RESEARCH ARTICLE

"Look who's talking!" Gaze Patterns for Implicit and Explicit Audio-Visual Speech Synchrony Detection in Children With High-Functioning Autism

Ruth B. Grossman, Erin Steinhart, Teresa Mitchell, and William McIlvane

Conversation requires integration of information from faces and voices to fully understand the speaker's message. To detect auditory-visual asynchrony of speech, listeners must integrate visual movements of the face, particularly the mouth, with auditory speech information. Individuals with autism spectrum disorder may be less successful at such multisensory integration, despite their demonstrated preference for looking at the mouth region of a speaker. We showed participants (individuals with and without high-functioning autism (HFA) aged 8–19) a split-screen video of two identical individuals speaking side by side. Only one of the speakers was in synchrony with the corresponding audio track and synchrony switched between the two speakers every few seconds. Participants were asked to watch the video without further instructions (implicit condition) or to specifically watch the in-synch speaker (explicit condition). We recorded which part of the screen and face their eyes targeted. Both groups looked at the in-synch video significantly more with explicit instructions. However, participants with HFA looked at the in-synch video less than typically developing (TD) peers and did not increase their gaze time as much as TD participants in the explicit task. Importantly, the HFA group looked significantly less at the mouth than their TD peers, and significantly more at non-face regions of the image. There were no between-group differences for eye-directed gaze. Overall, individuals with HFA spend less time looking at the crucially important mouth region of the face during auditory-visual speech integration, which is maladaptive gaze behavior for this type of task. Autism Res 2015. © 2015 International Society for Autism Research, Wiley Periodicals, Inc.

Keywords: face perception; audio-visual integration; high-functioning autism; eye tracking; mouth-directed gaze

Introduction

During daily communication, spoken words are combined with facial and manual gestures to allow for more complete understanding of our messages. Visual speech information (e.g., lip movements) is rapidly integrated with auditory signals and supports comprehension, particularly in noisy environments [Sumby & Pollack, 1954]. Individuals with autism spectrum disorder (ASD) have demonstrated difficulty integrating information from different sensory modalities (see [Iarocci & McDonald, 2006b] for review), although the evidence is not clear cut.

There are data showing that low-level integration of simple, nonverbal visual and auditory stimuli, such as beeps and flashes are preserved in ASD [van der Smagt, van Engeland, & Kemner, 2007; Zainal, Magiati, Tan, Sung, Fung, & Howlin, 2014]. Conversely, studies have demonstrated reduced robustness in this population of the McGurk effect [McGurk & MacDonald, 1976], in which auditory /ga/ is blended with visual speech /ba/ to be perceived as /da/ [de Gelder, Vroomen, & van der Heide, 1991; Hampson, van Anders, & Mullin, 2006; Irwin, 2006; Magnée, de Gelder, van Engeland, & Kemner, 2008; Zainal et al., 2014]. This task is often used to determine how people integrate speech information from voices and faces, as necessary in every-day speech. Children with ASD tend to give perceptual priority to auditory over visual speech information in this type of task. However, this deficit is not always present [Nishiyama & Kanne, 2014] and may be driven mostly by reduced lipreading skills [Medeiros & Winsler, 2014]. This lipreading deficit may result in greater reliance on auditory over visual aspects of audio-visual (AV) speech [Iarocci, Rombough, Yager, Weeks, & Chua, 2010] and explain why individuals with high-functioning autism (HFA) have reduced attention to visual speech [de Gelder et al., 1991] and do not derive as much comprehension benefit from lip movements as typically developing (TD) peers in a speech-in-noise paradigm [Smith

From the Emerson College, Department of Communication Sciences and Disorders, 120 Boylston Street, Boston, Massachusetts (R.B.G.); University of Massachusetts Medical School Shriver Center, 200 Trapelo Rd, Waltham, Massachusetts (R.B.G., E.S., T.M., W.M.)

Received October 28, 2013; accepted for publication November 25, 2014

Address for correspondence and reprints: Ruth B. Grossman, Emerson College, 120 Boylston Street, Boston, MA 02116. E-mail: Ruth_grossman@emerson.edu

Published online 00 Month 2014 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.1447

& Bennetto, 2007], which may have negative impact on comprehension during social communication.

Receptive speech tasks in which the auditory and visual channels are out of temporal synchrony place high demands on AV speech integration. The visual system (high spatial acuity) and the auditory system (high temporal acuity) both contribute significantly to resolving conflicting intermodal timing information and individuals with ASD may have difficulty with this type of multi-modal speech integration [Iarocci & McDonald, 2006b; Iarocci et al., 2010; Shams, Kamitani, & Shimojo, 2004]. However, the evidence is not without contradiction. Some studies show reduced ability to detect AV asynchrony for speech and non-speech (flashes and beeps) stimuli with short (40–320 ms) temporal offsets [Piven & Sasson, 2014] or large (3,000 ms) temporal offset [Ulloa & Pineda, 2007]. Other data show that children with HFA are as capable as their TD peers at detecting AV asynchrony with a range of short temporal offsets [Grossman, Schneps, & Tager-Flusberg, 2009].

The lack of consensus across studies may stem from differences in experimental paradigms and participant selection. Grossman et al. [2009] studied adolescents, presenting phrase-level speech with audio lagging behind video by 4–12 (120–400 ms) frames. Participants with HFA detected asynchrony as accurately as TD peers for all audio-lag rates. By contrast, Bebko et al. [2006] used significantly longer temporal offsets (3 sec) in a preferential looking paradigm with preschool-aged children. Here, the ASD group—in contrast to the TD group—exhibited no preference for looking at synchronous linguistic stimuli, but preserved preferential looking toward synchronous nonlinguistic stimuli. de Boer-Schellekens et al. [2013] found that adolescents with ASD were less sensitive to very small differences in AV synchrony of nonlinguistic and linguistic (syllables) stimuli than TD peers. Overall, individuals with ASD seem to have some ability to integrate auditory and visual information, but may attend to each of these channels differently across stimuli and task demands. The relative attentional allocation on visual speech can be better understood through analysis of gaze patterns to speaking faces, specifically the degree of visual fixation to the mouth region, which contains most relevant information for integration of auditory-visual speech signals.

