Severe Multisensory Speech Integration Deficits in High-Functioning School-Aged Children with Autism Spectrum Disorder (ASD) and Their Resolution During Early Adolescence

John J. Foxe1,2,4,5, Sophie Molholm1,2,4,5, Victor A. Del Bene1,2,3, Hans-Peter Frey1,2, Natalie N. Russo1,2,6, Daniella Blanco1,2,4,5, Dave Saint-Amour7,8 and Lars A. Ross1,2,9

1Department of Pediatrics, 2Department of Neuroscience, The Sheryl and Daniel R. Tishman Cognitive Neurophysiology Laboratory, Children’s Evaluation and Rehabilitation Center (CERC), 3Ferkauf Graduate School of Psychology, Albert Einstein College of Medicine, Bronx, NY 10461, USA, 4Department of Psychology, 5Department of Biology, The Cognitive Neurophysiology Laboratory, Program in Cognitive Neuroscience, City College of the City University of New York, New York, NY 10031, USA, 6Department of Psychology, Syracuse University, Syracuse, NY 13244, USA, 7Centre de Recherche, CHU Sainte-Justine, 3175, Côte-Sainte-Catherine Montréal, Montréal, QC, Canada H3T 1C5, 8Département de Psychologie, Université du Québec à Montréal (UQAM), Montréal, QC, Canada H3C 3P8 and 9The Gordon F. Derner Institute of Advanced Psychological Studies, Adelphi University, Garden City, NY 11530, USA

Address correspondence to John J. Foxe. Email: john.foxe@einstein.yu.edu

Under noisy listening conditions, visualizing a speaker’s articulations substantially improves speech intelligibility. This multisensory speech integration ability is crucial to effective communication, and the appropriate development of this capacity greatly impacts a child’s ability to successfully navigate educational and social settings. Research shows that multisensory integration abilities continue developing late into childhood. The primary aim here was to track the development of these abilities in children with autism, since multisensory deficits are increasingly recognized as a component of the autism spectrum disorder (ASD) phenotype. The abilities of high-functioning ASD children (n = 84) to integrate seen and heard speech were assessed cross-sectionally, while environmental noise levels were systematically manipulated, comparing them with age-matched neurotypical children (n = 142). Severe integration deficits were uncovered in ASD, which were increasingly pronounced as background noise increased. These deficits were evident in school-aged ASD children (5–12 year olds), but were fully ameliorated in ASD children entering adolescence (13–15 year olds). The severity of multisensory deficits uncovered has important implications for educators and clinicians working in ASD. We consider the observation that the multisensory speech system recovers substantially in adolescence as an indication that it is likely amenable to intervention during earlier childhood, with potentially profound implications for the development of social communication abilities in ASD children.

Keywords: autism spectrum disorders, cross-modal, development, sensory integration, speech-in-noise

Introduction

Effective decoding of the speech signal is critical to normal human communication. Although we experience speech as an auditory phenomenon, it is clear that visualizing the articulations of a speaker can play a hugely important role in the intelligibility of this signal (Sumby and Pollack 1954; Erber 1969). This is especially the case when the acoustic environment is noisy, such as when multiple individuals are speaking simultaneously, or in settings like a busy street or on a windy day (Ross et al. 2007a). These multisensory integrative abilities allow us to better extract the basic information content in speech (i.e. what the words are), but it is also the case that information about the emotionality, intonation, intentionality, excitability state, and stress patterns of the speaker are afforded through both the auditory and visual channels (Chen et al. 2010; Stienent et al. 2011). This latter class of multisensory prosodic information provides a large part of the socially relevant content in the speech signal, and it has been known for decades now that individuals with autism spectrum disorder (ASD) have significant deficits in extracting this class of information (Kujala et al. 2005; Paul et al. 2005). It is also the case that basic unisensory processing disturbances have been seen in ASD (Fiebelkorn et al. 2013), and it is not surprising then that researchers have asked whether some of these social communication deficits might not arise from a more basic deficit in multisensory integration processes (Iarocci and McDonald 2006; Foxe and Molholm 2009).

