Sensitivity to Audio-Visual Synchrony and Its Relation to Language Abilities in Children with and without ASD

Giulia Righi, Elena J. Tenenbaum, Carolyn McCormick, Megan Blossom, Dima Amso, and Stephen J. Sheinkopf

Autism Spectrum Disorder (ASD) is often accompanied by deficits in speech and language processing. Speech processing relies heavily on the integration of auditory and visual information, and it has been suggested that the ability to detect correspondence between auditory and visual signals helps to lay the foundation for successful language development. The goal of the present study was to examine whether young children with ASD show reduced sensitivity to temporal asynchronies in a speech processing task when compared to typically developing controls, and to examine how this sensitivity might relate to language proficiency. Using automated eye tracking methods, we found that children with ASD failed to demonstrate sensitivity to asynchronies of 0.3s, 0.6s, or 1.0s between a video of a woman speaking and the corresponding audio track. In contrast, typically developing children who were language-matched to the ASD group, were sensitive to both 0.6s and 1.0s asynchronies. We also demonstrated that individual differences in sensitivity to audiovisual asynchronies and individual differences in orientation to relevant facial features were both correlated with scores on a standardized measure of language abilities. Results are discussed in the context of attention to visual language and audio-visual processing as potential precursors to language impairment in ASD. ***Autism Res*** *2018, 11: 645–653.* © 2018 International Society for Autism Research, Wiley Periodicals, Inc.

Lay Summary: Speech processing relies heavily on the integration of auditory and visual information, and it has been suggested that the ability to detect correspondence between auditory and visual signals helps to lay the foundation for successful language development. The goal of the present study was to explore whether children with ASD process audio-visual synchrony in ways comparable to their typically developing peers, and the relationship between preference for synchrony and language ability. Results showed that there are differences in attention to audiovisual synchrony between typically developing children and children with ASD. Preference for synchrony was related to the language abilities of children across groups.

Keywords: autism; audio-visual synchrony; eye-tracking; language development

Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social communication and interaction, as well as restricted and repetitive patterns of behavior, interests, or activities [APA, 2013]. ASD is often accompanied by deficits in language processing. Though deficits crossing all language domains are now well documented in the ASD literature [see Eigsti et al., 2011 for a review], the etiology of these deficits remains a source of debate.

Speech perception is a complex phenomenon that encompasses numerous processes and signals. Naturalistic face-to-face speech processing is a multisensory experience that relies heavily on the integration of auditory and visual information [Erber, 1975; Schwartz et al, 2004; Sumby & Pollack, 1954]. Sensitivity to the multimodality of speech arises early in life, and numerous studies have shown that typically developing infants show a preference for auditory and visually congruent stimuli [Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 2003]. Furthermore, it has been suggested that infants' preference for multimodally congruent stimuli and their ability to detect correspondence between auditory and visual signals helps to lay the foundation for successful language development [Bahrick & Lickliter, 2012]. Processing of

From the Department of Psychiatry and Human Behavior, Warren Alpert Medical School of Brown University (G.R., E.J.T., C.M., S.J.S.); Bradley Hospital, Providence, RI (G.R.); Brown Center for the Study of Children at Risk at Women and Infants Hospital (E.J.T., C.M., S.J.S.); Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI (D.A.); Department of Pediatrics, Women and Infants Hospital, Providence, RI (S.J.S.); Purdue University (C.M.); Psychology Department, Castleton University, Castleton, VT (M.B.); Rhode Island Consortium for Autism Research and Treatment, East Providence, RI (G.R., E.J.T., C.M., S.J.S.)

Received February 18, 2017; accepted for publication November 28, 2017

Address for correspondence and reprints: Giulia Righi, Emma Pendleton Bradley Hospital, 1011 Veteran's Memorial Parkway, East Providence, RI 02915. E-mail: giulia.righi@lifespan.org

Published online 13 January 2018 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.1918

multimodal signals has received much attention in the ASD literature, and most studies to date point to both atypical behavioral responses and/or atypical processing of audiovisual stimuli. Unlike typically developing infants, infants who later develop ASD showed reduced looking to relevant facial features when presented with speaking faces, but did not show comparable reductions in looking patterns when presented with faces without associated speech [Shic et al., 2014]. Similarly, children and adolescents with ASD display reduced attention to salient facial features when presented with dynamic speaking faces [Johnels et al., 2014; Norbury et al., 2009]. In the context of speech perception, differences in response to synchrony between auditory and visual signals between individuals with ASD and controls have been documented across the developmental span [see Stevenson et al., 2016 for a review], with few exceptions [Grossman, Steinhart, Mitchell, & Mcilvane, 2015]. These differences have been reported from as early as 9 months of age and extend to children and adolescents, with atypical responses to mismatched auditory/visual signals in a McGurk paradigm that indicate reduced influence of visual information [Guiraud, Tomalski, Kushnerenko, Ribeiro, Davies, & Charman, 2012; Mongillo et al., 2008]. Children and adolescents with ASD also display atypical multisensory processing with younger participants failing to show preferences for synchronized audio-visual presentations of socially salient stimuli [Bebko et al., 2006], and older participants failing to show expected behavioral facilitatory effects of a synchronous multimodal signal in the context of a noisy environment [Irwin et al., 2011; Foxe et al., 2015; Smith & Bennetto, 2007; Stevenson et al., 2017a].

