Audiovisual speech perception development at varying levels of perceptual processing

Kaylah Lalonde^{a)}

Department of Speech and Hearing Sciences, Indiana University, 200 South Jordan Avenue, Bloomington, Indiana 47405, USA

Rachael Frush Holt

Department of Speech and Hearing Science, Ohio State University, 110 Pressey Hall, 1070 Carmack Road, Columbus, Ohio 43210, USA

(Received 17 February 2015; revised 4 January 2016; accepted 25 March 2016; published online 8 April 2016)

This study used the auditory evaluation framework [Erber (1982). *Auditory Training* (Alexander Graham Bell Association, Washington, DC)] to characterize the influence of visual speech on audiovisual (AV) speech perception in adults and children at multiple levels of perceptual processing. Six- to eight-year-old children and adults completed auditory and AV speech perception tasks at three levels of perceptual processing (detection, discrimination, and recognition). The tasks differed in the level of perceptual processing required to complete them. Adults and children demonstrated visual speech influence at all levels of perceptual processing. Whereas children demonstrated the same visual speech influence at each level of perceptual processing, adults demonstrated greater visual speech influence on tasks requiring higher levels of perceptual processing. These results support previous research demonstrating multiple mechanisms of AV speech processing (general perceptual and speech-specific mechanisms) with independent maturational time courses. The results suggest that adults rely on both general perceptual mechanisms that apply to all levels of perceptual processing and speech-specific mechanisms that apply when making phonetic decisions and/or accessing the lexicon. Six- to eight-year-old children seem to rely only on general perceptual mechanisms across levels. As expected, developmental differences in AV benefit on this and other recognition tasks likely reflect immature speech-specific mechanisms and phonetic processing in children. © 2016 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4945590>]

[MSS]

Pages: 1713–1723

I. INTRODUCTION

Speech is an inherently multimodal signal (Munhall and Vaikiotis-Bateson, 2004; Yehia *et al.*, 1998). The changes in the shape of the vocal tract that give rise to the structured acoustic signal also produce visible speech cues, creating a robust and reliable—and thus highly advantageous—multimodal signal (Rosenblum, 2005; Rowe, 1999; Sumby and Pollack, 1954). Listeners detect, discriminate, and recognize speech-in-noise better by combining the audible portions of the signal with speechreading than by listening alone (Grant and Seitz, 2000; Lalonde and Holt, 2015; MacLeod and Summerfield, 1987; Sumby and Pollack, 1954).

Audiovisual (AV) processing plays a vital role in speech and language development. Infants and children must learn language in complex, noisy listening environments, despite having poor listening strategies and being less adept than adults at controlling their attention and focusing on relevant portions of signals (Bargones and Werner, 1994; Buss *et al.*, 2012; Leibold and Buss, 2013; Leibold and Neff, 2007, 2011; Leibold and Werner, 2006). Research suggests that infants and children can use visual

speech to compensate for degraded auditory cues in noisy environments. Infants use visual speech cues to parse competing auditory speech signals and to learn native acoustic-phonetic category boundaries (Hollich *et al.*, 2005; Teinonen *et al.*, 2008). Children discriminate and recognize AV speech in noise better than A-only speech in noise (e.g., Lalonde and Holt, 2015; Ross *et al.*, 2011).

The developmental literature on AV speech processing suggests a U-shaped trajectory, wherein infants perform more like adults than do children (Jerger *et al.*, 2009; Markovitch and Lewkowicz, 2004). Infants demonstrate remarkable early AV processing abilities. In the first year of life, they are sensitive to the natural correspondence between auditory and visual speech signals (Kuhl and Meltzoff, 1982, 1984; Patterson and Werker, 1999, 2003), are sensitive to McGurk illusions (Burnham and Dodd, 2004; Desjardins and Werker, 2004; Rosenblum *et al.*, 1997), and use visual speech cues to help parse competing auditory speech signals (Hollich *et al.*, 2005). And yet, measures of AV processing in children emphasize adult-child differences and protracted development well into adolescence (Maidment *et al.*, 2015; Ross *et al.*, 2011; Sekiyama and Burnham, 2008; Tremblay *et al.*, 2007; Wightman *et al.*, 2006). Children sometimes fail to show the same visual influence demonstrated by infants (McGurk and MacDonald, 1976; Rosenblum *et al.*, 1997) or younger children (Jerger *et al.*, 2009).

^{a)}Current address: Department of Speech and Hearing Sciences, University of Washington, 1417 N. E. 42nd Street, Seattle, WA 98105, USA. Electronic mail: klalonde@uw.edu

One proposed explanation for this U-shaped developmental trajectory invokes dynamic systems theory (Jerger *et al.*, 2009; Smith and Thelen, 2003). According to this theory, development is characterized by periods of disorganization and behavioral regression, followed by the emergence of new states of organization (Markovitch and Lewkowicz, 2004; Smith and Thelen, 2003). From this perspective, the plateau of the function (at which children demonstrate little AV benefit) reflects a period of transition, rather than a “loss” of skill. For example, between 6 and 9 years of age, early literacy training prompts phonological reorganization (Anthony and Francis, 2005; Morais *et al.*, 1986) making it harder to access visual phonetic information from long-term memory (Jerger *et al.*, 2009). Dynamic systems theory posits that multiple factors interact to cause these non-monotonic functions (Jerger *et al.*, 2009; Smith and Thelen, 2003). In speech perception, these include perceptual skills, phonological knowledge, and general attentional resources (Jerger *et al.*, 2009). Another potential explanation for the U-shaped trajectory is that we lack developmentally appropriate methods for assessing children in certain age ranges. Therefore, developmental differences may actually reflect susceptibility to non-sensory cognitive-linguistic task demands (Lalonde and Holt, 2015).

A number of methods have been used to investigate AV perceptual development, and developmental differences vary depending on the task used. For instance, some measures of AV processing suggest that children integrate auditory and visual signals *more* than adults: children and adolescents judge auditory and visual stimuli as simultaneous across a larger range of asynchronies than adults (Lewkowicz and Flom, 2014; Pons *et al.*, 2014; van Wassenhove *et al.*, 2007) and are more susceptible to AV flash-beep illusions than adults (Innes-Brown *et al.*, 2011; Tremblay *et al.*, 2007). Other measures indicate that children integrate auditory and visual signals *less* than adults: children under 12 years of age are less susceptible to McGurk effects (Hockley and Polka, 1994; Sekiyama and Burnham, 2008) and demonstrate less AV benefit to speech identification/recognition than adults (e.g., Maidment *et al.*, 2015; Ross *et al.*, 2011; Wightman *et al.*, 2006).