For example, recent electrophysiological work has pointed to deficits in the integration of simple somatosensory and auditory inputs in high-functioning ASD children (Russo et al. 2010), as well as deficits in the processing of fundamental audiovisual inputs (Brandwein et al. 2013). With regard to multisensory integration of speech inputs, there is now a considerable collection of studies on this issue in ASD. By far the most common method applied in assaying multisensory speech perception in ASD has involved the well-known “McGurk illusion,” where dubbing of an auditory phoneme (e.g., /ba/) onto a video of a speaker making an incongruent articulatory movement corresponding to a different phoneme (e.g., /va/) can lead to strong illusory auditory perceptions (McGurk and MacDonald 1976; Saint-Amour et al. 2007). These illusory multisensory fusions provide a powerful window onto the ongoing influence that visual inputs exert over our auditory speech perceptions, and a nice tool for testing the integrity of these processes in clinical populations. However, when it comes to McGurk studies in ASD participants, the results have been somewhat mixed. Early work suggested lower sensitivity to McGurk fusions in ASD children (de Gelder et al. 1991), a finding that was replicated in a number of subsequent studies (Mongillo et al. 2008; Irwin et al. 2011). Others, however, have claimed that these apparent differences in multisensory integrative abilities are accounted for by basic deficits in speechreading (Williams et al. 2004; Iarocci et al. 2010). Another issue that arises is the span of ages that have been assessed in many of these studies, which are all of relatively limited sample size. In almost all cases, data from children as young as...
4 or 5 years of age are lumped together with that from children in their early teenage years and late adolescence. Further complicating this picture, when adults with ASD were assessed for McGurk sensitivity, results suggested that multisensory speech processing was essentially fully intact (Keane et al. 2010), whereas a similar study in adults with Asperger's syndrome pointed to some subtle differences in McGurk sensitivity (Saastad et al. 2012). These latter studies suggest that there is likely a developmental trajectory to McGurk sensitivity in ASD, with older individuals showing normal or near-normal performance levels. Indeed, work has also shown that the sensitivity of neurotypical children to McGurk fusions continues to develop quite late into childhood, with one study showing that as late as 9–10 years of age, adult-like sensitivity to McGurk fusions was still not fully established (Tremblay et al. 2007).

To our knowledge, only one study has attempted to assess the development of McGurk sensitivity in ASD individuals. Taylor et al. (2010) proposed that one clear possibility with the existing literature on audiovisual integration was that the deficits seen in previous studies might, in fact, reflect a developmental delay, rather than a frank-and-fixed deficit across the lifespan. In a relatively large sample of ASD children (n = 24) that ranged in age from 8 to 16 years, their work with the McGurk effect suggested that children at the older end of their age range might be “catching up” with their typical peers. In their experiment, children were simply required to repeat the phonemes they heard. The authors then used regression analyses to assess whether there was a significant developmental trajectory for a given ability. When the children repeated auditory-only phonemes, both groups performed equivalently and with high rates of accuracy. Indeed, performance was essentially as high in the youngest children tested as it was in the oldest, although the ASD group did show marginal evidence for improvement with age, whereas the neurotypical children did not. Interestingly, both groups showed significant developmental improvements across the age range when they were measured for visual-only accuracy (i.e. speechreading) and the ASD children were, indeed, significantly worse at speechreading than the neurotypical children. However, there was no Group × Age interaction, so the rate of improvement across childhood was equivalent across both groups. In contrast, whereas sensitivity to McGurk fusions remained constant across ages in the neurotypical group, there was an observable increase in sensitivity as a function of age in the ASD group. Inspection of their regression fits suggests that the ASD group did not reach typical McGurk sensitivity levels until the 14- to 16-year-age bracket, a level of sensitivity that was already present in neurotypicals as young as 8–10 years in their sample.

Of course, the possibility that ASD children may catch up during adolescence has potentially very significant implications for treatment and intervention. If it is, indeed, the case that ASD individuals catch up without actually undergoing explicit intervention strategies designed to ameliorate their multisensory deficits, then the capacity might be there for earlier intervention and significant amelioration of these deficits at younger ages. In turn, improvements in multisensory speech capacities could generalize to other aspects of social interaction and impact these truly debilitating aspects of ASD.

In our view, a key limitation of the use of the McGurk paradigm to test speech integration is that neither the visual nor auditory inputs are degraded or embedded in noise and, in this regard, this paradigm does not test multisensory speech integration under the environmental conditions where one would expect it to be most important for function. Multisensory audiovisual improvement during speech recognition is substantially greater when the speech is presented in noisy and distracting environments (Ross et al. 2007a, 2007b; Ma et al. 2009; Ross et al. 2011), and it is precisely such environments that are thought to present the greatest difficulties for individuals with autism (Russo et al. 2009). There are 2 studies, to date, that we are aware of, which have tested speech integration abilities in variable noise settings in ASD. In the first of these, 18 high-functioning ASD adolescents aged between 12.4 and 19.5 years were tested in a speech-in-noise design using short 5-to-7 word sentences, where the task for the participants was to identify 3 keywords in each sentence (Smith and Bennetto 2007). While both ASD and neurotypical adolescents had thresholds of about −19 dB signal-to-noise ratio (SNR) in an auditory-alone condition when the visual input was available, thresholds for neurotypical participants dropped to approximately −26 dB, whereas the ASD participants achieved a rather more modest improvement to around −22 dB. The ASD participants also showed significantly poorer speechreading skills, and hierarchical regression analysis showed that this unisensory deficit accounted for a good deal of the variance in their audiovisual speech perception abilities, although unisensory factors did not predict all of the variance, suggesting a specific multisensory integrative component.