These processing atypicalities have been primarily ascribed to deficits in two mechanisms, which are not necessarily separate: the perception of temporal synchrony, and the allocation of attention. Impairments in detecting the temporal characteristics of multimodal signals have been documented extensively in ASD [see Stevenson et al., 2014 for a review]. A recent study has demonstrated that temporal processing is correlated with sensory integration, and the ability to accurately detect speech in noise in ASD only [Stevenson et al., 2017b].

Closely tied to the ability to process speech is the tendency to attend to visual features associated with the auditory signal. A number of studies have demonstrated that attention to the mouth and eyes is correlated with greater success in language learning [Tenenbaum et al., 2014; Young et al., 2009]. This correlation is often attributed to the processing gain one achieves by attending to these informative regions of the face. Information about how sounds are formed in the mouth as well as information about the speaker's affect or intentions can all support language learning. An alternative account for the connections between attention to the mouth and eyes and successful language learning may be related to the processing of audio-visual information. That is, rather than increased access to information leading to gains in language development, poor processing of audio-visual synchrony could be driving both aversion from these regions of the face and limited success in language learning. If for instance one has difficulty processing the audio and visual streams simultaneously, averting one's attention from those regions that signal synchrony might improve the listening experience. This improvement may unfortunately be coupled with the loss of information available in those regions that could otherwise have contributed to greater learning success.

To our knowledge only one study has explicitly examined the association between response to audio-visual asynchrony and language skills [Patten et al., 2014]. In this study utilizing a preferential looking paradigm, 3–6-year-old children with ASD were presented with videos of a doll being handled by an adult. In the synchronous condition, the doll was bounced when its name was heard, whereas in the asynchronous condition, the doll was bounced 700 ms after its name was heard. Even though, on average, the participants looked significantly more at the synchronous video, individual differences in looking time were captured by a positive correlation between time spent looking at the synchronous video and receptive language abilities measured using a standardized instrument. While these results are intriguing, the use of dolls rather than naturalistic talking faces leaves open the question of how visual information in speech may affect processing for children with ASD.

The goal of the present study was twofold. First, we sought to evaluate the response of young children with ASD to temporal asynchrony in a speech-processing task including faces. Second, we sought to examine how response to multimodal asynchrony relates to language proficiency. The use of faces as stimuli is particularly relevant because of their naturalistic validity in language learning. Based on previous findings [Bebko et al., 2006] we hypothesized that children with ASD would not demonstrate the expected preferential looking to synchrony at delays of 300 ms, 600 ms, or 1,000 ms in a speech-processing task. Furthermore, we hypothesized that spontaneous preferential looking to the synchronous video would be positively related to language abilities for all children [Patten et al., 2014].

Methods

Participants

Data was successfully collected from seventy-seven participants who were part of ongoing studies in two

Table 1. Age and Test Results

		ASD ($n = 45$)		TD ($n = 32$)			
		Mean (SD)	Range	Mean (SD)	Range	F	*P*
# Male		36		21			
Age in years		5.12 (1.53)	2.40–8.98	3.04 (1.70)	1.09–7.23	31.60**	<0.01
PLS composite age equivalent		34.80 (21.56)	8–90	40.86 (22.19)	13–95	1.35	0.28
PLS composite standard scores		72.20 (28.49)	50–116	94.79 (37.48)	82–136	9.15**	<0.01
Cognitive standard scores		69.73 (25.43)	10–122	106.34 (15.25)	74–137	48.67**	<0.01
ADOS	Social affective	13.17 (4.53)	2–20				
	Repetitive behaviors	4.34 (1.96)	1–8				
	Comparison score	7.44 (1.72)	4–10				

Cognitive scores are standard scores from the Bayley Scales of Infant Development (ASD: $n = 2$; TD: $n = 11$), the Wechsler Preschool and Primary Scale of Intelligence-III (ASD: $n = 20$; TD: $n = 8$), or the Stanford Binet Intelligence Scales (ASD: $n = 21$; TD: $n = 9$).

PLS, Preschool language scales; ADOS, Autism Diagnostic Observation Schedule.

$^{**}P < 0.01$.

laboratories. One study exploring word learning in ASD and was conducted at the Developmental Cognitive Neuroscience (DCN) Lab at Brown University ($n = 30$). The other was a study of biomarkers within ASD ($n = 47$) that was conducted at the Brown Center for the Study of Children at Risk (BCC) at Women and Infants Hospital in Providence, RI. Many of the participants for both studies were recruited through the Rhode Island Consortium for Autism Research and Treatment (RI-CART), a state-wide consortium that includes a registry of individuals with a diagnosis of ASD who are interested in participating in research studies. Both studies included an ASD sample and a typically developing (TD) control group and both involved a battery of eye tracking experiments such that these stimuli were presented following 5–10 min of eye tracking for other studies. The groups did not differ on the measures of interest in this study (see analysis below) and we therefore collapsed across samples. In total, the ASD group consisted of 45 participants (M = 5.12 years, SD = 1.53; 36 male, 9 female) with clinical diagnoses of ASD that were confirmed with research reliable administration of the Autism Diagnostic Observation Schedule 2nd Edition [Lord et al., 2012]; Module 1 ($n = 24$), Module 2 ($n = 10$), or Module 3 ($n = 10$). Included in the 45 participants with ASD was one child with a community diagnosis of autism though we were not able to obtain results from their ADOS and one child with ASD who failed to complete language testing. Data from these two participants were included in all analyses for which data was available. One additional child with ASD was tested but not included in the analysis due to equipment failure.