Eyetracking studies of individuals with HFA have documented a preference to gaze at the mouth and avoid the eye region of a face [Jones, Carr, & Klin, 2008; Joseph & Tanaka, 2003; Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Langdell, 1978; Neumann, Spezio, Piven, & Adolphs, 2006; Paul, Campbell, Gilbert, & Tsiouri, 2013; Pelphrey, Sasson, Reznick, Paul, Goldman, & Piven, 2002; Spezio, Adolphs, Hurley, & Piven, 2007], although recently, several studies have not shown this effect [Bar-Haim, Shulman, Lamy, & Reuveni, 2006; Nishimura, Rutherford, & Maurer, 2008; Rutherford & McIntosh, 2007; Rutherford & Towns, 2008]. However, these findings are not based on auditory-visual speech integration tasks. Eyetracking evidence for visual speech processing shows that TD individuals prefer gazing at central face areas in tasks requiring greater reliance on visual speech, such as scenarios with low-intensity auditory signals [Buchan, Pare, & Munhall, 2007] or specifically the mouth [Lansing & McConkie, 2003; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998]. TD participants derive significant comprehension benefit from mouth movements [Sumby & Pollack, 1954] and show reduced AV integration when relying on peripheral face regions [Paré, Richler, Hove, & Munhall, 2003]. If individuals with ASD have an a priori preference for gazing at the mouth, this should provide them with an advantage in speech reading tasks, which has not been noted in the literature. Conversely, a deficit in AV integration might mitigate a potential mouth-gaze advantage.

Task design and response demands may also influence performance on AV speech integration. Summerfield and McGrath [1984] showed that TD individuals are susceptible to instruction bias when attending to AV information. When asked to report what they heard, participants showed greater reliance on the auditory component, than when the instructions were phrased more neutrally. Iarocci et al. [2010] emphasize the importance of taking this potential bias into account when investigating intermodal sensory integration in children with ASD. The type of behavioral response required (e.g. verbal or manual button presses) may also influence the performance [Nuske, Vivanti, & Dissanayake, 2014]. To determine the ability of individuals with ASD to integrate AV speech information, studies should use ecologically valid stimuli, such as continuous speech, in tasks involving instructions and response types that impose no sensory bias. Eyetracking analysis will allow us to determine whether reduced ability to integrate bimodal speech in this population is based on reduced gaze to relevant face areas or a separate sensory integration deficit.

We created an AV integration task for continuous, natural speech without required task response. To investigate the effects of task instructions, we presented implicit (no instructions given other than "look and listen") and explicit (look at the person who's talking) conditions. We used eyetracking to determine whether individuals with HFA use mouth-directed gaze to process AV asynchrony and whether gaze patterns to the crucially important mouth region change with explicit task instructions. We hypothesize that participants with HFA will: (1) gaze at the in-synch speaker in the implicit task less than their TD peers; (2) increase gaze

to the in-synch face with explicit instructions, but less so than TD participants; (3) show no overall increased gaze to the mouth, and (4) show decreased gaze to the eyes compared to TD peers.

Methods

We enrolled children and adolescents with HFA ($N = 30$) and TD controls ($N = 30$) aged 8–19 years, matched on age, sex, IQ, and receptive vocabulary skills. All participants passed vision, color vision, and hearing screenings. Participants were recruited through local schools, advertisements in magazines, newspapers, the internet, autism advocacy groups, and word of mouth. All descriptive characteristics are in Table 1. Informed consent was obtained under a protocol of the University of Massachusetts Medical School Institutional Review Board.

Table 1. Descriptive Characteristics of Participant groups

	HFA ($n = 30$) M(SD)	TD ($n = 30$) M(SD)
Age	11:10(1:4)	12:5(0:11)
	Range: 8:5–19:0	Range: 8:6–17:11
Sex	28 male	25 male
	2 female	5 female
IQ	104(15.9)	109(11.2)
	Range: 80–137	Range: 82–128
PPVT-III	107.8(20.5)	113.9(11.9)
	Range: 80–154	Range: 90–135

Diagnosis of HFA

Participants with ASD met DSM-IV criteria for autistic disorder, based on expert clinical impression and confirmed by the Autism Diagnostic Observation Schedule—Module 3 (ADOS, [Lord, Rutter, DiLavore, & Risi, 1999]) administered by experienced examiners. Participants with known genetic disorders were excluded to reduce heterogeneity of the cohort. Based on ADOS algorithm scores, 15 participants met criteria for autism and 15 met criteria for ASD. We also conducted standardized IQ [Leiter International Performance Scale Revised (Leiter-R, [Roid & Miller, 1997])] and receptive vocabulary (Peabody Picture Vocabulary Test [PPVT-III; Dunn & Dunn, 1997]) tests to verify that participants in the ASD group also had language and cognitive skills within normal limits, allowing us to describe them as having HFA. Using multivariate ANOVA with group as the independent variable we verified that the HFA and TD groups did not differ significantly in age, F (1, 59) = 0.84, $P = 0.36$, IQ, F (1,59) = 1.93, $P = 0.17$, or receptive vocabulary ability, F (1,59) = 1.98, $P = 0.16$. A chi-squared analysis showed that the groups did not differ in distribution of gender (χ^2 (1,$N = 60$) = 1.46, $P = 0.42$).