Smith and Bennetto did not specifically examine the extent of the multisensory deficit as a function of age in their study, and so one might expect that they would also have observed an amelioration of the deficit as their participants aged. It is also puzzling that almost all of their children were in the age range in which the work of Taylor et al. (2010) suggests that multisensory speech mechanisms should have normalized. One possible explanation clearly arises in the differences between what is being tested with the McGurk paradigm and the speech-in-noise paradigm. Another possible source of this apparent speech integration deficit may relate to fixation patterns, which are well known to be altered in ASD. There were no eye-tracking or other formal measures taken to ensure that gaze-direction differences were not present between groups, so this remains a possible source of the discrepancy between theirs and the Taylor study, although the Taylor study did not use eye-tracking techniques either.

More recently, the abilities of ASD children to benefit from visual inputs during a simple phoneme recognition task under variable noise conditions were assessed in a relatively small and younger ASD cohort (n = 13; aged 5–15 years), where participants were explicitly monitored using eye tracking (Irwin et al. 2011). As above, Irwin was understandably concerned that some of the previous findings of integration deficits in ASD might have been due to the well-known differences in fixation patterns that are often reported in this population and the lack of appropriate fixation verification in the vast majority of the previous studies. When these authors removed trials where fixations were inaccurate, and such trials were, indeed, found to be considerably more common in their ASD cohort, they nonetheless found that the ASD group benefitted less from available visible articulatory inputs than did their typical development (TD) counterparts.

Work from our research group has shown that multisensory speech integration continues to develop well into the later
childhood years in typically developing children and adolescents (Ross et al. 2011), and this is also the case for the integration of very basic tones and flashes (Brandwein et al. 2011). Similarly, late developmental emergence of visuo-haptic integration during shape size discriminations has also been seen (Gori et al. 2008). Thus, there appears to be a particularly elongated developmental trajectory for even the most basic multisensory processes (Gori et al. 2012). Another facet of multisensory integration that shows a protracted developmental trajectory is the so-called temporal window of integration (TWIN). As one might expect, inputs from 2 sensory modalities are optimally integrated when they occur in very close temporal proximity (Meredith et al. 1987) and the efficacy of integration decreases as the 2 elements are presented with greater temporal separations. When basic auditory and visual stimuli are used, this TWIN has a characteristic temporal aperture. It was recently shown that children as old as 10 or 11 years have a longer TWIN than adults when auditory inputs precede visual inputs (Hillock et al. 2011). It is becoming clear that any consideration of potential multisensory deficits in ASD will need to contend with these protracted developmental changes evident in neurotypicals.

Also worth pointing out is that many of the studies of audiovisual condition (AV) speech using the McGurk illusion or phoneme recognition use very limited and essentially very simplistic stimulus sets. In many of the studies, between 3 and 6 possible phoneme-voicing pairs were used and these were also often repeated many times over in the service of collecting sufficient trial numbers for analysis. As Keane et al. (2010) has pointed out, audiovisual integration deficits may well be uncovered using more challenging tasks and more varied stimulus sets. Certainly one possibility that needs to be entertained when children are being asked to repeat a phoneme from a limited set of possibilities is whether neurotypicals might not better realize the closed nature of the stimulus set than ASD participants and, in turn, make better guesses. One could certainly also raise questions about the real-world applicability of such an experimental set-up. In the present study, we used a large stimulus set of real monosyllabic words (n = 300), where no word was presented more than once. The use of real words of high frequency from the lexicon, in an entirely unpredictable open set of stimuli, was designed to much better approximate real-world naturalistic circumstances.

It is also the case that, in many of the preceding multisensory studies in children with ASD, the age ranges of the samples were typically between 5 and 20, with samples generally being quite small (range = 9–24; mean = 14.5). In an earlier study, in a considerably larger cohort of neurotypical children, adolescents, and adults (n = 58), we found that multisensory speech processing showed a very steep improvement over these exact age ranges. Typical children in the 10- to 11-year-age group were still developing their integrative abilities, whereas 12–14 year olds had attained near-adult levels. Here, we sought to comprehensively assess multisensory speech-in-noise abilities in a large cohort of children and adolescents with ASD (n = 84), and to track the developmental trajectory of their multisensory speech abilities relative to a large cohort of neurotypical control participants (n = 142). The intention was to have sufficiently dense sampling to assess whether ASD individuals would show a typical growth function in their abilities to integrate seen and heard speech, and whether they would, indeed, catch up to their typical peers by late childhood or adolescence.