Thirty-two TD participants (M = 3.04 years, SD = 1.70; 21 male, 11 female) were included in the final sample. One additional TD participant was tested but not included in the analysis due to failure to complete the experiment. Three TD participants completed the experiment but did not return for language and cognitive testing. TD participants were recruited for the word learning study in the DCN lab based on language level and for the biomarkers study at the BCC on chronological age (though many were also matched on language level within that sample). The final combined samples did not differ in language abilities, but TD participants were younger on average and had higher cognitive scores than ASD participants. Means and outcomes of one-way ANOVAs exploring group differences in age, cognitive, and language test scores are provided in Table 1.

Cognitive and Language Tests

All participants underwent cognitive and language testing administered by trained research staff and supervised by a licensed clinical psychologist. Cognitive tests were either the Bayley Scales of Infant Development, 3rd Edition [Bayley, 2005], the Wechsler Preschool and Primary Scale of Intelligence-III [Wechsler, 2002], or the Stanford Binet Intelligence Scales, 5th Edition [Roid, 2003] depending on the age and verbal ability of the child and the study from which they were recruited. All participants were tested using the Preschool Language Scales (PLS)– 5th Edition [Zimmerman, Steiner, & Pond, 2011].

The PLS is a standardized assessment of receptive and expressive language that requires a trained examiner to administer and takes 45–60 min to complete. The exam is appropriate for children from birth to 7 years 11 months and covers the range from pre-verbal and interaction-based skills to emerging language and early literacy. Specific items explore attention, vocal/gesture behaviors, levels of play, vocabulary comprehension and production, letter naming, and use of irregular grammatical markings and sentence structure. Scores are reported as age-equivalent and standard scores in the receptive and expressive domains as well as a total

Figure 1. Screen capture from a sample trial of the experiment.

composite score. The measure was normed on a sample of 1,400 typically developing children.

Stimuli

Stimuli consisted of two identical videos presented side-by-side on a single computer screen while the audio track matching one or both videos was played. The videos were each 14 sec long and showed a woman speaking in highly animated child directed speech about topics of interest to young children (e.g., a visit to Grandma's house, a trip to the doctor, lunchtime, bath time). A screen shot of a sample trial is shown in Fig. 1. The videos were recorded spontaneously (i.e., without a script) to be as natural as possible. Twelve videos were used to make the stimuli. All participants saw three trials from each of four conditions in a passive viewing task. Conditions differed by length of delay of the asynchronous video. In one condition, both videos were synchronous with the audio track (0s Delay). In the remaining three conditions, one of the videos preceded the audio track by 0.3s (0.3s Delay), 0.6s (0.6s Delay), or 1 full second (1.0s Delay). The video length was trimmed from longer recordings to allow for audio to play the full length of the trial. Position of the synchronous video (left or right side of the screen) was counterbalanced across trials. Four versions of the experiment were created and each video was shown at each possible delay across participants. That is, all participants saw the same twelve videos but the length of the delay for a given video differed by experiment version.

Eye Tracking

Eye tracking at the DCN lab was completed on a SensoMotoric Instruments (SMI) RED500 Remote Eye Tracking system (SensoMotoric Instruments, Inc., Boston, MA). This system includes a remote-controlled infrared eye camera with automatic eye and head tracker. Tracking relies on a binocular image of the pupil and corneal reflection collected at a rate of 60 Hz with spatial resolution of 0.03° and gaze position accuracy of 0.4°. Blink recovery time is at maximum 4 ms and tracking recovery time for excessive movement is at maximum 90 ms.

An experimenter was seated at a computer adjacent to the display monitor, but hidden from view with a dark curtain. Calibration and stimulus presentation were displayed using the Experiment Center software provided by SMI. A two-point calibration was used. This involved presentation of an animated image designed to attract the infant's attention to fixation points at the top left and lower right corners of the screen. Calibration was then verified with a four-point display of animated objects. Average deviation from calibration in the X plane was 1.58° (SD = 1.69) and in Y, 1.74° (SD = 1.57). Deviation from calibration did not differ significantly between groups, X: $F(1,45) = 0.04$, $P = 0.85$; Y: $F(1,45) = 1.44$, $P = 0.24$.

Stimuli were presented on a 19" (48.26 cm) computer monitor. Children were seated approximately 70 cm from the display monitor. The woman's face was 24 cm high × 15 cm wide, subtending a visual angle of 19° × 12° at this distance.