Some of the task-related differences in children’s AV processing likely reflect variation in the mechanisms the tasks require and how quickly these mechanisms develop. Multiple mechanisms underlie AV speech perception, including general perceptual mechanisms and speech-specific mechanisms (Klucharev *et al.*, 2003; Vroomen and Stekelenburg, 2011). General perceptual mechanisms refer to the non-phonetic aspects of AV benefits (but see Liberman and Mattingly, 1985; Fowler, 1986). For example, some benefits occur because cues in the visual signal—including correlations between the amplitude envelopes of the auditory and visual speech signals and visible preparatory articulatory gestures (such as lip closure in preparation for bilabials)—reduce temporal uncertainty regarding the auditory signal (Bernstein *et al.*, 2004; Grant and Seitz, 2000). These cues help listeners identify which parts of the auditory signal have the most favorable signal-to-noise ratios (SNRs). Speech-specific (phonetic) mechanisms-related to the

complex, linguistic nature of AV speech processing—are separable from these non-phonetic mechanisms (Klucharev *et al.*, 2003). Speech-specific mechanisms are relevant to tasks that require phonetic decisions (as opposed to acoustic decisions), because phonetic decisions are based on speech-specific articulatory and/or phonetic representations in long-term memory (Diehl and Kluender, 1989; Eskelund *et al.*, 2011; Schwartz *et al.*, 2004; Summerfield, 1992; Tuomainen *et al.*, 2005). Even when both tasks use the same stimuli, the developmental trajectory for tasks that require phonetic decisions is slower than that of tasks that do not (Tremblay *et al.*, 2007).

Variation in methodology and developmental results across AV speech perception studies can shed light on differences in the development of AV speech perception mechanisms. However, without a systematic way to characterize or classify methodologies, such variation makes it difficult to integrate findings, compare abilities across age groups, and draw conclusions about how perception develops. Research on unimodal auditory speech perception development is subject to the same methodological variability and developmentally based methodological restrictions. Erber’s (1982) auditory evaluation framework has provided a useful way to characterize a child’s level of perceptual development. In this framework, response tasks and perceptual abilities are categorized as reflecting one of four levels of perception: (1) detection, which requires only basic sensory reception; (2) discrimination, which requires an adequate representation of the physical properties of the stimulus without necessarily attaching meaning; (3) identification/recognition, which is the ability to label the sound heard; and (4) comprehension, which is the ability to attach meaning to the sounds. These tasks form a hierarchical continuum. Listeners must detect a stimulus and discriminate various stimulus properties in order to recognize that it maps onto something in memory. However, although listeners may recognize a stimulus that they are detecting, they can still be aware of the presence of a stimulus (detect it) without mapping it to a representation in memory (recognition). This work aims to apply Erber’s (1982) framework to AV speech perception development.

There is precedence for using Erber’s (1982) framework to address developmental issues. This framework forms the basic structure of current auditory training protocols for aural (re)habilitation of children with hearing impairment (Erber, 1982; Moog *et al.*, 1995; Rossi, 2003; Stout and Windle, 1986; Wilkes, 2001). It also parallels the model of perceptual development by Aslin and Smith (1988), which characterized three levels of perceptual development: (1) sensory-primitive, corresponding to detection; (2) perceptual representation, corresponding to discrimination; and (3) cognitive-linguistic, corresponding to identification/recognition and comprehension. These levels are not independent and the distinction between these levels is not always clear. For example, one must have adequate sensory primitives to develop perceptual representations, perceptual representations help to develop cognitive-linguistic knowledge, and reorganization of perceptual representations is mediated by emerging cognitive and linguistic skills. Nevertheless, Erber (1982) and Aslin and Smith (1988) used these parallel

frameworks to integrate findings across studies using disparate research methods, compare findings across research studies, compare abilities across age groups, and draw conclusions about development.

The purpose of the current study is to use Erber's (1982) auditory hierarchy and Aslin and Smith's (1988) model of perceptual development to characterize the influence of visual speech on AV speech perception in adults and children at multiple levels of perceptual processing. In doing so, we focus on one slice of the developmental spectrum: those children at the bottom of the U-shaped function whose phonological representations are reorganizing (Anthony and Francis, 2005; Morais *et al.*, 1986) and who are performing less like adults than older and younger children on many AV speech perception measures (e.g., Jerger *et al.*, 2009). We measured 6- to 8-year-old children and adults' auditory and AV speech perception at three levels of perceptual processing: detection, discrimination, and recognition.¹ Importantly, across the three levels of perceptual processing, we kept constant: the stimuli (auditory and AV monosyllabic words), experimental paradigm (yes/no response), SNR, and chance level of performance (50%). The tasks were as similar as possible, with the caveat that discrimination testing inherently requires two stimulus intervals. The major difference between tasks was the level of perceptual processing required to complete them, allowing assessment of the effects of visual speech on performance at varying levels of perceptual processing. Although the levels of perceptual processing are not entirely distinct from one another, we expect there to be greater reliance on more complex perceptual mechanisms as the perceptual processing demands increase, which in turn will be reflected in developmental differences. We expect that adults will demonstrate visual speech influence at all levels of perceptual processing, and that children will show reduced benefit relative to adults at some levels of perceptual processing. Because access to phonological representations can be difficult for children in this age range, we might see greater developmental differences in visual speech influence on higher-level tasks that require more complex perceptual mechanisms than on lower-level tasks.

II. METHOD

A. Participants

Two groups participated in this experiment: 24 adults and 25 six- to eight-year-old children. Two adults were excluded from data analysis due to ceiling performance in one or more A-only conditions. One child participant quit the experiment after 5 min of testing. The analysis includes results for 22 adults (8 male) and 24 children (15 male). Adults ranged from 18 to 31 years of age [mean = 23.29 yr, standard deviation (SD) = 3.92 yr]. Children ranged from 6;1 (years; months) to 7;11 (mean = 7.44 yr, SD = 0.95 yr).

All participants (or their caregivers) reported native English background and passed hearing and vision screenings. Caregivers also reported no developmental problems in their children. We screened hearing in each ear under TDH-39 headphones (Telephonics, Huntington, NY) at 25 dB hearing level (HL) at 250 Hz and 20 dB HL at octave

intervals from 500 through 8000 Hz (re: ANSI, 2004). We screened vision using Precision Vision "Patti Stripes" screening paddles (Patti Stripes Square Wave Grating paddles, 2016) (as in Lalonde and Holt, 2015). This screener requires participants to discriminate between a grey paddle and a paddle with thin black and white vertical stripes, and to identify the striped paddle. All participants passed the screening by correctly identifying an 8.4 cycles per degree, striped paddle on four of five trials. Children also passed a language screening—the Clinical Evaluation of Language Fundamentals-Fourth Edition screening test (CELF-4; Semel *et al.*, 2004).

B. Stimuli and equipment

The stimuli for all tasks consisted of 22 professional AV recorded monosyllabic words from the Lexical Neighborhood Test (LNT; Kirk *et al.*, 1995) spoken by a professional female announcer and described previously in Holt *et al.* (2011). Pilot testing with ten young adults with normal hearing indicated that the 22 words chosen for the experiment were approximately equally difficult to one another in A-only conditions at -7 dB SNR (mean recognition accuracy = 16% to 40%) and in AV conditions at -11 dB SNR (mean recognition accuracy = 40% to 80%). See the supplemental materials for more information on stimulus selection.²

We embedded the speech stimuli in a steady-state noise filtered to match the long-term average spectrum of the speech. The noise always began before and ended after the target stimulus, but the duration of the interval between the onset of the masker and the onset of the stimulus varied from trial to trial (0.341 to 1.091 s), creating temporal uncertainty and thereby increasing the benefit of the visual stimulus on the detection task (Grant and Seitz, 2000).