### Methods

#### Participants

A total of 236 children ranging in age from 5 to 17 years participated in this study. The data of 10 participants (<5% of the sample) were not entered into the final analysis due to technical difficulties or noncompliance. Of the final 226 participants, 84 had a diagnosis of ASD (age range: 5–17 years; M = 10.82; SD = 2.86), and 142 met criteria for TD (age range: 5–17 years; M = 11.32; SD = 3.4). Demographics for the 2 groups as a function of the 5 age groups considered in our analyses are presented in Table 1.

All participants were native English speakers. Participants were excluded from this study if they had a history of seizures or had uncorrected vision problems. TD children were excluded if they had a history of psychiatric, educational, attentional, or other developmental difficulties as assessed by a history questionnaire and were also excluded if their parents endorsed 6 or more items of inattention or hyperactivity on a DSM-IV checklist for attention deficit disorder (with and without hyperactivity). Diagnoses of ASD were obtained by a trained clinical psychologist using the Autism Diagnostic Interview-R (Lord et al. 1994) and the Autism Diagnostic Observation Schedule (ADOS-G; Lord et al. 2000).

The full-scale, verbal (VIQ), and performance (PIQ) intelligence quotients were assessed in 108 TD and 64 ASD children with the Wechsler Abbreviated Scales of Intelligence (WASI). The descriptive statistics for 5 TD and ASD subgroups are summarized in Table 1. All children had normal or corrected-to-normal vision, and audiometric threshold evaluation confirmed that all children had within-normal limits hearing. The parents of all child participants provided written informed consent in accordance with the tenets of the 1964 Declaration of Helsinki. All procedures were approved by the institutional review board(s) of the City College of New York and the Albert Einstein College of Medicine.

### Table 1

<table>
<thead>
<tr>
<th>Age</th>
<th>TD</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>5–6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>7–9</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>10–12</td>
<td>11.37 (0.72)</td>
<td>27</td>
</tr>
<tr>
<td>13–15</td>
<td>14.14 (0.85)</td>
<td>31</td>
</tr>
<tr>
<td>16–17</td>
<td>16.68 (0.48)</td>
<td>20</td>
</tr>
<tr>
<td>M</td>
<td>142</td>
<td>84</td>
</tr>
</tbody>
</table>

Note: The number of TD (nTD) and ASD (nASD) participants in respective age groups and the number of participants for whom IQ scores were obtained (nIQ). VIQ: verbal IQ; PIQ: performance IQ; FSIQ: full-scale IQ assessed using the WASI. Asterisks denote significant differences (uncorrected, t-tests α = 0.05) between TD and ASD groups for a given measure (*P < 0.05; **P < 0.01). M: Overall weighted mean VIQ, PIQ and FSIQ scores.
Stimuli and Task
Stimulus materials consisted of digital recordings of 300 simple monosyllabic words spoken by a female speaker. This set of words was a subset of the stimulus material created for a previous experiment in our laboratory (Ross et al. 2007a, 2007b) and used in a previous study (Ross et al. 2011). These words were taken from the “MRC Psycholinguistic Database” (Coltheart 1981) and were selected from a well-characterized normed set based on their written word frequency (Kucera and Francis 1967). The subset of words for the present experiment is a selection of simple, high-frequency words from a child’s everyday environment and is likely to be in the lexicon of children in the age range of our sample. The recorded movies were digitally remastered, so that the length of the movie (1.3 s) and the onset of the acoustic signal were similar across all words. Average voice onset occurred at 520 ms after movie onset (SD = 30 ms). The words were presented at approximately 50 dBA sound pressure level (SPL), at 7 levels of intelligibility including a condition with no noise (NN) and 6 conditions with added pink noise at 50, 53, 56, 59, 62, and 65 dBA SPL sound pressure. Noise onset was synchronized with movie onset. The SNRs were therefore NN, 0, −3, −6, −9, −12, and −15 dBA SPL. These SNRs were chosen to cover a performance range in the auditory-alone condition from 0% recognized words at the lowest SNR to almost perfect recognition performance with NN. The movies were presented on a monitor (NEC Multisync FE 2111SB) at 80 cm distance from the eyes of the participants. The face of the speaker extended approximately 6.4° of visual angle horizontally and 8.58° vertically (hairline to chin). The words and pink noise were presented over headphones (Sennheiser, model HD 555).

