influences of background noise on infants and children. *Current Directions in Psychological Science*, 26(5), 451–457.
- Eskelund, K., Tuomainen, J., & Anderson, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research*, 208, 447–457.
- Flom, R., & Bahrnick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, 43(1), 238–252.
- Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America*, 109(5), 2272–2275.
- Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108(3, Pt. 1), 1197–1208.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Harrison, M. (2011). Facilitating communication in infants and toddlers with hearing loss. In R. Seewald & A. M. Tharpe (Eds.), *Comprehensive handbook of pediatric audiology* (pp. 631–647). San Diego, CA: Plural.
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers*, 27, 46–51.
- Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, 76, 598–613.
- Jaime, M., Bahrnick, L. E., & Lickliter, R. (2010). The critical role of temporal synchrony in the salience of intersensory redundancy during prenatal development. *Infancy*, 15(1), 61–82.
- Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication*, 44, 19–30.
- Kim, J., & Davis, C. (2014). How visual timing and form information affect speech and non-speech processing. *Brain and Language*, 137, 86–90.
- Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, 18, 65–75.
- Knowland, V. C., Evans, S., Snell, C., & Rosen, S. (2016). Visual speech perception in children with language learning impairments. *Journal of Speech, Language, and Hearing Research*, 59, 1–14.

- Kuhl, P. K., & Meltzoff, A. N. (1982). 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, 361–381.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest: Test for random and fixed effects for linear mixed effects models (R package Version 2.0-2.5). *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>
- Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually-salient speech cues. *Journal of Speech, Language, and Hearing Research*, 58, 135–150.
- Lalonde, K., & Holt, R. F. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *The Journal of the Acoustical Society of America*, 139(4), 1713–1723.
- Lalonde, K., & Werner, L. A. (2019). Perception of incongruent audiovisual English consonants. *PLOS ONE*, 14(3), e0213588. <https://doi.org/10.1371/journal.pone.0213588>
- Lapierre, M. A., Piotrowski, J. T., & Linebarger, D. L. (2012). Background television in the homes of US children. *Pediatrics*, 130(5), 839–846.
- Leibold, L. J., Bonino, A. Y., & Buss, E. (2016). Masked speech perception thresholds in infants, children, and adults. *Ear and Hearing*, 37(3), 345–353.
- Lewkowicz, D. J. (1986). Developmental changes in infants' bisen-sory responses to synchronous durations. *Infant Behavior & Development*, 9, 335–353.
- Lewkowicz, D. J. (1988a). Sensory dominance in infants: I. Six-month-old infants' responses to auditory-visual compounds. *Developmental Psychology*, 24, 155–171.
- Lewkowicz, D. J. (1988b). Sensory dominance in infants: II. Ten-month-old infants' responses to auditory-visual compounds. *Developmental Psychology*, 24, 172–182.
- Lewkowicz, D. J. (1992a). Infants' responses to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior & Development*, 15, 297–324.
- Lewkowicz, D. J. (1992b). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Perception & Psychophysics*, 52, 519–528.
- Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1094–1106.
- Lewkowicz, D. J. (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(4), 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. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, 46(1), 66–77.
- Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental Psychobiology*, 56(2), 292–315.
- Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, 13(11), 470–478.
- Lewkowicz, D. J., & Kraebel, K. S. (2004). The value of multisensor redundancy in the development of intersensory perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 655–678). Cambridge, MA: MIT Press.
- Lewkowicz, D. J., Sowinski, R., & Place, S. (2008). The decline in cross-species intersensory perception in human infants: Underlying mechanisms and its developmental persistence. *Brain Research*, 1242, 291–302.
- Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology*, 16(6), 597–607.
- Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLOS ONE*, 4(3), e4638. <https://doi.org/10.1371/journal.pone.0004638>
- Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98, 185–199.
- Manlove, E. E., Frank, T., & Vernon-Feagans, L. (2001). Why should we care about noise in classrooms and child care settings? *Child & Youth Care Forum*, 30(1), 55–64.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
- Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience*, 25(25), 5884–4893.