The visual stimuli included the speaker's entire face, which extended 6 in. in height and 4.75 in. in width on a 19 in. monitor. When participants were seated at a computer, their faces were approximately 16 in. from the monitor. Auditory stimuli were routed through an audiometer to two speakers located at $\pm 45^\circ$ azimuth relative to the listener in a double-walled sound-attenuating chamber. Stimuli were presented and data recorded using E-prime software (Version 2.0; Psychology Software Tools, 2007) on an Intel desktop computer. An E-prime response box with keys marked "Y" for "yes" and "N" for "no" was used to record responses.

C. Tasks

This section describes the three tasks used in the study, to guide the reader as to what occurred across the different stages of the procedures described in Sec. IID. The tasks used to assess each of the three levels of perceptual processing had a yes/no response format, with equal probabilities of "yes" and "no" trials. Each trial began with a simple fixation cue: a grey "X" on a black screen, followed by the stimulus interval. At the end of the stimulus interval, the screen turned yellow as a cue for the listener to respond. Participants responded by pressing the "Y" button for "yes" or the "N" button for "no." They were instructed to attend to both the auditory and visual stimuli, but to *respond based on what they heard*. We chose

to have participants respond based on what they heard because we were interested in the influence of the visual information on auditory performance. These explicit, uniform instructions were necessary for experimental control, particularly with children.

In A-only conditions, participants were presented with a still image of the speaker throughout each stimulus interval. In the AV conditions, the visual signal complemented the auditory signal on 50% of the AV trials (AV-match) and opposed the auditory signal on the remaining trials (AV-mismatch). If the visual signal was always consistent with the auditory information, participants could simply base their responses on the visual speech stimulus. Thus, the AV-mismatch trials were used to avoid ceiling-level performance on tasks requiring lower levels of perceptual processing.

1. Detection

Consistent with Aslin and Smith’s (1988) sensory primitive level, the detection task requires only basic sensory reception (awareness of speech within noise). On each detection trial, a 2.37-s interval of the masker noise followed the fixation cue. This duration corresponds to the longest of the 22 stimulus videos. During half of the trials, an auditory word was embedded in the noise (“yes” trial); during the other half, only the noise was presented (“no” trial). Participants were instructed to press the “Y” button if they heard speech in the noise, and the “N” button if they did not hear speech in the noise.

Examples of each trial type are shown in Table I. The AV condition contained four types of trials presented in a random order from trial to trial. The first two trial types in the AV detection condition (see Table I) were AV-match trials because the correct response based on the visual information matched the correct response based on the auditory information (either both signals were present or both were absent). The other two trial types were AV-mismatch trials because participants would answer incorrectly if they answered based on the visual stimulus. In this sense, the “match” and “mismatch” refer to a match or mismatch in the presence/absence of the speech signal.

2. Discrimination

Following the fixation cue, each discrimination trial contained two 2.37-s noise intervals separated by an 800-ms

TABLE I. Detection trial types and signals.

Trial Type	Auditory Signal	Visual Signal
A-only		
Yes Trial	Noise + “bath”	Still image
No Trial	Noise	Still image
AV		
Match		
Yes Trial	Noise + “bath”	Video of “bath”
No Trial	Noise	Still image
Mismatch		
Yes Trial	Noise + “bath”	Still image
No Trial	Noise	Video of “bath”

silent inter-stimulus interval. Each noise interval contained one LNT word. During half of the trials, the same auditory word was presented in both intervals (“yes” trial); during the other half, different auditory words were presented in each interval (“no” trial). Participants pressed the “Y” button if they heard the same word in both intervals, and the “N” button if they did not hear the same word in both intervals. Consistent with the perceptual representation level given by Aslin and Smith (1988), listeners can rely on any perceptually salient difference between the auditory signals in the two intervals to discriminate, therefore the discrimination task requires an adequate representation of the acoustic properties of the stimulus, but does not require internal phonetic representations of speech.

Table II shows examples of each trial type. The AV condition contained eight types of trials presented in a random order from trial to trial. The first four trial types were AV-match trials because the correct response based on the visual information matched the correct response based on the auditory information. The other four trial types were AV-mismatch trials because participants would answer incorrectly if they answered based on the visual stimulus. Note that some of the AV-match trials had incongruent auditory and visual speech signals (i.e., incongruent-incongruent trials). The “match” and “mismatch” refer to a response match and response mismatch, rather than a match or mismatch between the auditory and visual words.

3. Recognition

On each recognition trial, an LNT word embedded in a 2.37-s interval of the masker noise followed the fixation cue. At the end of the stimulus interval, a probe word was presented auditorily in quiet, cueing the participant to respond. The mean duration of the silent interval between the end of the noise interval and the beginning of the probe was 0.53 s (SD = 0.19 s). Participants could respond as soon as they heard enough of the probe to answer. On half of the trials, the auditory speech signal in the noise masker matched the probe (“yes” trials); on the other half of the trials, it did not

TABLE II. Discrimination trial types and examples. Note: C = A and V Congruent, I = A and V Incongruent.

Trial Type	Auditory Signal	Visual Signal
A-only		
Yes Trial	“bath bath”	Still image
No Trial	“bath want”	Still image
AV		
Match		
C-C Yes Trial	“bath bath”	Video of “bath bath”
C-C No Trial	“bath want”	Video of “bath want”
I-I Yes Trial	“bath bath”	Video of “want want”
I-I No Trial	“bath want”	Video of “want bath”
Mismatch		
C-I Yes Trial	“bath bath”	Video of “bath want”
C-I No Trial	“bath want”	Video of “bath bath”
I-C Yes Trial	“bath bath”	Video of “want bath”
I-C No Trial	“bath want”	Video of “want want”

(“no” trials). Participants pressed the “Y” button if the auditory word in the noise interval matched the probe and the “N” button if the auditory word in the noise interval did not match the probe. The participant must rely on an internal representation of the stimulus in order to compare between the stimulus in noise and the probe in quiet.

Examples of each trial type are shown in Table III. The AV condition contained four types of trials presented in a random order from trial to trial. The first two trial types were AV-match trials because the correct response based on the visual information matched the correct response based on the auditory information (either both the auditory and visual signals matched the probe or neither matched the probe). The other two trial types were AV-mismatch trials because participants would answer incorrectly if they answered based on the visual stimulus.

D. Procedures

Participants completed a pre-test adaptive procedure to determine a participant-specific A-only masked speech detection threshold. Each participant’s threshold served as the participant-specific SNR for further testing in each of six conditions: A-only detection, A-only discrimination, A-only recognition, AV detection, AV discrimination, and AV recognition. Participants were tested in a randomized block design, with one exception: to decrease memory demand for task instructions, participants completed testing in both modalities for a given task before switching to another task.

The experimenter sat in the test booth with children throughout testing, providing verbal instructions and encouragement, but participants usually completed the experiments relatively independently. Children took a break for a game and/or snack between conditions. When necessary, they also took a break half way through the AV conditions (which were longer than the A-only conditions). With adults, the experimenter controlled the experiment from a second room of the testing suite and offered a break between test conditions.