Eye tracking at the BCC was completed on a SensoMotoric Instruments (SMI) RED250-mobile eye tracking system (SensoMotoric Instruments, Inc., Boston, MA). This system includes a remote-controlled infrared eye camera with automatic eye and head tracker. Tracking relies on a binocular image of the pupil and corneal reflection collected at a rate of 60 Hz with spatial resolution of 0.03° and gaze position accuracy of 0.4°. Blink recovery time is at maximum 4 ms and tracking allows for free head movement (40 cm × 20 cm at 70 cm).

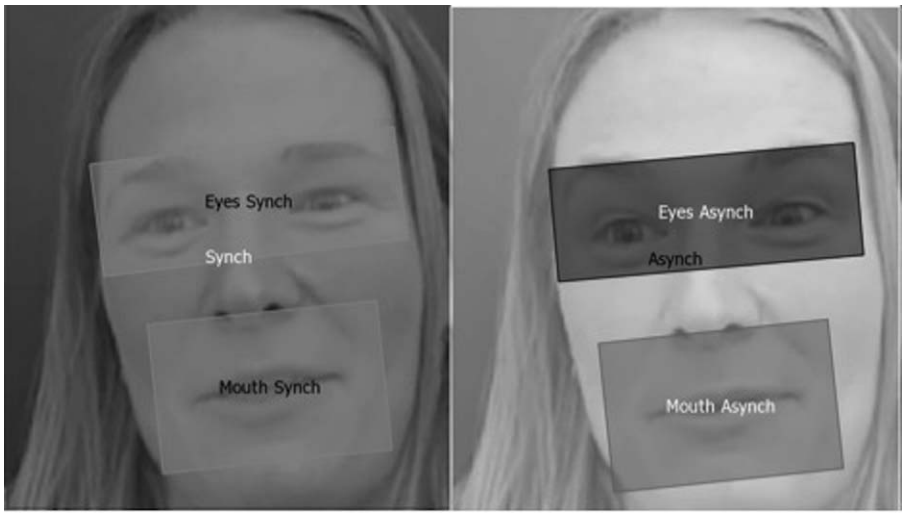

Figure 2. Screen capture from SMI BeGaze software (SensoMotoric Instruments, Inc., Boston, MA) illustrated areas of interest.

As in the DCN Lab, the experimenter was seated at a computer adjacent to the display monitor, but hidden from view with a dark curtain. Calibration and stimulus presentation were displayed using the Experiment Center software provided by SMI. A five-point calibration was used. This involved presentation of an animated image designed to attract the infant's attention to fixation points at the four corners and center of the screen. Calibration was assessed online and the experiment did not proceed until deviations in the X and Y planes were less than 1°. Average deviation from calibration in the X plane was 0.29° (SD = 0.22) and in Y, 0.34° (SD = 0.27). Deviation from calibration did not differ significantly between groups, X: $F(1,28) = 0.00$, $P = 0.95$; Y: $F(1,28) = 0.19$, $P = 0.67$. Stimuli were presented on a 27" (68.58 cm) computer monitor. Children were seated approximately 90 cm from the display monitor. The woman's face was 30.5 cm high × 19 cm wide, subtending a visual angle of 19° × 12° at this distance.

Procedure

Cognitive and language testing were completed in one visit and eye tracking was administered in a separate visit. Order of the visits differed by study (DCN lab completed eye tracking and then testing, BCC was the reverse). Following successful completion of the eye tracking experiment for which the child was recruited, parents were given the option to continue with this experiment. Calibration had been completed at the start of the eye tracking session and was not repeated. This experiment began directly with an attention capturing video presented at the center of the screen. Each of the twelve trials was preceded by a centrally fixated attention getter. The experiment lasted 3–5 min.

Analysis

Eye tracking data was preprocessed using SMI BeGaze native software (SensoMotoric Instruments, Inc., Boston, MA). Across labs, fixations were defined as at least 100 ms spent fixating a 100 pixel area. Two areas of interest surrounding the left and right side videos (thus covering the full screen) were created. These were identified as synchronous or asynchronous (where the synchronous video was randomly assigned to the right or left position for the 0s Delay condition). Within the two sides, areas of interest surrounding the eyes and mouth were defined dynamically across trials. Figure 2 shows a screen capture including the defined areas of interest. Time spent fixating the synchronous and asynchronous eyes and mouth were calculated as the percentage of successfully tracked trial time (maximum 14s) that each participant spent looking at those specific areas of interest (e.g., synchronous eyes/time spent fixating some area of the screen). Proportion of time spent fixating the synchronous video was also calculated relative to total tracked time (synchronous/synchronous + asynchronous). Trials on which the participant failed to look at the screen for at least 500 ms and trials on which a participant's proportion of looks to the synchronous video was greater than two standard deviations from the group mean were excluded from analysis. This resulted in the exclusion of 46/924 total trials across all participants. There was not a significant difference in the number of usable trials across groups, ASD: $M = 11.24$, TD: $M = 11.48$, $F(1,77) = 0.297$, $P = 0.59$, partial $\eta^2 = 0.06$. In the event that excluded trials resulted in zero usable trials for a given delay (ASD: $n = 3$, TD: $n = 1$), that participant's data was included for analyses with relevant data (e.g., overall fixation time), but excluded from analyses without (e.g., repeated measures ANOVAs).