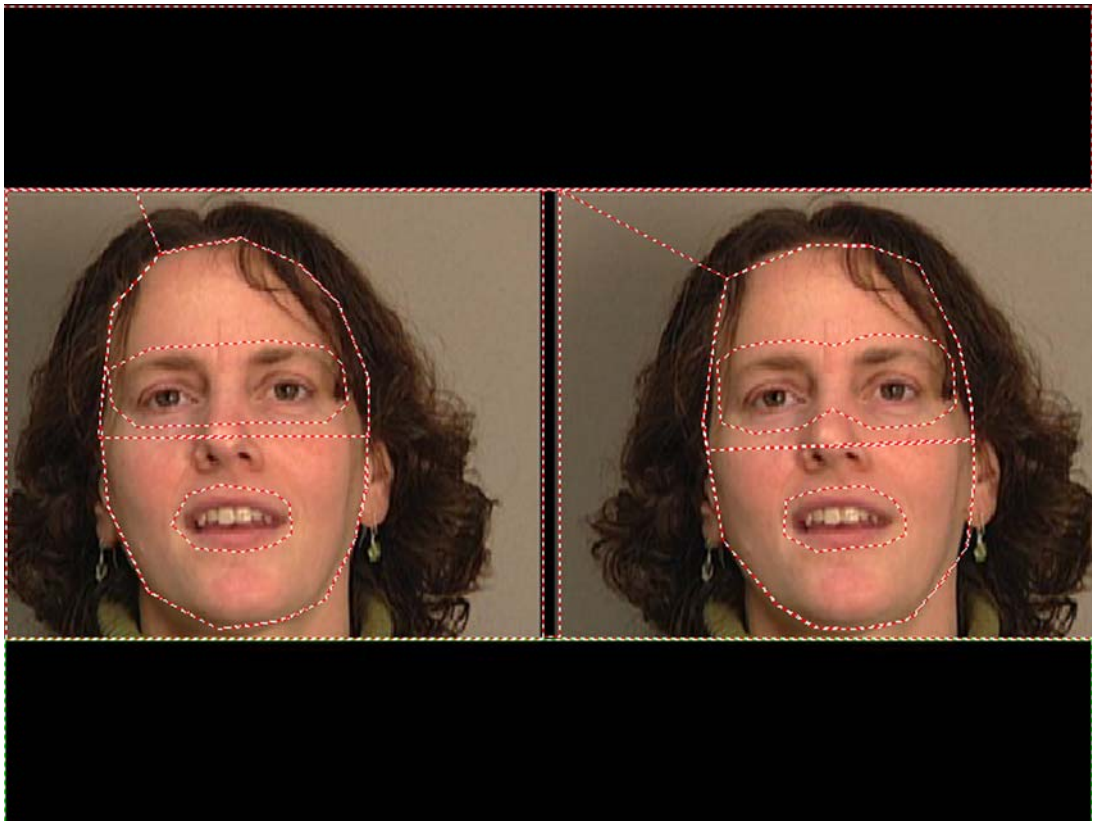
Stimuli

The video showed a woman's head and neck against a neutral background, speaking in simple, clear language, using high-frequency vocabulary, and sentence structure [Grossman et al., 2009]. We presented the same video in side-by-side frames on a computer monitor, with one of the two videos lagging behind the other by 10 frames, or 330 ms. We chose to delay the audio, rather than the video, because an audio delay was found to produce more reliable detection levels [Grossman et al., 2009] and not to result in age-related differences that could have been a confound [Kozlowski & Cutting, 1977]. The 330-ms delay was chosen, because it is significantly longer than the temporal binding windows for low-level bimodal stimuli, such as flashes and beeps [Hall, Szechtman, & Nahmias, 2003] and syllables [Nuske et al., 2014] in this population (<184 ms). We are, therefore, confident that eye-gaze patterns recorded in our task were not affected by low-threshold differences in temporal binding. In addition, our prior study showed that cohorts with and without HFA could detect onset asynchrony above chance (>63%) at this audio delay [Grossman et al., 2009], thereby making this task difficult enough to maintain attention and avoid ceiling-level performance while also enabling participants to detect the synchronous speaker at above-chance levels.

We created two versions of the stimuli, one showing the left video lagging behind the right and the other showing the right video lagging behind the left. We alternated presentation of the two versions in the implicit condition across participants and showed the alternate version in the explicit condition. A single audio track accompanied the two videos and switched from being in synch with the left or right video every 8–18 sec with a pseudo-random distribution of longer and shorter intervals. The audio was always in synch with one of the two videos and lagged behind the other by ten frames. As the audio-switch edit points contained visually noticeable "blips," we inserted additional edit points—or "blips"—into the video every 8 or 9 sec, not all of which were followed by an audio synchrony switch. This served to disguise audio switching points and effectively eliminated potential cueing of participants. The complete stimulus video was 4 min and 37 sec long and shown using Presentation software (Neurobehavioral Systems, http://www.neurobs.com).

Procedure

Participants sat in front of a monitor connected to an ISCAN RK-826 PCI Pupil/Corneal Reflection 60 Hz Tracking System (ISCAN Corp., Woburn, MA) at a comfortable viewing distance. The infrared light source and



COLOR

Figure 1. Screenshot of video stimuli with regions of interest.

eye camera were mounted below the monitor, affording participants free range of head movements. All participants successfully completed a calibration task guiding them to look at the four corners and center of the screen.

In the first (implicit) condition, participants were introduced to the task and provided minimal instructions ("look and listen" to the video of a woman talking about baking dessert), so we could record unbiased gaze behavior. No overt response was required. After completion of the implicit condition, we introduced an unrelated distractor task, followed by presentation of the second (explicit) condition. We told participants they would see the same video as before, but that the audio would be in synch with only one of the two faces at a time and that synchrony would switch back and forth between the two sides at random intervals. Their instructions were to "listen carefully and look only at the person speaking." Again, no response, other than eye gaze was required, thereby eliminating sensory bias in response type between the two conditions.