1. Familiarization and training

Before each test condition, participants received instructions and completed familiarization and training trials. Following verbal and written instruction, participants completed ten familiarization trials and up to ten training trials at 0 dB SNR with immediate feedback. The LNT words that

were not selected as test stimuli served as the target stimuli in practice and training. Participants were required to perform correctly on five training trials in a row to qualify for testing in each condition (as in [Holt and Lalonde, 2012](#); [Lalonde and Holt, 2014](#); [Trehub et al., 1995](#)). Participants met this criterion within ten training trials (mean = 5.14 for adults and 5.35 for children). Following training, participants were informed that the background noise would be louder (or for the adaptive portion, that the noise would get softer and louder on each trial), but that they should continue to perform the same task.

2. Adaptive pre-testing procedure

We measured masked thresholds for A-only speech detection using a two-down/one-up, adaptive staircase procedure with a yes/no response that converges on the 70.7% correct point on the psychometric function ([Levitt, 1971](#)). The speech level was fixed at 55 dB A (calibrated daily at the intended location of the listener’s head). The noise level varied adaptively.

One of the 22 target words was presented on each yes trial; the word presented varied randomly from trial to trial. Each threshold search began at –5 dB SNR. The step size was 8-dB until the first reversal, 4-dB until the second reversal, and 2-dB for the remaining six reversals (as in [Hall et al., 2002](#)). The staircase threshold was calculated as the average of the last four reversals. The overall threshold was calculated as the average of three staircase threshold estimates. To avoid conducting fixed-level testing at SNRs with inaudible auditory stimuli, we discarded any estimates less than –20 dB SNR. Further, if the highest and lowest estimates differed by more than 3 dB, a fourth estimate was obtained. In each age group, 15 thresholds were discarded (4 that were less than –20 dB SNR and 11 because the range was greater than 3 dB). When estimates were discarded, a fourth estimate was obtained, and the three most similar estimates were incorporated in the overall average threshold ([Grant and Seitz, 2000](#)).

3. Fixed-level testing

All fixed-level testing was conducted at the individual participant’s A-only masked detection threshold SNR with target speech set to 55 dBA. Participants completed 50 A-only trials per task (equal proportions of yes and no trials randomly intermixed) and 100 AV test trials per task. On each AV trial, there was equal probability that each trial type would be selected (i.e., AV-match yes, AV-match no, AV-mismatch yes, and AV-mismatch no).³

4. Catch trials

In addition to the 100 AV test trials, 12 catch trials were randomly intermixed in each AV condition to assess attention to the visual stimulus. During catch trials, participants were presented with the same stimuli as in test trials and were instructed to use the “Y” and “N” buttons to indicate whether the speaker’s mouth moved during the stimulus interval. The mouth moved during half of the catch trials,

TABLE III. AV recognition trial types and examples.

Trial Type	Auditory Signal	Visual Signal	Probe
A-only			
Yes Trial	“bath”	Still image	“bath”
No Trial	“want”	Still image	“bath”
AV			
Match			
Yes Trial	“bath”	Video of “bath”	“bath”
No Trial	“want”	Video of “want”	“bath”
Mismatch			
Yes Trial	“bath”	Video of “want”	“bath”
No Trial	“want”	Video of “bath”	“bath”

and there was equal probability that a catch trial would occur with each type of stimulus (i.e., AV-match, AV-mismatch). To keep participants from accidentally answering the test question on catch trials (or vice versa), the response box contained two pairs of “Y” and “N” keys, one yellow pair and one red pair. During test trials, the screen turned yellow as a cue to respond; during catch trials, the screen turned red as a cue to respond. Participants were instructed to answer the test question (e.g., “Did you hear a word?” for detection) with the yellow keys when the screen turned yellow, and answer the catch trial question (“Did her mouth move?”) with the red keys when the screen turned red. Participants did not know whether they were completing a test trial or a catch trial until the end of the stimulus, because the intent was to measure the degree to which participants attended to the visual stimulus when they thought they were performing the primary task. Only the pair of buttons appropriate for a given trial (red or yellow) was active, so participants could not make errors by choosing the wrong color.

III. ANALYSIS AND RESULTS

Sensitivity was measured in d' (Green and Swets, 1966). A value of 0.01 was added to hit and false alarm rates of 0.0 and subtracted from hit and false alarm rates of 1.0 (Hautus, 1995), so d' could range from -4.65 to 4.65 . Data were analyzed using a 2 (age: adults, 6- to 8-year-old children) $\times$ 3 (level of processing: detection, discrimination, recognition) $\times$ 2 (modality: A-only, AV) mixed analysis of variance (ANOVA). The AV d' was calculated based on the AV-match trials for each condition. The pattern of results is the same whether we analyzed modality as a 2-level variable (A-only, AV) or a 3-level variable (A-only, AV-match, AV-mismatch). Therefore, for ease of reading, we chose to eliminate the AV-mismatch trials from the analysis.

The results are shown in Fig. 1, for adults (left panel) and children (right panel) for A-only (grey), AV (solid black) conditions. The data represented by the dotted black points will be discussed in Sec. IV. The analysis revealed significant age effects, $F_{1,44} = 40.336$, $p < 0.001$, $\eta_p^2 = 0.478$; level of processing effects, $F_{2,88} = 3.573$, $p = 0.033$, $\eta_p^2 = 0.075$; and modality effects, $F_{1,88} = 93.322$, $p < 0.001$, $\eta_p^2 = 0.679$. *Post hoc* comparisons with Bonferroni corrections for multiple comparisons demonstrated enhanced sensitivity for the AV modality relative to the A-only modality and for adults relative to children ($p < 0.001$). After correcting for multiple comparisons, the differences in performance between levels of perceptual processing were not significant. The main effects were complicated by significant interactions between level of processing and modality, $F_{2,176} = 6.855$, $p = 0.002$, $\eta_p^2 = 0.135$; level of processing and age group $F_{1,176} = 6.633$, $p = 0.002$, $\eta_p^2 = 0.131$, and a 3-way interaction between age, modality, and level of processing, $F_{2,88} = 4.930$, $p = 0.009$, $\eta_p^2 = 0.101$. To investigate these interactions, we analyzed the adult and child data in separate two-way ANOVAs.

FIG. 1. Mean detection, discrimination, and recognition sensitivity for adults (left) and children (right) in A-only (solid grey) and AV (solid black) conditions. The dotted grey lines represent a subset of the AV discrimination trials (AV congruent-congruent). Error bars represent the 95% confidence intervals.

A. Adults

In the adult group, there was a significant effect of modality, $F_{1,42} = 51.625$, $p < 0.001$, $\eta_p^2 = 0.711$. *Post hoc* comparisons with Bonferroni corrections for multiple comparisons demonstrated enhanced sensitivity for AV conditions relative to A-only conditions ($p < 0.001$). These effects were complicated by a significant modality $\times$ processing level interaction, $F_{2,84} = 3.006$, $p < 0.001$, $\eta_p^2 = 0.406$.

To investigate the interaction, paired-samples t-tests were carried out for each level of processing, with modality as the independent variable. The modality effects were significant at all levels of perceptual processing: detection, $t_{21} = 2.856$, $p = 0.009$; discrimination, $t_{21} = 2.727$, $p = 0.013$; and recognition, $t_{21} = 9.505$, $p < 0.001$. The modality effect was greater for the recognition task (Cohen's $d = 0.811$) than the detection (Cohen's $d = 0.609$) and discrimination (Cohen's $d = 0.581$) tasks. The mean AV benefit relative to the A-only condition was greater for recognition (mean [SD] d' difference = 1.344 [0.663]) than detection (0.418 [0.686]) and discrimination (0.462 [0.794]).