The main experiment consisted of 3 randomly intermixed conditions: In the auditory-alone condition (A-alone), the auditory words were presented in conjunction with a still image of the speakers face; in the AV, the auditory words were presented in conjunction with the corresponding video of the speaker articulating the words. Finally, in the visual-alone condition (V-alone), only the video of the speaker’s articulations was presented. The word stimuli were presented in a fixed order, and the condition (the noise level and whether it was presented as A-alone, V-alone, or AV) was assigned to each word randomly. Stimuli were presented in 15 blocks of 20 words with a total of 300 stimulus presentations. There were 140 stimuli for the A and AV conditions (20 stimuli per condition and intelligibility level) and 20 stimuli for the V condition that was presented without noise.

Task
Participants were instructed to watch the screen and report which word they heard (or saw in the V-alone condition). If a word was not clearly understood, participants were encouraged to make their best guess. An experimenter, seated approximately 1 m distance from the participant at 90° to the participant-screen axis, monitored participant’s adherence to maintaining fixation on the screen. Only responses that exactly matched the presented word were considered correct. Any other response was recorded as incorrect.

Eye Tracking
Eye movements were recorded using an EyeLink 1000 system (SR Research, Ontario, Canada), at a sampling rate of 500 Hz. A small target sticker was placed on the participants’ forehead, allowing the system to compensate for head movements of up to 20 cm. To prevent larger head movements, the children had to place their heads on a comfortable chin rest. Before each set of 5 blocks of stimuli (or more often if necessary), the eye-tracking system was calibrated using a 9-point calibration. Saccades and fixations were defined by the Eyelink system using the default settings. Eye-movement data were collected for 127 participants; however, 6 datasets had to be removed due to recording errors (5 TD and 1 ASD).

Eye-tracking data were analyzed using custom Matlab scripts (Mathworks, Natick, MA USA) in the same age groups as for the behavioral analysis. As a first step, we determined the proportion of fixations on the different parts of the speaker’s face. This was accomplished by selecting 3 rectangular patches, each covering the face, mouth, or eyes, and determining the proportion of fixations within these patches. Since lips and jaw move during speech production, the mouth region was defined vertically from the bottom of the lower jaw to just below the nose (the nasolabial angle of the philtrum). These measures were taken from the still image of the speaker before articulation started (i.e. with the mouth closed). The proportions of fixations in the different groups were statistically compared using a z-statistic for proportions, while the numbers of fixations were compared using a t-test for independent samples. Data were analyzed for each experimental condition separately as well as for all conditions combined. To be as sensitive as possible to any possible differences between the participant groups, no correction for multiple comparisons was employed.

In addition to the statistical analyses, fixation distribution maps were created by convolving, at each fixation location, a unit impulse with a 2D Gaussian with half-width at half-height of 1° visual angle. The size of the Gaussian was chosen in accordance with previous studies (Frey et al. 2008). A value of 1 indicates that all participants fixated the same location during each fixation. Lower values indicate that participants either are not consistent in where they look or fixate different parts of the face in successive fixations. For example, if there are 2 small objects of interest in a scene, which are consistently fixated (equally often) by all participants, then the fixation map will have 2 peaks with a height of about 0.5. These maps, therefore, provide information about how consistent fixation locations are between participants and throughout the different trials.

Analyses of Task Performance
We submitted percent correct responses for each condition to a repeated-measures analysis of variance (RM-ANOVA) with factors of stimulus condition (auditory vs. audiovisual), SNR level (7 levels), and the-between-subjects factor of the group (TD vs. ASD) as well as age as a covariate. Performance in the V-alone condition was analyzed separately, because it was only presented without noise. Violations of the sphericity assumption of the RM-ANOVA were corrected by adjusting the degrees of freedom with the Greenhouse-Geisser correction method. We expected significant main effects of condition, SNR level, group, and age, as well as an interaction between condition and SNR level replicating previous findings (Ross et al. 2007a, 2007b; Ma et al. 2009; Ross et al. 2011). We expected a decrease in the ability to benefit from visual speech to manifest itself as an interaction of the group factor with the condition and SNR level. To provide the reader with an easy-to-interpret characterization of the group differences, we displayed A and AV speech perception performance as well as AV gain as it unfolded over all intelligibility conditions (Fig. 1). This analysis was performed for 3 subgroups of ages: 7–9, 10–12, and 13–15 years of age. Participants between ages of 5 and 6 and 16 and 17, either end of the age distribution, were excluded from statistical analyses due to low ASD participant numbers. Audiovisual enhancement (or AV gain) was operationalized here as the difference in performance between the AV and A-alone conditions (AV−A-alone). This analysis was performed at the 4 lowest SNRs, because the variance at higher SNRs becomes increasingly constrained by ceiling performance, and for consistency with our prior work (Ross et al. 2011).