- Morronegiello, B. A., Fenwick, K. D., & Chance, G. (1998). Cross-modal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior & Development*, 21(4), 543–553.
- Munhall, K. G., & Vaikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 177–188). Cambridge, MA: MIT Press.
- Nam, Y., & Polka, L. (2016). The phonetic landscape in infant consonant perception is an uneven terrain. *Cognition*, 155, 57–66.
- Nomikou, I., & Rohlffing, K. J. (2011). Language does something: Body action and language in maternal input to 3-month-olds. *IEEE Transactions on Autonomous Mental Development*, 3, 113–128.
- Nozza, R. J., Rossman, R. N., Bond, L. C., & Miller, S. L. (1990). Infant speech-sound discrimination in noise. *The Journal of the Acoustical Society of America*, 87, 339–350.
- Oster, M.-M., & Werner, L. A. (2017). The influence of target and masker characteristics on infants' and adults' detection of speech. *Journal of Speech, Language, and Hearing Research*, 60, 3625–3631.
- Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28, 381–393.
- Paris, T., Kim, J., & Davis, C. (2013). Visual speech form influences the speed of auditory speech processing. *Brain and Language*, 126, 350–356.
- 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, 237–247.
- Patterson, M. L., & Werker, J. F. (2003). Two-month-olds match vowel information in the face and voice. *Developmental Science*, 6(2), 191–196.
- Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, 68, 169–181.
- Picard, M. (2004). Characteristics of the noise, reverberation time and speech-to-noise ratio found in day-care centers. *Canadian Acoustics*, 32(3), 30–31.

- Reynolds, G. D., Bahrick, L. E., Lickliter, R., & Guy, M. W. (2014). Neural correlates of intersensory processing in five-month-old infants. *Developmental Psychobiology*, 56(3), 355–372.
- Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, 33, 2329–2337.
- Saffran, J., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology* (Vol. 2, pp. 59–108). New York, NY: Wiley.
- Schroeder, C. E., & Foxe, J. J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology*, 15, 454–458.
- Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12, 106–113.
- Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, 93, B69–B78.
- Shaw, K. E., & Bortfield, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in Psychology*, 6, 1844.
- Sokol, S. (1978). Measurement of infant visual acuity from pattern reversal evoked potentials. *Vision Research*, 18(1), 33–39.
- Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, 15, 626–636.
- Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific model of perception. *Neuropsychologia*, 50, 1425–1431.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215.
- ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., & Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychological Rehabilitation*, 63, 43–50.
- Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, 96, B13–B22.
- Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, 11, 233–241.
- Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Cross-modal enhancement of signal detection in young and older adults: Does signal content matter? *Ear and Hearing*, 32, 650–655.
- Valliant-Molina, M., & Bahrick, L. E. (2012). The role of intersensory redundancy in the emergence of social referencing in 5 1/2-month-old infants. *Developmental Psychology*, 48(1), 1–9.
- Voss, P. (2005). Noise in children's daycare centres. *Magazine of the European Agency for Safety and Health at Work*, 23–25. Retrieved from <http://akustiknet.dk/publikationer/noisedaycare.pdf>
- Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, 118, 75–83.
- Werner, L. A. (1995). Observer-based approaches to human infant psychoacoustics. In G. M. Klump, R. J. Dooling, R. R. Fay, & W. C. Stebbins (Eds.), *Methods in comparative psychoacoustics* (pp. 135–146). Boston, MA: Birkhauser.
- Werner, L. A. (2013). Infants' detection and discrimination of sounds in modulated maskers. *The Journal of the Acoustical Society of America*, 133(6), 4156–4167.
- Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *The Journal of the Acoustical Society of America*, 119(6), 3940–3949.
- Yehia, H., Rubin, P. E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26, 23–43.
- Zion-Golumbic, E., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *Journal of Neuroscience*, 33, 1417–1426.