B. Children

In the child group, there were significant effects of modality, $F_{1,46} = 19.236$, $p < 0.001$, $\eta_p^2 = 0.653$ and level of processing, $F_{2,46} = 13.726$, $p < 0.001$, $\eta_p^2 = 0.374$. Sensitivity was enhanced in the AV condition relative to the A-only condition. *Post hoc* comparisons with Bonferroni corrections for multiple comparisons indicated greater sensitivity on the lower-level detection task than the higher-level discrimination and recognition tasks ($p = 0.001$). In contrast to the adults, the interaction of task and modality was not significant.

C. Levels of perceptual processing

To assess whether there were greater developmental differences in visual speech influence at higher levels of perceptual processing, separate 2 (age: adults, children) $\times$ 2 (modality: A-only, AV) mixed ANOVAs were carried out for each level of processing. Adults showed better sensitivity than children at all levels of processing: detection, $F_{1,44} = 5.462$, $p = 0.024$, $\eta_p^2 = 0.110$; discrimination, $F_{1,44} = 24.383$, $p < 0.001$, $\eta_p^2 = 0.357$; and recognition, $F_{1,44} = 20.585$, $p < 0.001$, $\eta_p^2 = 0.474$. Sensitivity was better in the AV condition than the A-only condition: detection, $F_{1,44} = 21.891$, $p < 0.001$, $\eta_p^2 = 0.332$; discrimination, $F_{1,44} = 23.947$, $p < 0.001$, $\eta_p^2 = 0.352$; and recognition, $F_{1,44} = 93.111$, $p < 0.001$, $\eta_p^2 = 0.679$. The age $\times$ modality interaction was significant only in the recognition condition, $F_{1,44} = 6.537$, $p = 0.014$, $\eta_p^2 = 0.129$, revealing greater benefit for AV stimuli relative to A-only stimuli in adults (mean = 1.344, SD = 0.663) than in children (mean = 0.781, SD = 0.815) at the recognition level.

D. Adaptive testing

The purpose of the adaptive testing was to choose a participant-specific test SNR that elicited group performance within a range that would allow effects of level of perceptual processing and AV integration to be observed. Although adults' A-only detection performance (mean = 73.8%, SD = 8.66%) was significantly better than children's (mean = 67.1%, SD = 8.81%), $t_{44} = 2.611$, $p = 0.012$, group performance was within an appropriate range. One-sample t-tests indicated that neither the adults' performance, $t_{21} = 1.690$, $p = 0.106$, nor the children's, $t_{23} = 2.014$, $p = 0.056$, differed significantly from the target (70.7%) level. Further, test thresholds were almost identical for adults (mean [SD] = -12.8 [1.25] dB SNR, range -10 to -15 dB SNR) and children (mean [SD] = -12.43 [1.36] dB SNR, range -10 to -15 dB SNR).

E. Catch trials

Although participants did not perform at ceiling, both adult and child groups performed significantly above chance on catch trials at all levels of processing (statistics displayed in Table IV), confirming that the participants were indeed attending to the visual aspect of the stimuli.

IV. DISCUSSION

The purpose of this investigation was to characterize the influence of visual speech on AV speech perception in adults and children at multiple levels of perceptual processing by

carefully controlling the stimuli, experimental paradigm, SNR, and chance level of performance across levels of perceptual processing. Although adults and children demonstrated visual speech influence at all levels of perceptual processing, there were developmental differences in the size of visual speech influence as a function of level of perceptual processing. Whereas children demonstrated the same visual speech benefit at each level of perceptual processing, adults demonstrated greater visual speech benefit on tasks requiring higher levels of perceptual processing. As discussed below, these results suggest that adults and children differ in the degree to which they rely on general perceptual and speech-specific mechanisms for AV speech processing.

A. Mature AV speech processing

Adults benefited from matched visual speech at all levels of perceptual processing, but they benefited more on tasks requiring higher levels of perceptual processing (recognition). Similar AV benefits have been observed in other studies. Young adults demonstrate a 1- to 2.5-dB detection advantage when the visual and auditory stimuli matched relative to when the auditory stimulus is presented alone or an auditory stimulus and a mismatched visual stimulus are presented in a 2-IFC paradigm (Bernstein *et al.*, 2004; Grant and Seitz, 2000). These detection benefits occur because cues in the visual signal reduce temporal uncertainty regarding the auditory signal (Bernstein *et al.*, 2004; Grant and Seitz, 2000), helping listeners identify which parts of the auditory signal have the most favorable SNRs. This is considered a general perceptual (non-phonetic) mechanism, because reduced temporal uncertainty explains AV benefit to nonspeech stimuli just as well as speech stimuli (Eramudugolla *et al.*, 2011) and benefit is much greater with visual speech than with an orthographic representation of the stimulus (Grant and Seitz, 2000). Further, some benefit is observed when the visual speech is replaced with a simpler visual stimulus that provides cues about onset and amplitude variation (a static or dynamic square or Lissajous stimulus) (Bernstein *et al.*, 2004; Tye-Murray *et al.*, 2011).

General perceptual mechanisms certainly do not only apply to detection tasks. Their ability to effectively improve the SNR of the auditory signal enhances intelligibility and thus performance on higher-level speech perception tasks, such as discrimination and recognition (Schwartz *et al.*, 2004). For example, ambiguous visual speech signals help adults detect pre-voicing cues, allowing better discrimination between syllables with and without plosives and thus, improving syllable identification accuracy (Schwartz *et al.*,

TABLE IV. Catch trial accuracy and t-values from one-sample t-tests comparing catch trial performance to chance.

Level of perceptual processing	Adults		Children	
	Mean (SD)	t	Mean (SD)	t
Detection	92.05% (12.98%)	15.193 ^a	81.25% (19.23%)	7.960 ^a
Discrimination	96.41% (11.63%)	18.726 ^a	79.35% (22.66%)	6.344 ^a
Recognition	92.42% (16.65%)	11.952 ^a	90.97% (14.31%)	14.026 ^a

^a $p < 0.001$.

2004). However, there is evidence that additional mechanisms are likely involved in perceptual processing of more complex, higher-level tasks and signals. The recognition task requires participants to access phonetic and/or lexical representations in long-term memory. Some investigators have argued that the additional speech-specific processing that is used for the recognition level [Aslin and Smith's (1988) cognitive linguist level] requires an additional processing mechanism beyond the general perceptual mechanism; one that is speech-specific and related to the linguistic nature of speech (Driver and Noesselt, 2008; Miller and D'Esposito, 2005).

This multiple-mechanism account of AV speech processing is supported by neuroanatomical and neurophysiological studies demonstrating that multiple neural pathways underlie various AV phenomena (Arnal *et al.*, 2009; Driver and Noesselt, 2008; Eskelund *et al.*, 2011; Kayser *et al.*, 2007). Pathways involved in evaluating sensory correspondence and simple reactions to multimodal stimulation (such as speech and non-speech detection) differ from those involved in processing more complex multimodal relations (such as speech identification/recognition and achieving perceptual fusion) (Eskelund *et al.*, 2011; Miller and D'Esposito, 2005). Like Erber (1982) and Aslin and Smith's (1988) perceptual frameworks, these pathways/systems may be organized hierarchically, meaning that there must be sufficient evidence for sensory correspondence (which allows general perceptual benefit) for that information to converge at the phonetic level (which allows speech-specific benefit) (Vroomen and Stekelenburg, 2011).