Results
The Effect of SNR, Speaker Articulation, and age on Word Recognition Performance
In line with previous findings (Sumby and Pollack 1954; Ross et al. 2011), the RM-ANOVA returned main effects (Greenhouse-Geisser corrected for the violation of sphericity) of SNR ($F_{4.75,1066}=206.43; P<0.001; \eta^2 = 0.48$) and condition ($F_{1,225}=22.56; P<0.001; \eta^2 = 0.09$), showing that performance decreased as SNR decreased and was significantly better when visualized speech was present. The factors of condition and SNR showed a significant interaction ($F_{5,27,1175}=8.68; P<0.001; \eta^2 = 0.04$), reflecting the fact that AV gain was not uniform over SNRs. As expected, overall age affected

Cerebral Cortex February 2015, V 25 N 2 301
performance significantly ($F_{1, 223} = 112.73; P < 0.001; \eta^2 = 0.34$), and a significant interaction between age and SNR suggested that the effect of noise level on performance was related to age ($F_{4.75, 1060} = 7.48; P < 0.001; \eta^2 = 0.03$). A significant 3-way interaction between age, condition, and SNR ($F_{5.23, 1175} = 9.14; P < 0.001; \eta^2 = 0.04$) indicated that the interaction between condition and SNR was also related to the age of the observer.

**Differences Between TD and ASD Children**

The RM-ANOVA returned a significant main effect of group ($F_{1, 223} = 58.71; P < 0.001; \eta^2 = 0.21$) and interactions between group and condition ($F_{1, 223} = 27.23; P < 0.001; \eta^2 = 0.12$), as well as a significant 3-way interaction between group, condition, and SNR ($F_{5.23, 1175} = 9.14; P < 0.001; \eta^2 = 0.04$). To delineate group differences in more detail, we divided our sample into separate age groups (7–9; 10–12; 13–15) and plotted average auditory-alone performance, AV performance, and AV gain for each SNR (Fig. 1).

In both groups, A and AV performance were high (>80% of words correctly identified) when the words were presented without noise and showed only minimal improvement with age, indicating that the task difficulty was appropriate for all ages and both groups. Children with ASD performed lower in the A condition in all age groups, a difference that was on average 4.23% (see Table 2) and, therefore, numerically small but nevertheless significant (ages 7–9: $t_{61} = 2.18; P = 0.033$; ages 10–12: $t_{70} = 3.17; P = 0.002$; ages 13–15: $t_{50} = 2.39; P = 0.021$). Cohen’s $d$ amounted to $d = 0.56$ (7–9), $d = 0.74$ (10–12), and $d = 0.66$ (13–15), respectively. The top left panel of Figure 2A illustrates a gradual linear improvement in auditory-alone performance over age in both groups, with a delay of approximately 2–3 years in the ASD group.

We observed large group differences in the AV condition and, therefore, audiovisual gain between the ages of 7–9 ($t_{61} = 5.89; P < 0.001; d = 1.56$) and 10 and 12 ($t_{70} = 5.87; P < 0.001; d = 1.39$), while this performance difference was entirely absent in the oldest age group ($t_{50} = 1.03; P = 0.31$). Accordingly, group differences in audiovisual gain were also observed between 7 and 9 years ($t_{61} = 4.88; P < 0.001; d = 1.3$) and 10–12 years ($t_{70} = 4.53; P < 0.001; d = 1.3$), but not after the age of 12 years ($t_{50} = 0.26; P = 0.8$). Figure 2B illustrates that AV performance in TD children shows a sharper increase between 5 and 8 years and continues to improve more slowly well into the later teenage years. In contrast, in ASD, limited improvement in performance with age is observed until the

---

**Figure 1.** Multisensory speech recognition performance as a function of diagnosis and age group. Average word recognition performance (% correct) at each SNR and in the no-noise condition (NN) are plotted for both the A and AV conditions for 3 age groups (A: 7–9; B: 10–12; C: 13–15) for TD (blue) and ASD (red) participants. Multisensory gain is represented in the plot of the difference (AV – A). Bar graphs display gain averaged over the 4 lowest SNRs ($−15, −12, −9, \text{ and } −6$) for TD and ASD groups, as well as the results of 2-tailed $t$-tests with effect size (Cohen’s $d$).
Table 2

