The development of gaze to a speaking face

Julia Irwin, Lawrence Brancazio, and Nicole Volpe

Citation: The Journal of the Acoustical Society of America **141**, 3145 (2017); doi: 10.1121/1.4982727
View online: http://dx.doi.org/10.1121/1.4982727
View Table of Contents: http://asa.scitation.org/toc/jas/141/5
Published by the Acoustical Society of America

The development of gaze to a speaking face

Julia Irwin,[1,a)] Lawrence Brancazio,[2,b)] and Nicole Volpe[2]

[1]*Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511, USA*
[2]*Southern Connecticut State University, 501 Crescent Street, New Haven, Connecticut 06515, USA*

(Received 12 October 2016; revised 16 March 2017; accepted 17 April 2017; published online 8 May 2017)

When a speaker talks, the visible consequences of what they are saying can be seen. Listeners are influenced by this visible speech both in a noisy listening environment and even when auditory speech can easily be heard. While visible influence on heard speech has been reported to increase from early to late childhood, little is known about the mechanism that underlies this developmental trend. One possible account of developmental differences is that looking behavior to the face of a speaker changes with age. To examine this possibility, the gaze to a speaking face was examined in children from 5 to 10 yrs of age and adults. Participants viewed a speaker's face in a range of conditions that elicit looking: in a visual only (speech reading) condition, in the presence of auditory noise (speech in noise) condition, and in an audiovisual mismatch (McGurk) condition. Results indicate an increase in gaze on the face, and specifically, to the *mouth* of a speaker between the ages of 5 and 10 for all conditions. This change in looking behavior may help account for previous findings in the literature showing that visual influence on heard speech increases with development. © *2017 Acoustical Society of America.*

[http://dx.doi.org/10.1121/1.4982727]

[MSS] Pages: 3145–3150

I. INTRODUCTION

Visual information about speech influences what listeners hear (Desjardins *et al.*, 1997; Lachs and Pisoni, 2004; McGurk and MacDonald, 1976; MacDonald and McGurk, 1978; MacDonald *et al.*, 2000; Reisberg *et al.*, 1987). This visible articulatory information on a speaker's face is thought to be a central part of typical perceptual development and to foster native language acquisition (Legerstee, 1990) and has been demonstrated in infancy (Burnham and Dodd, 1998; Meltzoff and Kuhl, 1994; Rosenblum *et al.*, 1997). Both understanding and producing speech is likely influenced by experience with seeing other talkers. For example, blind individuals display differences in speech perception and production in comparison to sighted individuals (Ménard *et al.*, 2009). Further, visible speech can influence what is heard both in the context of a noisy background (e.g., Sumby and Pollack, 1954) and even in clear listening conditions. One powerful example of visual influence on what is heard in clear listening conditions is mismatched audiovisual (AV) speech. McGurk and MacDonald (1976) first discovered this by presenting mismatching audio and video consonant-vowel tokens to perceivers. Perceivers watching these dubbed productions sometimes reported hearing consonants that combined the places of articulation of the visual and auditory tokens (e.g., a visual /ba/ + auditory /ga/ heard as /bga/), "fused" the two places (e.g., a visual /ga/ + auditory /ba/ heard as /da/), or reflected the visual place information alone (a visual /va/ + auditory /ba/ heard as /va/).

In their classic paper described above, McGurk and MacDonald (1976) also reported that preschoolers (3–4 yr olds) and school-aged children (7–8 yr olds) are less influenced by visual speech information than are adults (also see Sekiyama and Burnham, 2008). Since then, a number of studies revealed that both visual influence in the context of a McGurk task and speechreading ability (identification of a syllable from visual information only) increases with age (Desjardins *et al.*, 1997; Hockley and Polka, 1994; LaLonde and Frush Holt, 2014; Massaro, 1984; Massaro *et al.*, 1986; Ross *et al.*, 2011; Tremblay *et al.*, 2007). According to this prior work, increased visual influence with development could be due to experience with producing speech sounds (motor experience: Desjardins *et al.*, 1997 report visual influence for children *if* they can produce the sound that they see on another speaker's face), ongoing perceptual learning/tuning with respect to visual speech during childhood (ability to pick up phonetic information from the visual signal: Ross *et al.*, 2011; Hockley and Polka, 1994), or that younger children were less attentive to the visual source, leading to an attenuated visual effect (changes in attention to the speaking face: Massaro, 1984).