Results

We recorded *x*- and *y*-coordinates of the point of regard for every time sample (60 Hz). A fixation was defined as point of regard maintaining maximum horizontal deviation of five pixels and maximum vertical deviation of three pixels for a minimum of 40 ms. We analyzed three eyetracking variables: (1) looking time to a specified region of interest (ROI, described below) as a percentage of overall looking time to the screen; (2) number of fixations to an ROI; and (3) time to first fixation (in seconds) to an ROI after each synchrony side change.

We defined six ROIs: eyes, mouth, upper face, lower face, non-face, and video (see Fig. 1). The eye ROI F1
encompassed the upper and lower lids, and eyebrows. The mouth ROI encompassed the lips and surrounding areas involved in speech. The upper face ROI was defined by a horizontal line across the tip of the nose as the lower border and encompassed the area up to the hairline and ears (including the eyes). The lower face ROI was defined with the same horizontal line as the upper border and encompassed the area down to and including the chin and both cheeks (including the mouth). The non-face ROI included the hair, neck, and background of each image. The video ROI encompassed the entire video of each speaker, providing an overall measure of whether participants were looking at the in-synch video. Initial analysis showed that upper face and lower face ROI results were driven almost exclusively by data from eye and mouth ROIs, so data

Table 2. Percent Looking Time

Task version	ROI	HFA ($N = 30$) Mean (StDev)	TD ($N = 30$) Mean (StDev)
Implicit	Eyes	7.53 (8.55)	7.57 (7.77)
	Mouth	3.1 (2.96)	8.53 (8.98)
	Side	31.25 (12.04)	44.7 (20.37)
	Non-face	4.73 (3.29)	2.92 (2.9)
	Upper face	11.62 (11.28)	13.58 (13.23)
	Lower face	14.9 (8.51)	28.2 (21.56)
Explicit	Eyes	6.1 (5.45)	8.25 (8.34)
	Mouth	3.27 (5.02)	11.3 (11.16)
	Side	35.23 (15.86)	56.68 (22.96)
	Non-face	6.25 (3.03)	3.95 (4.47)
	Upper face	11.82 (9.56)	14.5 (12.47)
	Lower face	17.17 (15.17)	38.23 (24.37)

Table 3. Number of Fixations

Task version	ROI	HFA ($N = 30$) Mean (StDev)	TD ($N = 30$) Mean (StDev)
Implicit	Eyes	142.07 (121.77)	150.53 (113.87)
	Mouth	62.03 (59.32)	99.43 (77.28)
	Side	702.43 (364.59)	729.37 (360.47)
	Non-face	145.5 (85.53)	99.03 (93.85)
	Upper face	247.63 (231.54)	260.87 (198.89)
	Lower face	309.3 (222.57)	369.47 (199.86)
Explicit	Eyes	114.1 (121.05)	134.93 (108.69)
	Mouth	57.87 (53.27	92.57 (88.66)
	Side	773.63 (347.43)	727.8 (413.62)
	Non-face	185.03 (85.0)	98.7 (113.71)
	Upper face	271.93 (228.77)	239.77 (180.55)
	Lower face	316.67 (236.37)	389.33 (289.67)

Table 4. Time to First Fixation, in Seconds

Task version	ROI	HFA ($N = 30$) Mean (StDev)	TD ($N = 30$) Mean (StDev)
Implicit	Eyes	3.24 (1.22)	3.24 (1.09)
	Mouth	3.76 (.99)	3.14 (1.3)
	Side	2.92 (.64)	2.76 (.54)
	Non-face	3.26 (1.01)	3.33 (1.24)
	Upper face	2.87 (1.1)	2.95 (1.25)
	Lower face	2.62 (1.02)	1.99 (.82)
Explicit	Eyes	3.41 (.99)	3.29 (.90)
	Mouth	3.19 (1.16)	3.02 (1.49)
	Side	2.84 (.51)	2.83 (.58)
	Non-face	3.02 (.65)	3.81 (1.43)
	Upper face	3.11 (1.24)	2.71 (.79)
	Lower face	2.41 (.87)	1.97 (.92)

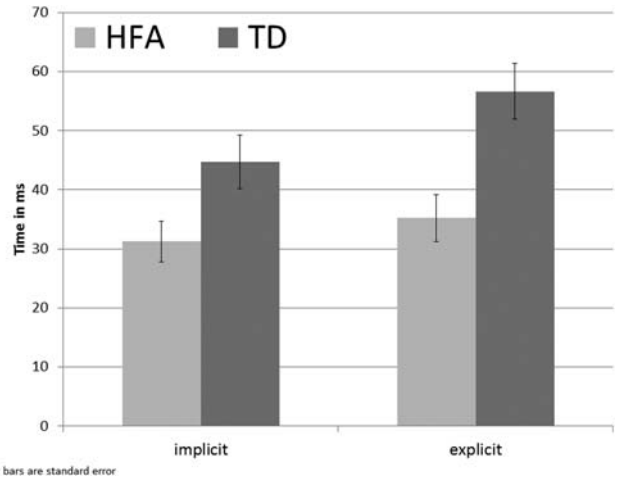


Figure 2. Percent looking time to in-synch video.

presentation will focus on eyes, mouth, non-face, and video ROIs. Results for all six ROIs are in Tables (2–4).

To investigate whether participants could perform the task, we conducted within-group paired *t*-tests, which revealed that both groups gazed at the in-synch video significantly longer than the out-of-synch video in the implicit condition (HFA: $t\,(1, 29) = 5.33$, $P < 0.0001$, TD: $t\,(1, 29) = 4.48$, $P < 0.0001$) and the explicit condition (HFA: $t\,(1, 29) = 5.31$, $P < 0.0001$, TD: $t\,(1, 29) = 7.7$, $P < 0.0001$). Based on these results we restricted further analyses of gaze patterns to only the in-synch videos of both conditions. Given the relatively large age range of participants, we explored correlations of our measures with age and found no significant correlations with looking patterns to any of the in-synch video ROIs.