<table>
<thead>
<tr>
<th>Age Group</th>
<th>TD (M)</th>
<th>ASD (M)</th>
<th>TD (SD)</th>
<th>ASD (SD)</th>
<th>TD ASD</th>
<th>TD ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>7–9</td>
<td>3.26±6.4</td>
<td>0.75±5.9</td>
<td>12.4±11.3</td>
<td>23.2±16.1</td>
<td>0.61</td>
<td>13.8</td>
</tr>
<tr>
<td>10–12</td>
<td>3.26±6.4</td>
<td>0.75±5.9</td>
<td>12.4±11.3</td>
<td>23.2±16.1</td>
<td>0.61</td>
<td>13.8</td>
</tr>
<tr>
<td>13–15</td>
<td>3.26±6.4</td>
<td>0.75±5.9</td>
<td>12.4±11.3</td>
<td>23.2±16.1</td>
<td>0.61</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Note: Mean percent correct word identification in the A, AV, and V conditions, as well as percent gain (AV−A), as a function of diagnosis (TD and ASD) and age group (7–9, 10–12, and 13–15 years). AV gain was calculated as the mean difference (lower−upper).
model and assessed if the variable group could contribute significantly to the success of the model after the entry of age and VIQ. As expected, the overall model including all 3 predictors was significant ($F_{3,96} = 25.34; P < 0.001$), explaining approximately 43% of the variance ($R^2 = 0.425$). Most importantly, predictor group accounted for an additional 20% of the variance ($R^2_{inc} = 0.28$) after age and VIQ had been entered and was therefore by far the strongest predictor in the model ($F_{(1,96)inc} = 34.41; P < 0.001$). Please see Table 4 for regression coefficients and change statistics.

**Table 3**

<table>
<thead>
<tr>
<th>TD</th>
<th>A</th>
<th>V</th>
<th>AV</th>
<th>Gain</th>
<th>A</th>
<th>V</th>
<th>AV</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIQ</td>
<td>0.3**</td>
<td>0.07</td>
<td>0.27**</td>
<td>0.1</td>
<td>0.37**</td>
<td>0.12</td>
<td>0.46***</td>
<td>0.28*</td>
</tr>
<tr>
<td>PIQ</td>
<td>0.33**</td>
<td>0.08</td>
<td>0.24*</td>
<td>0.1</td>
<td>0.46***</td>
<td>0.18</td>
<td>0.42**</td>
<td>0.28*</td>
</tr>
</tbody>
</table>

Note: Values represent partial correlation scores ($r$) between VIQ and PIQ and percent word identification A, V, AV, and AV gain in TD and ASD children. Correlations are adjusted for the effect of age.

Significant correlations are marked with an asterisk (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

model and assessed if the variable group could contribute significantly to the success of the model after the entry of age and VIQ. As expected, the overall model including all 3 predictors was significant ($F_{3,96} = 25.34; P < 0.001$), explaining approximately 43% of the variance ($R^2 = 0.425$). Most importantly, predictor group accounted for an additional 20% of the variance ($R^2_{inc} = 0.28$) after age and VIQ had been entered and was therefore by far the strongest predictor in the model ($F_{(1,96)inc} = 34.41; P < 0.001$). Please see Table 4 for regression coefficients and change statistics.

**Relationships Between A, V, and AV Performance**

The above statistical analysis suggested that age and group and IQ are significantly related to performance in speech perception. Here, we tested the relationships between our dependent measures for each group while controlling for age and IQ. In both groups, participants’ speech perception in the auditory-alone condition showed no significant relationship with speechreading abilities. However, the partial correlation procedure uncovered significant correlations between AV performance and speechreading in both TD ($r = 0.53; r^2 = 0.28; P < 0.001$) and ASD ($r = 0.33; r^2 = 0.11; P = 0.009$) children. AV performance correlated with auditory-alone performance in both the ASD ($r = 0.51; r^2 = 0.26; P = 0.001$) and TD groups ($r = 0.22; r^2 = 0.05; P = 0.02$).

**The Relationships Between Speech Perception in Noise, Symptom Severity, and Sensory Profile**

Based on the fact that speech perception and audiovisual speech processing are fundamental for communication, it was reasonable to assume that performance on these measures would relate to symptom severity in autism and social communication skills. We therefore conducted a partial correlation analysis (adjusted for age and VIQ) between the speech perception performance measures and overall symptom severity calibrated for age and language level from the ADOS (ADOS_CSS). We also computed partial correlation coefficients between total scores on subscale A (language and communication functioning) of the ADI and performance measures. This correlation analysis returned no significant effects and is reported in Supplementary Table 1.