The current study supports the proposal that adults rely on multiple mechanisms to obtain AV speech benefit: those that rely on general perceptual mechanisms and those related to the linguistic nature of speech. Our results showed approximately equal benefit for detection and discrimination. We might conclude that discrimination benefit resulted from general perceptual mechanisms of integration. After all, one can discriminate speech based on only the physical features of the stimulus. However, the discrimination results may reflect an important methodological difference between the AV-match discrimination condition and the AV-match detection and recognition conditions: The AV-match discrimination condition included both trials with congruent auditory and visual stimuli and trials with incongruent auditory and visual stimuli (but congruent correct responses; see Table II), whereas the matched AV stimuli in the detection and recognition tasks were always congruent. When we define the AV discrimination trials as only those with congruent auditory and visual signals in both intervals (dotted black points in Fig. 1), adults' mean AV discrimination benefit increases from 0.462 to 1.313 (d' difference). When we define AV discrimination benefit this way, adults demonstrate greater benefit on the discrimination task than the detection task (mean d' benefit = 0.418 for detection, 1.313 for discrimination and 1.345 for recognition). These additional benefits may relate to the more subtle/complex multimodal relations that must be processed for discrimination (but not detection). It is possible that speech discrimination

activates phonetic representations, and thus speech-specific mechanisms in adults.

B. AV speech processing development

Like adults, children demonstrated AV benefit at all levels of perceptual processing. However, they demonstrated the same amount of benefit across levels of perceptual processing. The pattern held when we defined the AV-match condition as only the congruent-congruent trials (Fig. 1, dotted black line). These results might be interpreted to suggest that AV speech benefits in 6- to 8-year-old children resulted from a single general perceptual mechanism that applies across all levels of perceptual processing; in contrast to the adults, children did not show evidence of using a speech-specific mechanism to achieve greater AV-match benefit at higher levels of perceptual processing. This general perceptual mechanism allows children to attain the same degree of AV benefit to detection and discrimination as adults, but the lack of a speech-specific mechanism in children explains why adults demonstrated greater AV recognition benefit than children.

Dynamic systems theory posits that developmental differences in visual speech influence might result from differences in perceptual skills, phonological knowledge, and/or general attentional resources (Jerger *et al.*, 2009). This study highlighted the role of developmental differences in phonological knowledge/representations in explaining adult-child differences in AV benefit: developmental differences in AV benefit were only observed on the recognition task (the task that requires phonetic decisions). Because children do not have as much prior linguistic experience as adults, they may be less prone to strategically use the visual phonetic information in the recognition task. Alternatively, children in this age range may not be able to readily access and retrieve the visual phonological knowledge needed for speech-specific mechanisms of AV benefit (Jerger *et al.*, 2009). As noted, early literacy training prompts phonological reorganization at this age (Jerger *et al.*, 2009; Smith and Thelen, 2003), making visual phonological information harder to access from long-term memory (Jerger *et al.*, 2009) and perhaps forcing children to rely on general perceptual mechanisms of AV benefit. On the basis of this interpretation, one might expect that preliterate children will show more evidence of a speech-specific mechanism than the 6- to 8-year-old children in the current investigation, because they have not yet reached this state of disorganization. However, it is also possible that young children's representations are not detailed and robust enough (or that children do not have enough linguistic experience) to use speech-specific mechanisms before this literacy-induced reorganization.

A previous investigation of ours partially addresses this question (Lalonde and Holt, 2015). We assessed whether 3- and 4-year-olds rely on visually salient speech cues and visual phonological knowledge to obtain AV speech discrimination and recognition benefit. Both 3- and 4-year-old children used visually salient speech cues to obtain AV benefit to discrimination: they demonstrated AV benefit for a visually salient place contrast, but not a less visually salient

manner contrast. Although both groups of children demonstrated AV benefit on the recognition task, only the 4-year-old children showed evidence of using the salient visual speech cues: 4-year-olds' substitution errors (like adults') were more likely to involve visually confusable phonemes in the AV condition than in the A-only condition (Lalonde and Holt, 2015).

It is unclear whether AV speech discrimination benefits result from general perceptual mechanisms or speech-specific mechanisms, because one can discriminate speech based on only the physical features of the stimulus, and visual saliency would likely affect both speech-specific and general perceptual mechanisms. At the recognition level, however, it is not possible to recognize speech based on only the physical features of the stimulus, suggesting that 4-year-olds used visual phonological knowledge to achieve AV benefit. The picture that emerges from these two studies is that 3-year-old and 6- to 8-year-old children rely only on general perceptual mechanisms for AV-benefit, while 4-year-old children and adults can rely on speech-specific mechanisms. Jerger *et al.* (2009) also demonstrated influence of visual speech on phonological processing in 4-year-olds, but not 5- to 9-year-olds. In that study, 4-year-old and 10- to 14-year-old children demonstrated visual speech influence in the form of greater facilitation in naming time for congruent AV than congruent A-only speech "distractors," but failed to demonstrate any visual speech influence in 5- to 9-year-old children (Jerger *et al.*, 2009). From a dynamic systems perspective, these studies suggest that perhaps our pre-literate 4-year-olds are in a period of relative stability in terms of phonological organization. Perhaps the 3-year-olds in the previous study were in a state of phonological reorganization secondary to the vocabulary growth spurt that begins around 18 months and extends to middle childhood (Walley *et al.*, 2003). More research assessing AV speech perception across levels of processing (using a design similar to the current investigation that controls the experimental method across the levels), and across a larger age range is necessary to more fully understand AV speech processing mechanism development.

V. SUMMARY AND FUTURE DIRECTIONS

Adults demonstrated greater visual speech influence on a higher-level recognition task than a lower-level detection task. Children demonstrated the same visual speech influence across all levels of perceptual processing. Results support the contention that both general perceptual and speech-specific mechanisms contribute to adults' AV speech processing benefit (Klucharev *et al.*, 2003). Adults rely on general perceptual mechanisms for detection benefits and rely on both speech-specific and general perceptual mechanisms for recognition benefit. Because there are multiple mechanisms (Eskelund *et al.*, 2011; Klucharev *et al.*, 2003), we must consider both low-level, general perceptual and higher-level, speech-specific AV interactions in spoken language processing. Six- to eight-year-old children seem to rely only on general perceptual mechanisms across levels. Children only demonstrated less visual speech influence than

adults on the higher-level recognition task, suggesting that developmental differences in AV benefit on this and other recognition tasks likely reflect immature speech-specific mechanisms and phonetic processing in children. These results reinforce the importance of assessing AV speech processing and its development at multiple processing levels: each level reflects different degrees of contribution from various mechanisms and pathways.