Our subsequent analyses focused on looking time (expressed as percent of looking to a given ROI relative to total looking time to the screen), number of fixations to an ROI, and time to the first fixation to the ROI.

Gaze Patterns to the video ROI

We conducted two (group) by two (condition: implicit, explicit) repeated measures ANOVA for percent looking time to the correct video. We found main effects for group ($F\ (1,58) = 16.59$, $P < 0.0001$, partial $\eta^2 = 0.22$) and condition ($F\ (1,58) = 16.13$, $P < 0.0001$, partial $\eta^2 = 0.22$). Both participant groups gazed at the correct side significantly longer in the explicit than the implicit condition ($P < 0.0001$), indicating they both modulated gaze behavior based on task instructions (Fig. 2). However, the TD group spent a greater percentage of their looking time gazing at the correct side than their HFA peers, regardless of condition. There was also a significant group by condition interaction ($F\ (1,58) = 4.1$, $P = 0.05$, partial $\eta^2 = 0.065$), showing that the increase in gaze to the in-synch video during the explicit condition was greater in the TD group than the HFA group. Follow-up within-group *t*-tests (FDR corrected) showed that the TD group gazed at the in-synch video significantly longer in the explicit than implicit condition ($t\ (1, 29) = 3.8$, $q = 0.008$), but the HFA group did not ($t\ (1, 29) = 1.65$, $q < 0.37$). There were no significant between or within group differences for number of fixations or time to first fixation to this ROI.

Gaze Patterns to Smaller ROIs

We conducted a two (group) by two (condition: implicit, explicit) by three (ROI: eyes, mouth, non-face) repeated measures ANOVA of percent looking time showing main effects for group (F (1,58) = 10.1, $P = 0.002$, partial $\eta^2 = 0.15$), condition (F (1,58) = 4.62, $P = 0.036$, partial $\eta^2 = 0.08$), and ROI (F (2,116) = 3.38, $P = 0.037$, partial $\eta^2 = 0.06$). We also identified a significant interaction between ROI and group (F (2,116) = 7.49, $P = 0.001$, partial $\eta^2 = 0.11$). A two (group) by two (condition: implicit, explicit) by three (ROI: eyes, mouth, non-face) repeated measures ANOVA of time to first fixation revealed no main effects for group, condition, or ROI, but a significant interaction between ROI and group (F (2,116) = 3.52, $P = 0.033$, partial $\eta^2 = 0.06$). A two (group) by two (condition: implicit, explicit) by three (ROI: eyes, mouth, non-face) Repeated measures ANOVA of number of fixations showed no main effects for group or condition, but a significant effect for ROI (F (2,116) = 13.35, $P < 0.0001$, partial $\eta^2 = 0.19$) and a significant ROI by group interaction (F (2,116) = 7.44, $P = 0.001$, partial $\eta^2 = 0.11$).

We then conducted more in-depth analyses using a series of two (group) by two (condition: implicit, explicit) repeated measures ANOVAs of percent looking time, number of fixations, and time to first fixation within and between groups for gaze patterns to the three within-video ROIs (eyes, mouth, nonface).

Gaze patterns to the eye ROI. There were no significant differences between or within groups in percent looking time, number of fixations, or time to first fixation to the eyes.

Gaze patterns to the mouth ROI. There were main effects for group with the TD group looking longer to the mouth ROI (F (1,58) = 14.46, $P < 0.0001$, partial $\eta^2 = 0.2$, Fig. 3) and making more fixations to the mouth than the HFA group (F (1,58) = 5.56, $P = 0.02$, partial $\eta^2 = 0.09$). There was also a main effect for condition (F (1,58) = 4.03, $P = 0.05$, partial $\eta^2 = 0.07$) with both groups making their first fixation to the mouth faster in the explicit vs. implicit condition. No other comparisons or interactions were significant for this ROI.

Gaze Patterns to the Non-Face ROI

A repeated measures ANOVA for percent looking time revealed a main effect for condition (F (1,58) = 4.81, $P = 0.03$, partial $\eta^2 = 0.08$), with both groups looking longer at the non-face during the explicit than implicit condition. As this was the largest within-video ROI, this effect may simply have been a corollary of increased looking time to in-synch video during the explicit task. Separate repeated measures ANOVAs for percent looking time, number of fixations, and time to first fixation revealed main effects for group, with the HFA group looking longer at the non-face than the TD group (F (1,58) = 9.06, $P = 0.004$, partial $\eta^2 = 0.14$), making more fixations in this ROI (F (1,58) = 12.06, $P = 0.01$, partial $\eta^2 = 0.17$), and being faster to fixate on the non-face than the TD group (F (1,58) = 4.62, $P = 0.004$, partial $\eta^2 = 0.07$), regardless of condition. The analyses revealed no significant group by condition interactions for this ROI.

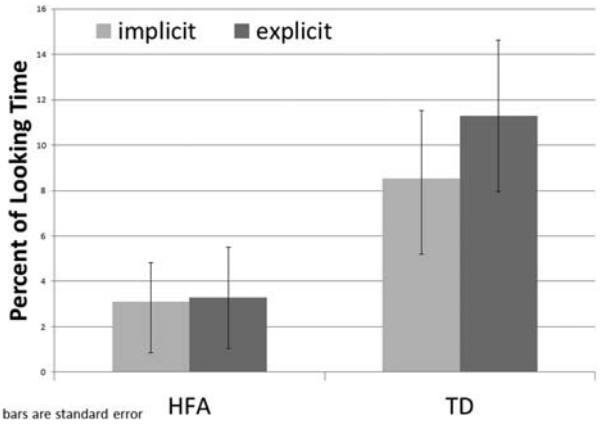


Figure 3. Percent looking time to mouth ROI.