Clearly, access to visual speech is necessary for AV speech perception. Adult listeners have been demonstrated to exhibit reduced visual influence on what is heard when asked to engage in a task that requires attention to another visual stimulus (e.g., attention to a shape on the face of the speaker; Alsius *et al.*, 2005). Previous research using eye tracking to examine gaze to a speaking face in adults indicates a pattern of reduced gaze on the eyes and increased gaze on the nose and mouth in the context of auditory noise

[a)] Also at: Southern Connecticut State University, 501 Crescent Street, New Haven, CT 06515, USA. Electronic mail: julia.irwin@haskins.yale.edu
[b)] Also at: Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA.

(Buchan *et al.*, 2008; Vatikiotis-Bateson *et al.*, 1998; Yi *et al.*, 2013). In clear auditory listening conditions, Lansing and McConkie (2003) found that before and after a sentence is spoken perceivers gaze to the eyes of a speaker, while gaze is largely toward the mouth of the speaker during the production of the sentence. Barenholtz *et al.* (2016) report that gaze to the speaker's mouth is greater during speech tasks for an unfamiliar language in monolinguals, but not in bilinguals. Looking behavior to the face of a speaker is likely also a factor in how much children are influenced by visible speech. In particular, if the developmental trends are due to attention to the face of the speaker, this could be reflected in gaze (as indicated by the work in adults of Vatkiotis-Bateson *et al.*, 1998 and Buchan *et al.*, 2008) or a more central attentional difference (e.g., Alsius *et al.*, 2005). A few studies have examined gaze to the face in infants and children. Lewkowicz and Hansen-Tift (2012) showed a shift in focus toward the mouth from the eyes of a speaker corresponding to the onset of producing speech in infants, suggesting that gaze to the mouth increases as children begin to speak. Typically developing children have been compared to those with autism spectrum disorders (ASDs; a social-communication disability whose hallmark is atypical gaze to other's faces and is frequently accompanied by delays in spoken language; Irwin and Brancazio, 2014; Johnels *et al.*, 2014). These studies indicate that typically developing children look more on the mouth of the speaker than children with an ASD. Yet to be examined is how a pattern of gaze to a speaking face in *typically developing* school-aged children changes with development (which in turn might account for the developmental differences in visual influence reported in previous research). Thus, using a cross-sectional design and visual tracking methodology we sought to provide more data on pattern of gaze to the speaking face in development during a set of auditory, visual, and AV tasks in children from 5 to 10 yrs of age and in adult participants.

II. METHOD

A. Participants

Seventy-four participants completed this study: 54 children, sixteen 5–6 yr olds (10 girls and 6 boys, mean age 6 yrs old), twenty 7–8 yr olds (6 girls and 14 boys, mean age 7 yrs, 11 months), eighteen 9–10 yr olds (9 girls and 9 boys, mean age 10 yrs old), and 20 adults (10 women and 10 men, mean age 22 yrs, 1 month) recruited from the community in the greater New Haven, CT area. The participants were reported by their parents or (in the case of the adults) by self-report to have normal or corrected-to-normal hearing and vision. In addition, all participants were reported by their parents or by self-report to have no history of vision, hearing, speech, language, or learning problems.

III. MATERIALS

A. Speech stimuli

The speech stimuli were created from a recording of the productions of an adult male, monolingual, native speaker of American English. This speaker was audio- and video-recorded in a sound-attenuated recording booth producing a randomized list of the consonant-vowel (CV) syllables /ma/, /na/, /ga/. The speaker produced each CV with as similar duration and intonation as possible.

1. Visual only (speechreading) stimuli

The visual only stimuli were silent versions of the speaker producing /ma/ and /na/. In this condition, participants were told that they would see a man saying some sounds that they would not be able to hear, and then asked to report what they thought the man was saying, for a total of 20 trials.

2. Speech in noise stimuli

Noise was added to the 60 dB /ma/ and /na/ tokens to create a range of signal-to-noise ratio levels at 5, 0, −5, −10, −15, and −20 dB, from less to more noisy. The AV stimuli were the same auditory tokens with video of the speaker producing the same CV syllables. For both auditory and AV stimuli, there were 24 trials.