Results

Preliminary analyses explored potential differences between laboratories. Neither total tracked time, $F(1,75) = 1.71$, $P = 0.20$, partial $\eta^2 = 0.02$, nor proportion of looks to the synchronous videos, $F(1,75) = 0.16$, $P = 0.70$, partial $\eta^2 = 0.002$, differed as a function of the lab in which participants were tested. Neither age, $F(1,75) = 0.23$, $P = 0.63$, partial $\eta^2 = 0.003$, nor ADOS severity scores, $F(1,39) = 0.75$, $P = 0.39$, partial $\eta^2 = 0.02$, differed by lab. PLS composite scores did differ between labs, $F(1,71) = 4.69$, $P = 0.03$, partial $\eta^2 = 0.06$, and the differences in cognitive capacity were marginally significant, $F(1,72) = 3.19$, $P = 0.08$, partial $\eta^2 = 0.04$. This difference was due to experimental design. The word learning study was specifically targeting early word learners whereas the biomarkers study included both low and high language abilities. Data for the remaining analyses were collapsed across sites.

Further preliminary analyses explored group differences in looking patterns overall. Participants with ASD did not differ from TD participants in average fixation duration, $F(1,74) = 0.002$, $P = 0.97$, partial $\eta^2 = 0.00$, (ASD: $M = 368.17$ ms, $SD = 130.66$; TD: $M = 366.53$ ms, $SD = 191.47$) or first looks to the screen within a trial, $F(1,74) = 1.64$, $P = 0.21$, partial $\eta^2 = 0.02$, (ASD: $M = 537.84$ ms, $SD = 1742.15$; TD: $M = 139.41$ ms, $SD = 286.44$). Participants with ASD looked less at the screen overall ($M = 6.59$s, $SD = 2.58$) than TD participants ($M = 8.56$s, $SD = 3.00$), $F(1,74) = 9.17$, $P = 0.003$, partial $\eta^2 = 0.11$. All eye tracking results are thus reported as proportion of total time spent looking at the screen that a participant was attending to a given area of interest.

Distribution of Attention to Areas of Interest

A 2 (Synchrony) × 4 (Delay: 0s, 0.3s, 0.6s, 1.0s) × 2 (Area of Interest: eyes vs. mouth) mixed-model repeated measures ANOVA with a between subjects factor of group was used to explore potential differences in attention to areas of interest as a function of diagnosis, synchrony, and delay. The between subjects factor of group was significant, $F(1,67) = 13.50$, $P = 0.000$, partial $\eta^2 = 0.17$ (ASD: $M = 0.16$, $SD = 0.04$; TD $M = 0.19$, $SD = 0.03$). The main effect of Synchrony was also significant, $F(1,67) = 9.30$, $P = 0.003$, partial $\eta^2 = 0.12$ (ASD: $M = 0.17$, $SD = 0.04$; TD $M = 0.19$, $SD = 0.04$), as was the interaction between Synchrony and Group, $F(1,67) = 4.41$, $P = 0.04$, partial $\eta^2 = 0.06$. There was a significant difference in looks to the synchronous vs. asynchronous videos within the TD group (Synchronous: $M = 0.21$, $SD = 0.05$; Asynchronous: $M = 0.17$, $SD = 0.04$). Participants with ASD did not demonstrate a preference for the synchronous videos (Synchronous: $M = 0.15$, $SD = 0.05$; Asynchronous: $M = 0.16$, $SD = 0.06$). The paired group comparison within the synchronous condition was also significant (ASD: $M = 0.15$, $SD = 0.05$; TD: $M = 0.21$, $SD = 0.05$). No other main effects or interactions were significant.

Bonferroni corrected planned comparisons between attention to the eyes and mouth by group revealed no significant differences in attention to the eyes between groups (ASD: $M = 0.18$, $SD = 0.02$; TD: $M = 0.19$, $SD = 0.02$, $P = 0.70$, $\eta 2 = 0.00$) but a significant difference in attention to the mouth with TD participants focusing greater proportions of their looks at the mouth than ASD participants (ASD: $M = 0.13$, $SD = 0.02$; TD: $M = 0.19$, $SD = 0.02$, $P < 0.05$, $\eta^2 = 0.07$).