We repeated the same set of between-within group analyses for all ROIs in both task conditions comparing the two HFA groups (autism vs. ASD, as differentiated by ADOS algorithm cut-off scores) with 15 participants per group to determine whether performance differed between these two subgroups. We found no significant differences for any measure or condition and no significant correlations of gaze pattern data with ADOS algorithm scores. This indicates that—as far as this task is concerned—participants with HFA showed a fairly homogeneous gaze pattern profile, regardless of subgroup or characteristics of social communication deficits.

Discussion

This study examined the gaze patterns of children with and without HFA during processing of bimodal (auditory and visual) asynchronous speech. We investigated the effect of task instructions through comparison of implicit vs. explicit conditions. Explicit instructions were designed to contain an equal number of visual (look) and auditory (talking) words, to avoid biasing attention to either sensory modality. Lastly, we eliminated potential response bias by focusing exclusively on eye tracking data.

ID: jwweb3b2server Time: 14:16 I Path: D:/JW/Support/Printer_Autopdf/3D_IN/JW-AUR#140014

Results show both groups looking significantly more to the in-synch than the out-of-synch video, in both conditions, indicating that in-synch speech draws the eye of both participant groups more than the novelty, or strangeness of out-of-synch speech, even in the absence of task instructions. However, confirming our first hypothesis, participants with HFA do not gaze at the in-synch speaker as much as their TD peers in the implicit task. When comparing gaze patterns across task conditions, individuals with HFA do not increase their gaze to the in-synch video as much as the TD cohort after explicit instructions to look at the in-synch speaker. These data support our second hypothesis that the gaze behavior of participants with HFA would be less responsive than TD peers to explicit task instructions. Although there is a main effect showing both groups gazing longer at the in-synch speaker in the explicit vs. implicit condition, Figure 2 shows that this increased gaze time is larger for TD participants than participants with HFA. Participants with HFA do gaze at the crucially important mouth region significantly faster in the explicit vs. the implicit task, indicating that they can and do change gaze behavior in response to explicit task instructions but may be less able to sustain this behavioral change.

Our third hypothesis was that children with HFA would show gaze patterns to the mouth that are similar to those of their TD peers in both task conditions. Contrary to our hypothesis, we detected significantly decreased gaze (looking time and number of fixations) to the mouth for HFA compared to TD participants, regardless of condition. These data are particularly startling given the existing literature on increased mouth-directed gaze in ASD vs. TD individuals [Jones et al., 2008; Joseph & Tanaka, 2003; Klin et al., 2002; Langdell, 1978; Neumann et al., 2006; Paul et al., 2013; Pelphrey et al., 2002; Spezio et al., 2007]. In contrast to prior studies, our task focused on visual speech, not emotion or identity recognition, which may prompt different gaze behavior in both participant cohorts. Nevertheless, participants with HFA neither implicitly nor explicitly modified gaze behavior in a way that would demonstrate understanding that the mouth was the primary source of visual information for AV speech integration.

TD individuals in our study did not gaze at the mouth faster, but did sustain longer visual attention to that region during the explicit condition. These data are well supported by existing literature. Lansing and McConkie [2003] found that TD adults gazed at the eyes of nonspeaking faces, but fixated longer on the mouth of speaking faces, particularly in more difficult comprehension contexts, such as low volume speech. The authors proposed that TD adults gazed at the eyes of resting faces to determine emotional or social information, but diverted their gaze to the mouth when asked to comprehend AV speech with degraded audio signal. Other data suggest that TD individuals obtain sufficient information for AV speech comprehension by fixating on the eyes and noting dynamic information from the lips through non-foveal vision [Vatikiotis-Bateson, Eigsti, & Yano, 1994]. These findings appear to be contradictory, until we take the difficulty of the task into account. Peripheral or non-foveal perception of the mouth may be sufficient in simpler speech recognition tasks [Massaro, 1998; Vatikiotis-Bateson et al., 1998], but increased difficulty of the speech recognition task drives TD observers to increase foveal fixations to the mouth region of the face while at the same time reducing fixation to other regions of the face [Lansing & McConkie, 2003]. Although the audio-quality of the stimuli used here was not degraded, the AV slip rate was difficult enough to reach above-chance, but still low (>63%), accuracy levels. We, therefore, propose that this high level of difficulty redirected gaze patterns of TD participants from the eyes to the mouth in an attempt to improve AV asynchrony detection, particularly in the explicit condition.

Participants with HFA did not follow this adaptive gaze pattern. In contrast, this cohort showed significantly fewer fixations to the mouth region and significantly more and longer fixations to the non-face region, which contains no relevant information for this task. Dynamic information from regions other than the central face or mouth region do not enhance AV speech recognition [Ijsseldijk, 1992; Marassa & Lansing, 1995], thereby making the gaze pattern of participants with HFA a maladaptive strategy for difficult AV speech tasks. One possible explanation for this gaze pattern difference is that individuals with HFA were actively avoiding looking at the mouth—focusing on the non-face area instead—to reduce the demands of multisensory integration, which have frequently been reported to be difficult for this population [Iarocci & McDonald, 2006a]. However, data also show that individuals with ASD have preserved multisensory integration abilities for low-level nonlinguistic stimuli [van der Smagt et al., 2007] and meaningful language contexts [Grossman et al., 2009], despite possible differences in the timing of that integration [Foss-Feig et al., 2010; Kwakye, Foss-Feig, Cascio, Stone, & Wallace, 2011]. These data collectively do not point to a clear deficit in multisensory integration for meaningful linguistic stimuli. In addition, participants with HFA in this study did modulate their gaze pattern in response to explicit instructions by gazing at the mouth faster than in the implicit task, thereby demonstrating a willingness to engage with task demands for multisensory integration. Although this cohort demonstrated the ability to shift gaze quickly to the mouth region in the explicit

condition, they did not sustain foveal gaze to this crucially important region, thereby significantly reducing their chance of successully integrating AV speech information and supporting comprehension.