3. AV match and mismatch (McGurk) stimuli

The mismatch stimuli were dubbed by placing the audio track such that the point of consonant release at the beginning of the vowel for a new auditory token matched the point of release for the original token, at the resolution of a single video frame, for a total of 12 trials. Mismatched stimuli were always a visual /ga/ token paired with an auditory /ma/. Matched stimuli replaced the audio from tokens of the same CV (e.g., a /ma/ visual token paired with a different auditory /ma/), for a total of 16 trials.

For the speech in noise and the AV match–mismatch conditions, participants were instructed to watch and listen to the video display. They were then told that they would hear a man saying some sounds that were not words and to say out loud what they heard.

B. Visual tracking methodology

Visual tracking was assessed with an ASL model 504 pan/tilt remote tracking system (Applied Science Laboratories, Bedford MA), a remote video-based single eye tracker that uses bright pupil, coaxial illumination to track both pupil and corneal reflections at 120 Hz. To optimize the accuracy of the pupil coordinates obtained by the optical camera, this model has a magnetic head tracking unit that tracks the position of a small magnetic sensor attached to the head of the participant.

1. Procedure

After parental consent and child assent (children) or individual consent (adults) was obtained in accordance with the Yale University Institutional Review Board, all participants completed the experimental tasks in the eye-tracker. The participant was placed 60 in. in front of the monitor and eye-tracker, after which calibration of the participant's fixation points in the eye-tracker was completed. Prior to any

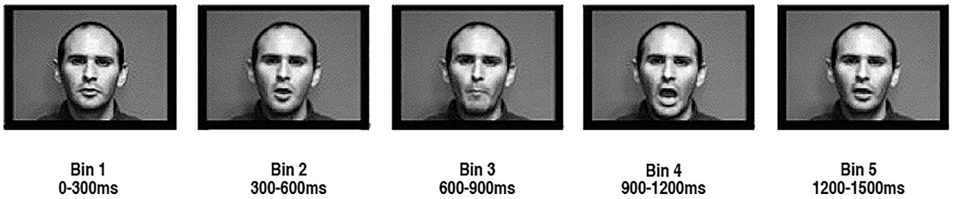

FIG. 1. Image of video frames corresponding to time bins for /ma/.

stimulus presentation for each task, directions appeared on the monitor. These directions were read aloud to the participant by a researcher to ensure that the participant understood the task. In addition, two practice items for each condition were completed with the researcher present to confirm that the participant understood and could complete the task.

After every five trials, participants were presented with a slide of animated shapes and faces to maintain attention to the task. Tasks were blocked, with stimuli presented in random order within block. The inter-stimulus interval for all trials within the blocks was 3 s. The blocks were presented in a semi-random order. The auditory-only stimuli had to be presented first to be sure that the participants could discriminate the difference between /ma/ and /na/. Other than the first, auditory-only block, the rest of the blocks were presented in random order. All audio stimuli were presented at a comfortable listening level (60 dB) from a centrally located speaker under the eye-tracker, and visual stimuli were presented at a 640 × 480 aspect ratio on a centrally located video monitor.

IV. RESULTS

A. Patterns of gaze

Gaze was analyzed for each of the four age groups at five time bins that corresponded to significant events in the speech signal. The first bin (0–300 ms) is initial neutral rest position, the second bin (300–600 ms) is opening prior to consonant closing gesture, the third bin (600–900 ms) is closure for the /m/ or /n/, the fourth bin (900–1200 ms) includes the peak mouth opening for the vowel, and the last bin (1200–1500 ms) is return to rest at the end of the vowel (see Fig. 1).