Distribution of Attention by Delay

As discussed above, the recognition of synchrony has been shown to differ by delay. To explore whether this was the case in the current study, we ran planned comparisons within groups using one-sample *t*-test with 0.50 (chance distribution of attention) as a test of preference for the synchronous video and Bonferroni corrected *alpha* of 0.006. TD participants were reliably orienting to the synchronous videos in the 1.0s Delay, $t(29) = 3.17$, $P = 0.004$, $d = 1.18$, and 0.6s Delay, $t(30) = 3.65$, $P = 0.001$, $d = 1.33$, conditions, but not in the 0.3s Delay condition, $t(29) = -0.96$, $P = 0.35$, $d = 0.36$,. As expected, there was no preference in the 0s Delay (synchronous) condition, $t(27) = 1.51$, $P = 0.14$, $d = 0.58$. ASD participants were not reliably orienting to the synchronous video at any delay, 1.0s Delay: $t(38) = 0.52$, $P = 0.61$, $d = 0.17$; 0.6s Delay: $t(40) = 1.75$, $P = 0.09$, $d = 0.55$; 0.3s Delay: $t(39) = -0.90$, $P = 0.37$, $d = 0.29$. As expected, there was no preference in the 0s Delay condition: $t(42) = -0.17$, $P = 0.67$, $d = 0.05$. Proportion of looks to the synchronous video by delay and group is shown in Fig. 3. These results suggest that TD participants were successful in discriminating between the synchronous and asynchronous videos and showed a preference for the synchronous videos at 0.6 s and 1.0 s delays while the ASD participants did not demonstrate any preference for the synchronous over the asynchronous video at any of the delays. Comparisons between groups using independent sample *t*-tests showed a trend for ASD participants to look less at the synchronous videos than TD participants in the 0.6s, $t(67) = 1.46$, $P = 0.15$, $d = 0.36$, and 1.0s Delay conditions, $t(61) = 1.65$, $P = 0.11$, $d = 0.42$, but these differences were not significant.

Relation between Preference for the Synchronous Video and PLS Scores

To explore whether preference for the synchronous video is related to language ability, we conducted linear

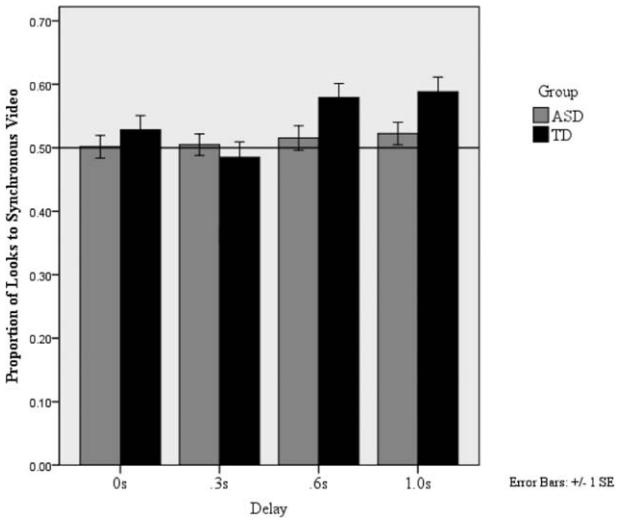

Figure 3. Proportion of looking time spent fixating the synchronous video by delay and group. Reference line shows 0.50 or chance distribution of attention. "Synchronous" video in the 0 delay condition was randomly assigned to the right or left side.

Table 2. Beta Weights for Linear Regression with Proportion of Looks to Synchronous Videos (Synchronous/Synchronous + Asynchronous) at Each Delay in Relation to PLS Composite Age Equivalent Scores, $F(4,49) = 5.76$, $P = 0.001$, $R^2 = 0.32$

Predictor	Beta Weights	*P*
Chronological age	**0.42**	0.001
0.3s delay	0.08	0.498
0.6s delay	0.10	0.394
1.0s delay	**0.30**	0.021

The bold values represent statistically significant results as it can be verified by the *P* values listed in the third column.

regression analyses on data collapsed across the two groups. Given the broad range in age of the participants and the fact that age tends to correlate with PLS scores (Zimmerman, Steiner, & Pond, 2011), we included chronological age as a covariate in the regression analysis. In the first regression analysis, we regressed overall attention to the synchronous videos (looks to the synchronous/synchronous + asynchronous) at 0.3s, 0.6s, and 1.0s delays on PLS-5 composite age-equivalent scores. The regression model was significant overall, $F(4,49) = 5.76$, $P = 0.001$, $R^2 = 0.32$, with age and proportion of looks to the synchronous video in the 1.0s Delay condition emerging as significant predictors of language scores (see Table 2). This suggests that attention to the synchronous videos in the 1.0s Delay condition was related to language abilities.

In light of the positive relation between attention to the synchronous video and language scores at the 1.0s delay, we ran a similar linear regression in which we broke down the proportion of looks to the synchronous

Table 3. Beta Weights for Linear Regression with Proportion of Looks to the Synchronous Eyes and Mouth (AOI/Tracked Time) in the 0.6s Delay Condition in Relation to PLS Composite Age Equivalent Scores, $F(3,63) = 10.36$, $P < 0.001$, $R^2 = 0.33$

Predictor	Beta	*P*
Chronological age	**0.47**	0.000
Mouth 1.0s delay	**0.35**	0.006
Eyes 1.0s delay	**0.42**	0.002

The bold values represent statistically significant results as it can be verified by the *P* values listed in the third column.

video in the 1.0s Delay condition by attention to the eyes and mouth (time spent fixating that AOI/tracked time). The overall model was again significant, $F(3,63) = 10.36$, $P < 0.001$, $R^2 = 0.33$, with Age and Attention to the Mouth and Eyes all contributing significantly to the prediction of language scores (see Table 3). This suggests that attention to both the eyes and mouth of the synchronous video at 1.0s Delay was related to superior language scores across the sample.