Our fourth hypothesis stated that participants with HFA would show the same eye-directed gaze patterns as their TD peers, which was confirmed by a lack of group differences in looking time, number of fixations, and time to first fixation toward the eyes. These data support existing findings that there are no significant deficits for eye-directed gaze in this population [Bar-Haim et al., 2006; Grossman, Smith, Steinhart, & Mitchell, 2012; Nishimura et al., 2008; Rutherford & McIntosh, 2007; Rutherford & Towns, 2008], but stand in contrast to findings of deficits in this area [e.g. Klin et al., 2002; Neumann et al., 2006; Paul et al., 2013; Pelphrey et al., 2002]. The results presented here are the first data on gaze patterns of children with HFA in a nonemotional visual speech task and it is possible that the elimination of emotional content and focus on speech processing enabled individuals with HFA to gaze at the eyes more freely than when they contain potentially overstimulating emotional information [Klin et al., 2002; Neumann et al., 2006; Pelphrey et al., 2002]. These data also resonate with the findings of Lansing and McConkie [2003] that TD observers gazed at eyes more before and after speech, but redirected their gaze from the eyes to the mouth region during a visual speech task. Both participant cohorts may have recognized the relatively reduced contribution of the eye region in a difficult AV speech integration task and reduced their fixation time to this ROI. Alternatively, task demands appropriately and robustly drove TD fixations away from the eyes and toward the mouth while the HFA cohort did not alter their gaze patterns effectively, thereby masking a potential group difference in eye-directed gaze. Future studies should follow up on this interaction between AV speech task difficulty and eye gaze, as well as investigate whether individuals with ASD have the flexibility to adapt gaze behavior based on other task demands. It would be interesting to investigate gaze patterns for in-synch speakers, to determine potential differences in mouth-directed and eye-directed gaze behavior between TD and HFA individuals for visual speech in a simpler task.

Conclusion

Participants with HFA fixate significantly less on the mouth for AV asynchronous speech than TD peers, which stands in contrast to prior findings of increased mouth gaze for neutral or emotional faces. By fixating less on the crucially important mouth region and more on the irrelevant non-face area, individuals with HFA do not maximize their ability to integrate visual speech cues. These findings suggest that the visual fixation patterns of individuals with HFA to speaking faces are less adapted to task demands and not sufficiently focused on integrating AV dynamic speech cues. A potential intervention target for face-gaze in this population could, therefore, be flexible deployment of gaze toward sources of greatest information, directing children with HFA to increase gaze toward eyes when processing emotional facial expressions, but toward the mouth to enhance AV speech comprehension.

Acknowledgments

Funding was provided by NIDCD Grant R21 DC010867-01 (R. Grossman, P.I.) and by NIH Intellectual and Developmental Disabilities Research Center P30 Grant HDP30HD004147. We thank Christopher Bianrosa, Kerri Green, Gregory Hurst, Emily Levoy, Matthew Schneps, and Kelly Wessel for their assistance in stimulus creation. We also thank the children and families who gave their time to participate in this study. We have no conflict of interest to declare.

References

Bar-Haim, Y., Shulman, C., Lamy, D., & Reuveni, A. (2006). Attention to eyes and mouth in high-functioning children with autism. Journal of Autism and Developmental Disorders, 36(1), 131–137.

Bebko, J.M., Weiss, J.A., Demark, J.L., & Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. Journal of Child Psychology and Psychiatry, 47(1), 88–98.

Buchan, J.N., Pare, M., & Munhall, K.G. (2007). Spatial statistics of gaze fixations during dynamic face processing. Soc Neurosci, 2(1), 1–13.

de Gelder, B., Vroomen, J., & van der Heide, L. (1991). Face recognition and lip-reading in autism. European Journal of Cognitive Psychology, 3(1), 69–86.

de Boer-Schellekens, L., Eussen, M., & Vroomen, J. (2013). Diminished sensitivity of audiovisual temporal order in autism spectrum disorder. Frontiers in integrative neuroscience, 7.

Dunn, L.M., & Dunn, L.M. (1997). Peabody Picture Vocabulary Test (3 ed.). Circle Pines, MN: American Guidance Service.

Foss-Feig, J., Kwakye, L., Cascio, C., Burnette, C., Kadivar, H., Stone, W., et al. (2010). An extended multisensory temporal binding window in autism spectrum disorders. Experimental Brain Research, 203(2), 381–389.

Grossman, R.B., Schneps, M.H., & Tager-Flusberg, H. (2009). Slipped lips: Onset asynchrony detection of auditory-visual language in autism. Journal of Child Psychology and Psychiatry, 50(4), 491–497.

Grossman, R.B., Smith, A., Steinhart, E., & Mitchell, T. (2012). Visual scanning of facial expression with high and low

intensity. Poster presented at Gatlinburg Conference, Annapolis, MD.

Hall, G.B., Szechtman, H., & Nahmias, C. (2003). Enhanced salience and emotion recognition in Autism: A PET study. Am J Psychiatry, 160(8), 1439–1441.

Hampson, E., van Anders, S.M., & Mullin, L.I. (2006). A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis. Evolution and Human Behavior, 27(6), 401–416.

Iarocci, G., & McDonald, J. (2006a). Sensory integration and the perceptual experience of persons with autism. Journal of Autism & Developmental Disorders, 36(1), 77–90.

Iarocci, G., & McDonald, J. (2006b). Sensory Integration and the Perceptual Experience of Persons with Autism. [Article]. Journal of Autism & Developmental Disorders, 36(1), 77–90.

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. Autism, 14(4), 305–320.

Ijsseldijk, F. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. Journal of Speech Language & Hearing Research, 35(2), 466–471.