There were two measures which were analyzed separately for the speechreading, speech in noise, and AV (McGurk) conditions: *face*, which was the percentage of time fixating on the face out of time gazing on the screen, and *mouth*, which was the percentage of time fixating on the mouth out of time gazing on the face (see Figs. 2, 3, and 4). For percentage of time on the *face*, all participants increased gaze to the face of the speaker once speech movement begins (see onset at 300 ms) until a plateau or slight decrease at 1200–1500 ms, where the speech signal is concluding. Critically, there was a significant main effect of age in all three tasks [speechreading: F (3, 70) = 5.68, $p < 0.01$; speech in noise: F (3, 70) = 7.25, $p < 0.001$; AV: F (3, 70) = 9.55, $p < 0.0001$]. In the three tasks, there was a developmental trend such that the younger children (5–6 and 7–8 yr olds) spent proportionally less time fixating on the face than older children (9–10 yr olds and adults). The difference between adults and 9–10 yr olds (higher mean fixations on the face for adults) was reliable only in the AV condition ($p < 0.01$). There were also significant age × time interactions in all three tasks, F (12, 280) = 4.36, $p < 0.0001$ for speech in noise; F (12, 280) = 5.59, $p < 0.0001$ for speechreading; F (12, 280) = 2.75, $p < 0.001$ for AV (McGurk) conditions, reflecting changes in the overall magnitude of the group differences as the stimulus was presented, but this did not alter the basic pattern of age-related differences.

For percentage of time on the *mouth* (Figs. 5, 6, and 7), there was again a significant main effect of age for speech reading F (3, 70) = 2.70, $p < 0.05$ and speech in noise, F (3, 70) = 5.61, $p < 0.001$, but not for AV, F (3, 70) = 1.3, ns. All three tasks had significant interactions of age × time (speech reading: F (12, 280) = 4.19, $p < 0.0001$; speech in noise: F (12, 280) = 5.43, $p < 0.0001$; AV: F (12, 280) = 2.75, $p < 0.001$). In all three tasks, the younger children (5–6 yr olds) spent less time fixating on the mouth region than the older children (ages 7–8 and 9–10 yrs old), who did not reliably differ. The interaction reflected the finding that the age difference emerged only when speech movement began (300 ms and onward); the older children exhibited a sharper increase in mouth fixations from the first bin to the third, with a more gradual increase for the younger children. Unlike in fixations on the face, adults did not exhibit a higher proportion of fixations on the mouth than the children; adults exhibited a relatively flat pattern of mouth fixations across the time bins; thus, by mid-stimulus (600–900 ms) they tended to have more mouth fixations than the youngest children (5–6 yrs old) but fewer than the older children (7–8 and 9–10 yrs old).

We also examined gaze to the eyes, nose, and to nonfocal areas of the face (areas other than the eyes, nose, and

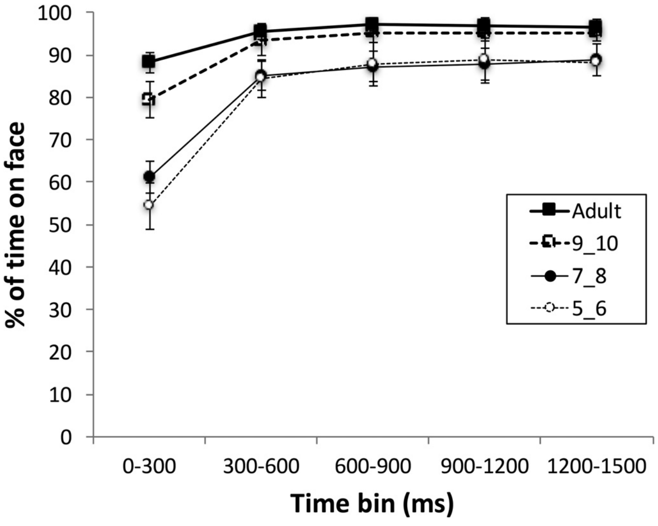

FIG. 2. Percent of time on face: speechreading condition.

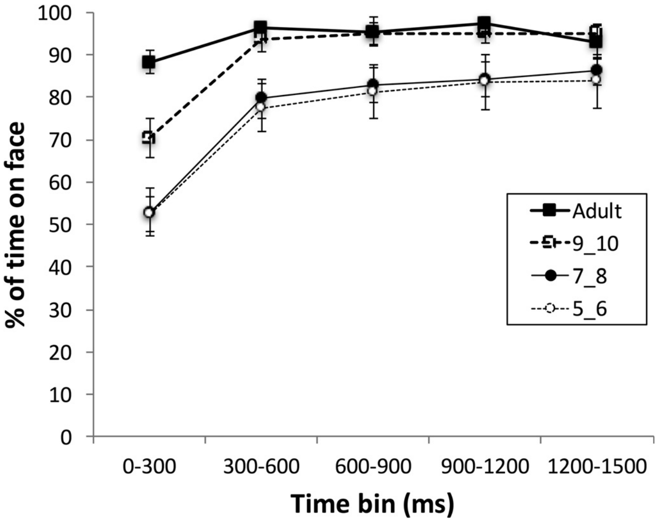

FIG. 3. Percent of time on face: speech in noise condition.