Discussion

The goals of the present study were twofold: (a) to examine audio-visual speech processing in children with ASD relative to TD controls; and (b) to capture the relation between audio-visual speech processing and language skills to further our understanding of factors that may lead to language deficits. In a preferential looking paradigm designed to explore audiovisual speech processing, typically developing children showed a reliable preference for synchronous audio and video stimuli in comparison to audiovisual asynchronies of 0.6s and 1.0s. In contrast, children with ASD did not display any reliable preferences for the synchronous or asynchronous videos at any delay tested. Regression analyses revealed positive correlations between preference for synchrony at 1.0s and standardized language scores. The present results provide further evidence for atypical processing of multimodal speech in children with ASD. To our knowledge, this is the first study to demonstrate a significant relationship between processing of audiovisual asynchrony during naturalistic speech perception and language abilities in children with and without ASD.

The lack of preference for synchronous audiovisual stimuli could be interpreted as an inability to detect the temporal characteristics of the auditory signal in relation to the videos. It has been posited that impaired multimodal processing is a potential precursor to the language deficits observed in ASD [Stevenson et al., 2016]. Within our sample, participants' proportion of looks directed towards the synchronous video was positively related to their overall language skills. Using naturalistic speech stimuli, this result extends prior work

demonstrated a positive relationship between audiovisual speech processing and receptive language abilities [Patten et al., 2014]. Even though causality cannot be inferred from the present results, our data support the hypothesized relationship between language deficits observed in ASD and atypical multisensory processing [Stevenson et al., 2016, 2017]. Interpreted in this way, these results together with prior research may help to generate hypotheses about low level deficits that may precede or influence the development of deficits in higher level language abilities in ASD.

A second related finding showed a positive relationship between language ability and time spent looking at the eyes and mouth regions of the speaker's face in the synchronous videos across all participants, regardless of diagnostic status. This finding is not surprising, as the majority of the information needed to determine synchronicity between audio and visual speech streams is contained in these regions [Foster & Oberlander, 2007; Graf et al., 2002]. Therefore, this result might be due to the fact that the children with more effective multisensory processing abilities are better able to utilize salient facial information, which in turn leads to better language outcomes.

Further, consistent with previous findings [Chawarska, Macari, & Shic, 2012] participants with ASD looked less at the mouth of the speaker than TD controls in this experiment. This provides additional support for the notion that children with ASD may be missing critical linguistic information by not attending to the mouth resulting in deficits in the processing of audiovisual synchrony and language ability. However, our data do not allow us to rule out the alternative account whereby atypical processing of audiovisual synchrony makes attention to the mouth aversive for individuals with ASD. Future studies should address this distinction more directly, perhaps by exploring physiological responses to synchronous and asynchronous stimuli.

The current study is not without limitations. The TD and ASD groups differed in average chronological age, with the ASD group containing older participants. In order to account for the variability in age between the TD and ASD groups, as well as the relationship between age and language level (Zimmerman et al., 2011), age was added in order to account for any within-group variability primarily due to this factor. Nevertheless, we believe that language-matching the two samples, even though it led to group differences in chronological age, was a reasonable strategy to ensure that participants were able to process the stimuli in comparable ways.

Though the present results cannot speak to the etiology of the observed differences in audiovisual speech processing in ASD, they validate the existence of a relationship between multisensory processing of audiovisual speech and language proficiency. Multisensory perceptual atypicalities manifest themselves both in lack of spontaneous preference for synchrony and in the allocation of attention to relevant features.

Acknowledgments

This research was supported by grants from the Autism Science Foundation (11–1013) and Autism Speaks (7897), NIMH T-32 5T32MH019927-24.

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bahrick, L.E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. Bremner, D. J. Lewkowicz, & C. Spence (Eds.). Multisensory development (pp. 183–205). Oxford, England: Oxford University Press.

Bayley, N. (2005). Bayley scales of infant development (3rd ed.). San Antonio: PsychCorp.

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 and Allied Disciplines, 47, 88–98.

Chawarska, K., Macari, S., & Shic, F. (2012). Context modulates attention to social scenes in toddlers with autism. Journal of Child Psychology and Psychiatry, 53, 903–913. doi: 10.1111/j.1469-7610.2012.02538.x

Eigsti, I.M., De Marchena, A.B., Schuh, J.M., & Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. Research in Autism Spectrum Disorders, 5, 681–691.

Erber, N.P. (1975). Auditory-visual perception of speech. Journal of Speech and Hearing Disorders, 40, 481–492.

Foster, M.E., & Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. Language Resources and Evaluation, 41, 305–323.