Irwin, J. (2006). Audiovisual speech integration in children with autism spectrum disorders. Journal of the Accoustic Society of America, 120(5, Pt 2), 3348.

Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. Archives of General Psychiatry, 65(8), 946–954.

Joseph, R.M., & Tanaka, J. (2003). Holistic and part-based face recognition in children with autism. Journal of Child Psychology and Psychiatry, 44(4), 529–542.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Archives in General Psychiatry, 59(9), 809–816.

Kozlowski, L., & Cutting, J. (1977). Recognizing the sex of a walker from a dynamic point-light display. Perception & Psychophysics, 21(6), 575–580.

Kwakye, L.D., Foss-Feig, J.H., Cascio, C.J., Stone, W.L., & Wallace, M.T. (2011). Altered auditory and multisensory temporal processing in autism spectrum disorders. Front Integr Neurosci, 4, 129.

Langdell, T. (1978). Recognition of faces: An approach to the study of autism. Journal of Child Psychology and Psychiatry, 19(3), 255–268.

Lansing, C.R., & McConkie, G.W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. [Research Support, U.S. Gov't, P.H.S.]. Perception & psychophysics, 65(4), 536–552.

Lord, C., Rutter, M., DiLavore, P.C., & Risi, S. (1999). Autism Diagnostic Observation Schedule - WPS (ADOS-WPS). Los Angeles, CA: Western Psychological Services.

Magnée, M.J., de Gelder, B., van Engeland, H., & Kemner, C. (2008). Audiovisual speech integration in pervasive developmental disorder: Evidence from event-related potentials. Journal of Child Psychology and Psychiatry, 49(9), 995–1000.

Marassa, L.K., & Lansing, C.R. (1995). Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible. Journal of Speech Language & Hearing Research, 38(6), 1387–1394.

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

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748.

Medeiros, K., & Winsler, A. (2014). Parent–child gesture use during problem solving in autistic spectrum disorder. Journal of Autism and Developmental Disorders, 44(8), 1946–1958.

Neumann, D., Spezio, M.L., Piven, J., & Adolphs, R. (2006). Looking you in the mouth: Abnormal gaze in autism resulting from impaired top-down modulation of visual attention. Social Cognitive and Affective Neuroscience, 1(3), 194–202.

Nishimura, M., Rutherford, M.D., & Maurer, D. (2008). Converging evidence of configural processing of faces in high-functioning adults with autism spectrum disorders. Visual Cognition, 16(7), 859–891.

Nishiyama, T., & Kanne, S. (2014). On the Misapplication of the BAPQ in a Study of Autism. Journal of Autism and Developmental Disorders, 44(8), 2079–2080.

Nuske, H., Vivanti, G., & Dissanayake, C. (2014). Brief Report: Evidence for normative resting-state physiology in autism. Journal of Autism and Developmental Disorders, 44(8), 2057–2063.

Paré, M., Richler, R., Hove, M., & Munhall, K.G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. Perception & psychophysics, 65(4), 553–567.

Paul, R., Campbell, D., Gilbert, K., & Tsiouri, I. (2013). Comparing spoken language treatments for minimally verbal preschoolers with autism spectrum disorders. Journal of Autism & Developmental Disorders, 43(2), 418–431.

Pelphrey, K.A., Sasson, N.J., Reznick, J.S., Paul, G., Goldman, B.D., & Piven, J. (2002). Visual scanning of faces in autism. Journal of Autism & Developmental Disorders, 32(4), 249–261.

Piven, J., & Sasson, N. (2014). On the misapplication of the broad autism phenotype questionnaire in a study of autism. Journal of Autism and Developmental Disorders, 44(8), 2077–2078.

Roid, G.H., & Miller, L.J. (1997). Leiter International Performance Scale—Revised. Wood Dale, IL: Stoelting Co.

Rutherford, M., & McIntosh, D. (2007). Rules versus Prototype Matching: Strategies of perception of emotional facial expressions in the autism spectrum. Journal of Autism and Developmental Disorders, 37(2), 187–196.

Rutherford, M., & Towns, A. (2008). Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. Journal of Autism and Developmental Disorders, 38(7), 1371–1381.

Shams, L., Kamitani, Y., & Shimojo, S. (2004). Modulation of visual perception by sound. In Calvert, G.A., Spence, C., & Stein, B.E. (Eds.), The handbook of multisensory processes (pp. 26–33). Cambridge, MA: MIT Press.

Smith, E.G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. Journal of Child Psychology and Psychiatry, 48(8), 813–821.

Spezio, M.L., Adolphs, R., Hurley, R.S., & Piven, J. (2007). Abnormal use of facial information in high-functioning autism. Journal of Autism & Developmental Disorders, 37(5), 929–939.

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

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. The Quarterly Journal of Experimental Psychology Section A, 36(1), 51–74.

Ulloa, E.R., & Pineda, J.A. (2007). Recognition of point-light biological motion: Mu rhythms and mirror neuron activity. Behavioural Brain Research, 183(2), 188–194.

van der Smagt, M.J., van Engeland, H., & Kemner, C. (2007). Brief report: Can you see what is not there? low-level auditory-visual integration in autism spectrum disorder. Journal of Autism & Developmental Disorders, 37(10), 2014–2019.

Vatikiotis-Bateson, E., Eigsti, I.M., & Yano, S. (1994). Listener eye movement behavior during audiovisual perception. Proceedings of the acoustical society of Japan, 94(3), 679–680.

Vatikiotis-Bateson, E., Eigsti, I.M., Yano, S., & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. [Comparative Study]. Perception & psychophysics, 60(6), 926–940.

Zainal, H., Magiati, I., Tan, J.-L., Sung, M., Fung, D.S., & Howlin, P. (2014). A Preliminary investigation of the spence children's anxiety parent scale as a screening tool for anxiety in young people with autism spectrum disorders. Journal of Autism and Developmental Disorders, 44(8), 1982–1994.