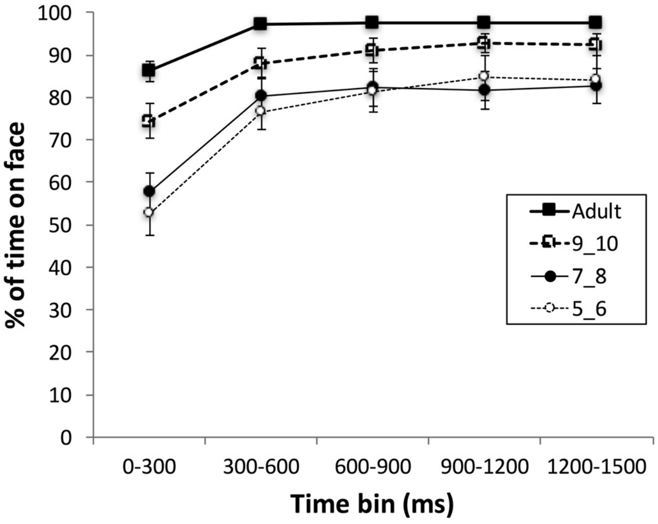

FIG. 4. Percent of time on face: AV (McGurk) condition.

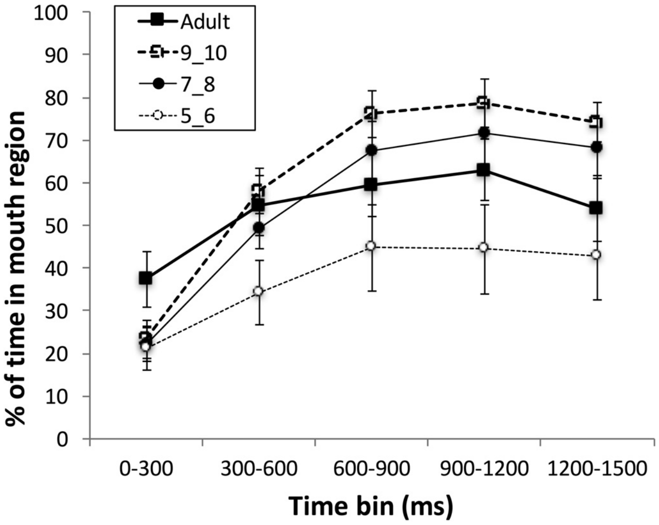

FIG. 5. Percent of time on mouth out of time on face: speechreading.

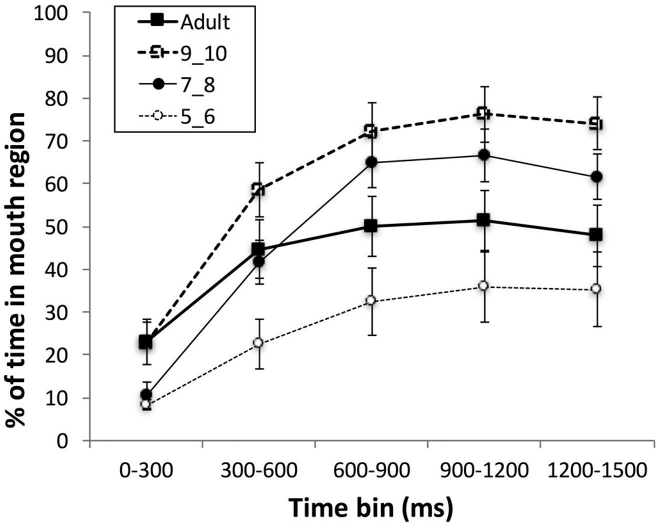

FIG. 6. Percent of time on mouth out of time on face: speech in noise condition.

mouth). There were no significant effects between groups for these areas.[1]

V. DISCUSSION

Developmental trends in visual influence on heard speech have been reported in the literature, with a range of possible explanations for this effect, including motor experience, ability to pick up phonetic information from the visual signal and attention. The current study investigated whether there are developmental changes in gaze to relevant areas of the speaking face, a necessary precursor to determining whether such changes could underlie changes in visual influence.