Foxe, J.J., Molholm, S., Del Bene, V.A., Frey, H.P., Russo, N.N., Blanco, D., … Ross, L.A. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. Cerebral Cortex, 25, 298–312. http://doi.org/10.1093/cercor/bht213

Graf, H.P., Cosatto, E., Strom, V., & Huang, F.J. (2002). Visual prosody: Facial movements accompanying speech. *In Proceedings of the 5th IEEE Automatic Face and Gesture Recognition, 2002* (pp. 396–401). IEEE.

Grossman, R.B., Steinhart, E., Mitchell, T., & Mcilvane, W. (2015). "Look who's talking!" Gaze Patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. Autism Research, 8, 307–316.

Guiraud, J.A., Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T. … the BASIS Team. (2012). Atypical audiovisual speech integration in infants at risk for

autism. PLoS One, 7, e36428. http://doi.org/10.1371/journal.pone.0036428

Irwin, J.R., Tornatore, L.A., Brancazio, L., & Whalen, D.H. (2011). Can children with autism spectrum disorders "hear" a speaking face?. Child Development, 82, 1397–1403. http://doi.org/10.1111/j.1467-8624.2011.01619.x

Johnels, J.A., Gillberg, C., Falck-Ytter, T., & Miniscalco, C. (2014). Face viewing patterns in young children with autism spectrum disorders: Speaking up for a role of language comprehension. Journal of Speech, Language, and Hearing Research, 57, 1679–1691. (June), http://doi.org/10.1044/2014_JSLHR-L-13-0268

Kuhl, P.K., & Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. Science, 218, 1138–1141. http://doi.org/10.1126/science.7146899

Kuhl, P.K., & Meltzoff, A.N. (1984). The intermodal representation of speech in infants. Infant Behavior & Development, 7, 361–381.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). Autism diagnostic observation schedule (ADOS-2) (2nd ed.). Los Angeles, CA: Western Psychological Corporation.

Mongillo, E.A., Irwin, J.R., Whalen, D.H., Klaiman, C., Carter, A.S., & Schultz, R.T. (2008). Audiovisual processing in children with and without autism spectrum disorders. Journal of Autism and Developmental Disorder, 38, 1349–1358. http://doi.org/10.1007/s10803-007-0521-y

Norbury, C.F., Brock, J., Cragg, L., Einav, S., Griffiths, H., & Nation, K. (2009). Eye-movement patterns are associated with communicative competence in autistic spectrum disorders. Journal of Child Psychology and Psychiatry and Allied Disciplines, 50, 834–842. http://doi.org/10.1111/j.1469-7610.2009.02073.x

Patten, E., Watson, L.R., & Baranek, G.T. (2014). Temporal synchrony detection and associations with language in young children with ASD. Autism Research and Treatment, 1–8. http://doi.org/10.1155/2014/678346

Patterson, M.L., & Werker, J.F. (2003). Two-month-old infants match phonetic information in lips and voice. Developmental Science, 6, 191–196. http://doi.org/10.1111/1467-7687.00271

Roid, G.H. (2003). Stanford-Binet intelligence scales: Riverside Publishing Itasca, IL.

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month olds who develop autism spectrum disorder. Biological Psychiatry, 75, 231–237. http://doi.org/10.1016/j.biopsych.2013.07.009

Schwartz, J.L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. Cognition, 93, B69–B78.

Smith, E.G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. Journal of Child Psychol Psychiatry, 48, 813–821. http://doi.org/10.1111/j.1469-7610.2007.01766.x

Stevenson, R.A., Siemann, J.K., Woynaroski, T.G., Schneider, B.C., Eberly, H.E., Camarata, S.M., & Wallace, M.T. (2014). Brief report: Arrested development of audiovisual speech perception in autism spectrum disorders. Journal of Autism and Developmental Disorders, 44, 1470–1477. http://doi.org/10.1007/s10803-013-1992-7

Stevenson, R.A., Segers, M., Ferber, S., Barense, M.D., Camarata, S., & Wallace, M.T. (2016). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. Autism Research, 9, 720–738. http://doi.org/10.1002/aur.1566

Stevenson, R.A., Baum, S.H., Segers, M., Ferber, S., Barense, M.D., & Wallace, M.T. (2017). Multisensory speech perception in autism spectrum disorder: From phoneme to whole-word perception. Autism Research, 10, 1280–1290. http://doi.org/10.1002/aur.1776

Stevenson, R.A., Segers, M., Ncube, B.L., Black, K.R., Bebko, J.M., Ferber, S., & Barense, M.D. (2017). The cascading influence of multisensory processing on speech perception in autism. Autism, 136236131770441. http://doi.org/10.1177/1362361317704413

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

Tenenbaum, E., Sobel, D., Sheinkopf, S., Malle, B., & Morgan, J. (2014). Attention to the mouth and gaze following in infancy predict language development. Journal of Child Language, 5, 413–427. http://doi.org/10.1017/S0305000914000725

Young, G.S., Merin, N., Rogers, S.J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. Developmental Science, 12, 798–814. http://doi.org/10.1111/j.1467-7687.2009.00833.x

Zimmerman, I.L., Steiner, V.G., & Pond, R.E. (2011). Preschool language scales (PLS-5) (5th ed.). Bloomington, MN: Pearson.