This study demonstrates that an AV speech processing framework based on Erber's (1982) auditory evaluation framework allows us to develop a more comprehensive understanding of AV speech processing mechanisms. It is possible to make the same comparisons across levels of perceptual processing using any method of assessing auditory and AV speech perception, and therefore, to apply this framework across the lifespan. In future research, we will observe how the development of AV speech perception interacts with phonetic development by assessing auditory and AV speech perception across levels of perceptual processing, using measures developmentally appropriate for infants, younger and older children, and younger and older adults. Future work will also use stimulus manipulations (such as backward and sine wave speech) and assess lexical effects at each level of perceptual processing to further explore developmental differences in AV speech processing mechanisms. As noted in Sec. I, the relationship between the levels of processing is complex. More work is necessary to fully understand the intricacies of how these mechanisms contribute to different tasks in adults and across development.

ACKNOWLEDGMENTS

This work was supported by a National Institute of Health–National Institute on Deafness and Other Communication Disorders predoctoral and postdoctoral training grants (NIH-NIDCD T32 DC00012, T32 DC005361) and the Ronald E. McNair Research Foundation. We are grateful for the contributions of Larry Humes, Jennifer Lentz, and David Pisoni, who contributed to the experimental design and interpretation of results; Karen Forrest, who allowed us to collect data in her laboratory; Kyle Enlow, Dave Montgomery, and Luis Hernandez for technical assistance; Lori Leibold and Andrea Hillock Dunn for advice regarding pediatric methodology; and Tessa Bent for comments on a previous draft of the manuscript. Portions of this work were presented at the annual meeting of the American Auditory Society, Scottsdale, AZ (March 2014) and the Acoustical Society of America meeting, Indianapolis, IN (October 2014).

¹We use the term "recognition" (as opposed to "identification") to describe the task in the current experiment, in keeping with the memory literature's definitions for yes/no paradigms (Craik and Lockhart, 1972).

²See supplementary material at <http://dx.doi.org/10.1121/1.4945590> for more information on stimulus selection.

³In all but the AV discrimination condition, participants completed exactly 25 trials of each type. In the AV discrimination condition, there were 8 trial types distributed over 100 trials. Therefore, there was some variation in the number of trials of each type across participants. All d' scores were based on a minimum of 16 yes and 16 no trials (mean = 25 trials each),

- with one exception. One participant completed only 2 congruent-incongruent no trials. Mean d' scores and the pattern of results were not different when this participant was excluded from the analysis. Therefore all of the subjects were included.
- ANSI (2004). S3.21-2004, *Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).
- Anthony, J. L., and Francis, D. J. (2005). "Development of phonological awareness," *Curr. Dir. Psychol. Sci.* **14**(5), 255–259.
- Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A.-L. (2009). "Dual neural routing of visual facilitation in speech processing," *J. Neurosci.* **29**(43), 13445–13453.
- Aslin, R. N., and Smith, L. B. (1988). "Perceptual development," *Annu. Rev. Psychol.* **39**, 435–473.
- Bargones, J. Y., and Werner, L. A. (1994). "Adults listen selectively; infants do not," *Psychol. Sci.* **5**(3), 170–174.
- Bernstein, L. E., Auer, E. T., and Takayanagi, S. (2004). "Auditory speech detection in noise enhanced by lipreading," *Speech Commun.* **44**, 5–18.
- Burnham, D., and Dodd, B. (2004). "Auditory-visual integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect," *Dev. Psychobiol.* **45**, 204–220.
- Buss, E., Hall, J. W., and Grose, J. H. (2012). "Development of auditory coding as reflected in psychophysical performance," in *Human Auditory Development*, edited by L. A. Werner, R. R. Fay, and A. N. Popper (Springer, New York), pp. 107–136.
- Craik, F. I. M., and Lockhart, R. S. (1972). "Levels of processing: A framework for memory research," *J. Verbal Learn. Verbal Behav.* **11**, 671–684.
- Desjardins, R. N., and Werker, J. F. (2004). "Is the integration of heard and seen speech mandatory for infants?," *Dev. Psychobiol.* **45**, 187–203.
- Diehl, R. L., and Kluender, K. R. (1989). "On the objects of speech perception," *Ecol. Psychol.* **1**(2), 121–144.
- Driver, J., and Noesselt, T. (2008). "Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments," *Neuron* **57**, 11–23.
- Eramudugolla, R., Henderson, R., and Mattingly, J. B. (2011). "Effects of audio-visual integration on the detection of masked speech and non-speech sounds," *Brain Cognit.* **75**, 60–66.
- Erber, N. P. (1982). "Glenondale auditory screening procedure," in *Auditory Training* (Alexander Graham Bell Association, Washington, DC), pp. 47–71.
- Eskelund, K., Tuomainen, J., and Anderson, T. S. (2011). "Multistage audio-visual integration of speech: Dissociating identification and detection," *Exp. Brain Res.* **208**, 447–457.
- Fowler, C. A. (1986). "An event approach to the study of speech perception from a direct-realist perspective," in *Status Report on Speech Research*, edited by I. G. Mattingly and N. O'Brien (Haskins Laboratories, New Haven, CT), pp. 139–169.
- Grant, K. W., and Seitz, P. F. (2000). "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.* **108**(3), 1197–1208.
- Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Wiley, New York).
- Hall, J. W., Grose, J. H., Buss, E., and Dev, M. B. (2002). "Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children," *Ear Hear.* **23**, 159–165.
- Hautus, M. J. (1995). "Corrections for extreme proportions and their biasing effects on estimated values of d' ," *Behav. Res. Methods Instrum. Comput.* **27**, 46–51.
- Hockley, N., and Polka, L. (1994). "A developmental study of audiovisual speech perception using the McGurk paradigm," *J. Acoust. Soc. Am.* **96**, 3309.
- Hollich, G., Newman, R. S., and Jusczyk, P. W. (2005). "Infants' use of synchronized visual information to separate streams of speech," *Child Dev.* **76**, 598–613.
- Holt, R. F., Kirk, K. I., and Hay-McCutcheon, M. J. (2011). "Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants," *J. Speech. Lang. Hear. Res.* **54**, 632–657.
- Holt, R. F., and Lalonde, K. (2012). "Assessing toddlers' speech discrimination," *Int. J. Pediatr. Otorhinolaryngol.* **76**, 680–692.
- Innes-Brown, H., Barutcu, A., Shivdasani, M. N., Crewther, D. P., Grayden, D. B., and Paolini, A. (2011). "Susceptibility to the flash-beep illusion is increased in children compared to adults," *Dev. Sci.* **14**(5), 1089–1099.
- Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., and Abdi, H. (2009). "Developmental shifts in children's sensitivity to visual speech: A new multimodal picture word task," *J. Exp. Child Psychol.* **102**(1), 40–59.
- Kayser, C., Pektov, C. I., Augath, M., and Logothetis, N. K. (2007). "Functional imaging reveals visual modulation of specific fields in auditory cortex," *J. Neurosci.* **27**(8), 1824–1835.
- Kirk, K. I., Pisoni, D. B., and Osberger, M. J. (1995). "Lexical effects on spoken word recognition by pediatric cochlear implant users," *Ear Hear.* **16**(5), 470–481.
- Klucharev, V., Möttönen, R., and Sams, M. (2003). "Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception," *Cognit. Brain Res.* **18**, 65–75.
- Kuhl, P. K., and Meltzoff, A. N. (1982). "The bimodal perception of speech in infancy," *Science* **218**(4577), 1138–1141.
- Kuhl, P. K., and Meltzoff, A. N. (1984). "The intermodal representation of speech in infants," *Infant Behav. Dev.* **7**, 361–381.
- Lalonde, K., and Holt, R. F. (2014). "Cognitive and linguistic sources of variance in 2-year-olds' speech-sound discrimination: A preliminary investigation," *J. Speech. Lang. Hear. Res.* **57**, 308–326.
- Lalonde, K., and Holt, R. F. (2015). "Preschoolers benefit from visually-salient speech cues," *J. Speech. Lang. Hear. Res.* **58**, 135–150.
- Leibold, L. J., and Buss, E. (2013). "Children's identification of consonants in a speech-shaped noise or a two-talker masker," *J. Speech. Lang. Hear. Res.* **56**, 1144–1155.
- Leibold, L. J., and Neff, D. L. (2007). "Effects of masker-spectral variability and masker fringes in children and adults," *J. Acoust. Soc. Am.* **121**(6), 3666–3676.
- Leibold, L. J., and Neff, D. L. (2011). "Masking by a remote-frequency noise band in children and adults," *Ear Hear.* **32**, 663–666.
- Leibold, L. J., and Werner, L. A. (2006). "Effect of masker-frequency variability on the detection performance of infants and adults," *J. Acoust. Soc. Am.* **119**, 3960–3970.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- Lewkowicz, D. J., and Flom, R. (2014). "The audiovisual temporal binding window narrows in early childhood," *Child Dev.* **85**(2), 685–694.
- Lieberman, A. M., and Mattingly, I. G. (1985). "The motor theory of speech perception revised," *Cognition* **21**(1), 1–36.
- MacLeod, A., and Summerfield, Q. (1987). "Quantifying the contribution of vision to speech perception in noise," *Br. J. Audiol.* **21**, 131–141.
- Maidment, D. W., Kang, H. J., Stewart, H. J., and Amitay, S. (2015). "Audiovisual integration in children listening to spectrally degraded speech," *J. Speech. Lang. Hear. Res.* **58**, 61–68.
- Markovitch, S., and Lewkowicz, D. J. (2004). "U-shaped functions: Artifact or hallmark of development?," *J. Cognitive Dev.* **5**(1), 113–118.
- McGurk, H., and MacDonald, J. (1976). "Hearing lips and seeing voices," *Nature* **264**(5588), 746–748.
- Miller, L. M., and D'Esposito, M. (2005). "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech," *J. Neurosci.* **25**(25), 5884–4893.
- Moog, J. S., Biedenstein, J., and Davidson, L. (1995). *Speech Perception Instructional Curriculum and Evaluation (SPICE)* (Central Institute for the Deaf, St. Louis, MO).
- Morais, J., Bertelson, P., Cary, L., and Algeria, J. (1986). "Literacy training and speech segmentation," *Cognition* **24**, 45–64.
- Munhall, K. G., and Vaikotis-Bateson, E. (2004). "Spatial and temporal constraints on audiovisual speech perception," in *The Handbook of Multisensory Processes*, edited by G. Calvert, C. Spence, and B. E. Stein (The MIT Press, Cambridge, MA), pp. 177–188.
- Patterson, M. L., and Werker, J. F. (1999). "Matching phonetic information in lips and voice is robust in 4.5-month-old infants," *Infant Behav. Dev.* **22**, 237–247.
- Patterson, M. L., and Werker, J. F. (2003). "Two-month-olds match vowel information in the face and voice," *Dev. Sci.* **6**(2), 191–196.
- Patti Stripes Square Wave Grating Paddles [Apparatus] (2016). La Salle, IL: Precision Vision. Available from <http://precision-vision.com/product/pattistripessquarewavegratingpaddles/> (Last viewed April 1, 2016).
- Pons, F., Andreu, L., Sanz-Torrent, M., Buil-Legaz, L., and Lewkowicz, D. J. (2014). "Perception of audio-visual speech synchrony in Spanish-speaking children with and without specific language impairment," *J. Child Lang.* **40**(3), 687–700.
- Psychology Software Tools. (2007). "E-Prime 2.0 [computer program]" (Psychology Software Tools: Pittsburgh, PA).