Our data showed an increase in gaze on the face, specifically in fixations on the *mouth* of a speaker, between the ages of 5 and 10 for speech reading, AV speech in auditory noise, and mismatched AV (McGurk) speech. These results reveal a potential explanatory factor for previously reported

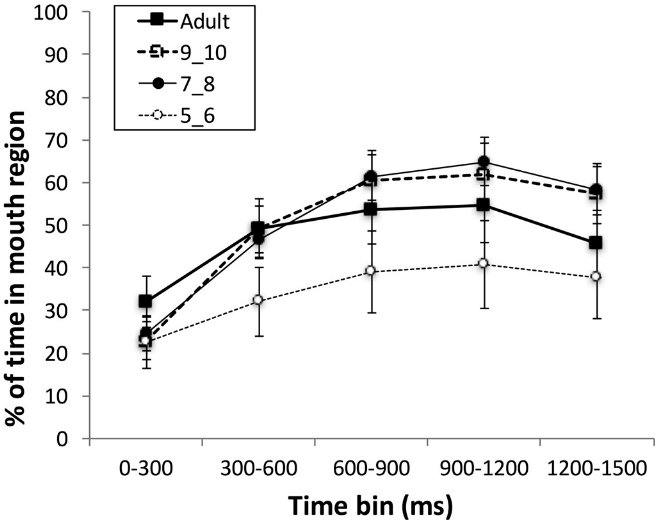

FIG. 7. Percent of time on mouth out of time on face: AV (McGurk) condition.

findings of increased visual influence with development: an increase in fixation (and possibly visual attention) to the mouth of the speaker. The current data suggest that while adults show a high percentage of time on face relative to children, they do not show a great deal of time fixating on the mouth. Interestingly, previous research by Vatikiotis-Bateson *et al.* (1998) and Paré *et al.* (2003) demonstrates that direct gaze on the mouth of the speaker is not required for influence of visual articulatory information in typically developing adults, however, this may not be the case for younger listeners. In particular, the youngest listeners (5–6 yr olds) appear to be less focused on the speaker's mouth, indicating either general poor attention to the mouth in the youngest listeners, or, potentially a developmental trend in focus to the speaker's articulators. Because previous research with adults has shown less visual influence on heard speech when visual attentional demands are increased (Alsius *et al.*, 2005), it is possible that greater focus on the mouth exhibited with development reduces attention load during AV perception for children, which may facilitate more effective pickup of visual phonetic information with age. Either of these possibilities might provide some understanding of the relationship between gaze to the speaking face in developmental disability as well. For example, children with an ASD could be less generally attentive to the mouth of the speaker (also see Irwin and Brancazio, 2014) or may be exhibiting a more immature developmental pattern, which can be assessed by looking at the pattern of gaze in adolescents and young adults on the autism spectrum.

Although the previous literature indicates that young children differ from older children and adults in visual influence on heard speech (e.g., Hockley and Polka, 1994; Massaro, 1984; Massaro *et al.*, 1986; McGurk and MacDonald, 1976; Sekiyama and Burnham, 2008), our behavioral data did not allow us to evaluate this finding in the current sample, primarily due to overall high performance in behavioral responding. An additional factor to consider is that the current stimuli were CV /ma/ and /na/'s. Connected speech, such as sentence-level stimuli, would be more akin to what many children and adults encounter in a communicative exchange. Future research should include somewhat more difficult visual tasks that would elicit a more variable pattern of results across participants, potentially with connected speech stimuli. Such work could determine whether age-related changes in looking to the face are linked with changes in AV perception, and in particular, whether this is due to a pickup of phonetic information or processing load.

ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. DC007339, DC011342, DC013864, and the LEARN Center at Haskins Laboratories. Thanks to Jon Close-Reilly for assistance with data coding and reduction and Amanda Sciola for help with editing.

[1] We attempted to compare the age groups on their behavioral responses on the visual tasks. However, the comparisons were not informative due to overall high performance on the three conditions (visual only median: 100% correct; AV speech in noise median: 100% correct; McGurk median: 92% visually influenced responses). All but four participants had at least 90% accuracy in visual only and all but three had at least 90% accuracy in AV speech in noise.

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Current Biol. **15**, 839–843.

Barenholtz, E., Mavica, L., and Lewkowicz, D. J. (**2016**). "Language familiarity modulates relative attention to the eyes and mouth of a talker," Cognition **147**, 100–105.

Buchan, J. N., Paré, M., and Munhall, K. G. (**2008**). "The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception," Brain Res. **1242**, 162–171.

Burnham, D., and Dodd, B. (**1998**). "Familiarity and novelty preferences in infants' auditory visual speech perception: Problems, factors, and a solution," Adv. Infancy Res. **12**, 170–187.

Desjardins, R. N., Rogers, J., and Werker, J. F. (**1997**). "An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks," J. Exp. Child Psychol. **66**, 85–110.

Hockley, N., and Polka, L. (**1994**). "A developmental study of audiovisual speech perception using the McGurk paradigm," J. Acoust. Soc. Am. **96**, 3309.

Irwin, J., and Brancazio, L. (**2014**). "Seeing to hear: Patterns of gaze to speaking faces in children with autism spectrum disorders," Front. Lang. Sci. **5**, 397.

Johnels, J. A., Gillberg, C., Falck-Ytter, T., and Miniscalco, C. (**2014**). "Face viewing patterns in young children with autism spectrum disorders: Speaking up for a role of language comprehension," J. Speech, Lang. Hear. Res. **57**, 2246–2252.

Lachs, L., and Pisoni, D. B. (**2004**). "Crossmodal source identification in speech perception," Ecol. Psychol. **16**(30), 159–187.

Lalonde, K., and Frush Holt, R. (**2014**). "Audiovisual speech integration development at varying levels of perceptual processing," J. Acoust. Soc. Am. **136**(4), 2263–2263.

Lansing, C. R., and McConkie, G. W. (**2003**). "Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences," Percept. Psychophys. **65**(4), 536–552.

Legerstee, M. (**1990**). "Infants use multimodal information to imitate speech sounds," Infant Behav. Develop. **13**(3), 343–354.

Lewkowicz, D. J., and Hansen-Tift, A. M. (**2012**). "Infants deploy selective attention to the mouth of a talking face when learning speech," Proc. Natl. Acad. Sci. U.S.A. **109**(5), 1431–1436.

MacDonald, J., Andersen, S., and Bachmann, T. (**2000**). "Hearing by eye: How much spatial degradation can be tolerated?," Perception **29**, 1155–1168.

MacDonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception processes," Percept. Psychophys. **24**(3), 253–257.

Massaro, D. W. (**1984**). "Children's perception of visual and auditory speech," Child Develop. **55**, 1777–1788.

Massaro, D. W., Thompson, L. A., Barron, B., and Laren, E. (**1986**). "Developmental changes in visual and auditory contributions to speech perception," J. Exp. Child Psychol. **41**, 93–113.

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Meltzoff, A. N., and Kuhl, P. K. (**1994**). "Faces and speech: Intermodal processing of biologically relevant signals in infants and adults," in *The Development of Intersensory Perception: Comparative Perspectives*, edited by D. J. Lewkowicz and R. Lickliter (Psychology Press, New York), pp. 335–398.

Ménard, L., Dupont, S., Baum, S. R., and Aubin, J. (**2009**). "Production and perception of French vowels by congenitally blind adults and sighted adults," J. Acoust. Soc. Am. **126**(3), 1406–1414.

Paré, M., Richler, R., ten Hove, M., and Munhall, K. G. (**2003**). "Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect," Percept. Psychophys. **65**, 553–567.

Reisberg, D., McLean, J., and Goldfield, A. (**1987**). "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," Language 97–114.

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (**1997**). "The McGurk effect in infants," Percept. Psychophys. **59**(3), 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," European J. Neurosci. **33**, 2329–2337.

Sekiyama, K., and Burnham, D. (**2008**). "Impact of language on development of auditory-visual speech perception," Develop. Sci. **11**(2), 306–320.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.

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**(8), e742.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., and Munhall, K. G. (**1998**). "Eye movement of perceivers during audiovisual speech perception," Percept. Psychophys. **60**, 926–940.

Yi, A., Wong, W., and Eizenmann, M. (**2013**). "Gaze patterns and audiovisual speech enhancement," J. Speech, Lang., Hear. Res. **56**, 471–480.