- Rosenblum, L. D. (2005). "Primacy of multimodal speech perception," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishing Ltd., Malden, MA), pp. 51–78.
- Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997). "The McGurk effect in infants," *Percept. Psychophys.* **59**, 347–357.
- Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., and Foxe, J. J. (2011). "The development of multisensory speech perception continues into the late childhood years," *Eur. J. Neurosci.* **33**, 2329–2337.
- Rossi, K. (2003). *Learn to Talk Around the Clock* (AG Bell, Washington, DC).
- Rowe, C. (1999). "Receiver psychology and the evolution of multicomponent signals," *Anim. Behav.* **58**, 921–931.
- Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," *Cognition* **93**, B69–B78.
- Sekiyama, K., and Burnham, D. (2008). "Impact of language on development of auditory/visual speech perception," *Dev. Sci.* **11**(2), 306–320.
- Semel, E., Wiig, E., and Secord, W. (2004). "Clinical evaluation of language fundamentals, fourth edition—Screening test (CELF-4 screening test)" (The Psychological Corporation/A Harcourt Assembly Company, Toronto, Canada).
- Smith, L. B., and Thelen, E. (2003). "Development as a dynamic system," *Trends Cogn. Neurosci.* **7**(8), 343–348.
- Stout, G. G., and Windle, J. V. E. (1986). *A Developmental Approach to Successful Listening: DASL* (Developmental Approach to Successful Listening, Houston, TX).
- Sumby, W. H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.* **26**(2), 212–215.
- Summerfield, Q. (1992). "Lipreading and audio-visual speech perception," *Philos. Trans. R. Soc. B* **335**(1273), 71–78.
- Teinonen, T., Aslin, R. N., Alku, P., and Csibra, G. (2008). "Visual speech contributions to phonetic learning in 6-month-old infants," *Cognition* **108**, 850–855.
- Trehub, S. E., Schneider, B. A., and Henderson, J. L. (1995). "Gap detection in infants, children," *J. Acoust. Soc. Am.* **98**, 2532–2541.
- Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., and Théoret, H. (2007). "Speech and non-speech audio-visual illusions: A developmental study," *PLoS One* **2**, e742.
- Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (2005). "Audio-visual speech perception is special," *Cognition* **96**, B13–B22.
- Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., and Hale, S. (2011). "Cross-modal enhancement of speech detection in young and older adults: Does signal content matter?," *Ear Hear.* **32**, 650–655.
- van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia* **45**, 598–607.
- Vroomen, J., and Stekelenburg, J. J. (2011). "Perception of intersensory synchrony in audiovisual speech: Not that special," *Cognition* **118**, 75–83.
- Walley, A. C., Metsala, J. L., and Garlock, V. M. (2003). "Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability," *Reading Writing* **16**, 5–20.
- Wightman, F., Kistler, D., and Brungart, D. (2006). "Informational masking of speech in children: Auditory-visual integration," *J. Acoust. Soc. Am.* **119**(6), 3940–3949.
- Wilkes, E. M. (2001). *Cottage Acquisition Scales for Listening, Language and Speech: User's Guide* (Sunshine Cottage School for Deaf Children, San Antonio, TX).
- Yehia, H., Rubin, P. E., and Vatikiotis-Bateson, E. (1998). "Quantitative association of vocal-tract and facial behavior," *Speech Commun.* **26**, 23–43.