**Figure 2.** Scatter plots of individual participant performance in the A, V, and AV conditions, and AV gain (AV–A), as a function of age in months. Explained variance ($R^2$) of the linear regression lines for TD (blue) and ASD (red) participants are displayed. Insert panels display the same data expressed as group averaged performance over 1-year increments in the A, AV, and V conditions, and the gain function (AV–A), for TD and ASD children from ages 5 to 17 years of age. For gain, the average of the lowest 4 SNR levels ($−15, −12, −9, and −6$ dBA) is represented, whereas for the main conditions, the average of all SNR levels is represented.
We also explored possible relationships between speech perception and aspects of behaviors and performance relating to sensory processing in ASD children as measured by an abbreviated version of the “Sensory Profile Caregiver Questionnaire.” After correcting for multiple comparisons (Bonferroni), no significant or near significant relationships between speech perception and the total scores of the sensory profile and 7 subscales were found (see Supplementary Table 2).

**Eye Tracking**

The first line of analysis examined fixations on the regions of interest in the 3 different age ranges. No significant differences between TD and ASD groups were uncovered in terms of numbers of fixations or the proportion of fixations on regions of interest in the age ranges of 7–9 (23 TD and 18 ASD), 10–12 (21 TD and 23 ASD), or 13–15 (27 TD and 9 ASD; Table 5). Combining all age groups, the distribution of fixations on the face during the speech stimulus was also statistically indistinguishable between groups (Fig. 3). No significant difference was found in the proportion of fixations on any part of the face (all \( z < 1.5 \)).

Although there were no significant differences in any experimental condition or for the different conditions combined, the difference in the proportion of fixations on the mouth, especially in the youngest age group, was numerically higher. We therefore examined whether the proportion of fixations on the mouth might be related to behavioral performance in the audiovisual task using partial correlation, while controlling for age. In the control group, there was no significant correlation between the proportion of fixations on the mouth and the mean gain at low SNR (\( r_{08} = 0.16; P > 0.19 \)). However, in the ASD group, we found a small, albeit significant, partial correlation between these 2 variables (\( r_{17} = 0.29; P = 0.04 \)). In the ASD group, there were 5 cases with notably lower proportions of fixations on the mouth (<10%). Reanalysis excluding these cases eliminated the correlation between mean gain at low SNR and the proportion of fixations on the mouth (\( r_{18} = 0.24; P = 0.12 \)). Notably, the multisensory behavioral performance differences in the youngest age groups remained (\( t_{15} = 2.82; P < 0.01 \) and \( t_{14} = 3.47; P < 0.01 \), respectively), while performance in the oldest age bracket was indistinguishable between groups (\( t_{33} = 0.71; P > 0.48 \)).

**Discussion**

We asked whether high-functioning children with ASD would show deficits in their abilities to integrate seen and heard speech, especially under noisy environmental conditions. Such deficits could have profound implications for rehabilitative efforts, particularly with regard to typical classroom settings and teaching strategies. The data point to robust and relatively severe deficits in multisensory speech perception across the early and middle school years (5–12 years of age). These deficits could not be explained by aberrant gaze fixation patterns and were not explainable in terms of unisensory functioning. That is, the multisensory deficits were not simply the result of poorer auditory-alone or visual-alone abilities. It is a reasonable assumption that these deficits are present at the earliest stages of speech acquisition, considerably earlier than the low cutoff age of the present sample, and are likely even more severe in younger children. The data also show that this deficit is compounded as the level of background noise is increased. It could be argued that this latter effect implicates attentional functions as opposed to multisensory mechanisms, on the
premise that ASD participants might have become more distractable under noisier conditions, but this seems highly unlikely based on the results of the unisensory auditory-alone condition. If a simple attentional account applied, then one would surely expect that a similar if not more severe decline in performance would be evident when only auditory words were presented in high noise. Yet, in the unisensory auditory-alone condition, ASD children showed a wholly similar monotonous function to that of their TD counterparts and appeared no more affected by increasing noise levels.

Perhaps the most remarkable aspect of the current data is the finding that ASD children of 13 years of age and above showed no evidence whatsoever of a multisensory speech integration deficit. Although this study used a cross-sectional approach, the apparent recovery of function in this older ASD group suggests a very rapid and quite discrete developmental change as ASD children enter adolescence. We can only speculate at this juncture as to what changes might drive such an apparently dramatic recovery of function. It is plausible, for example, that differential myelination patterns along the major white matter tracts that communicate between the auditory and visual systems could lead to differential developmental trajectories for multisensory integrative functions. Along very similar lines, recent work using diffusion-weighted imaging techniques has shown relationships between longitudinal measures of white matter tract development and the development of reading skills in 7- to 15-year-old neurotypical children (Yeatman et al. 2012), with considerable variability in the patterns observed across children. We will return to the issue of neural connectivity in more detail below.

While a structural account is appealing, there are certainly other possible mediating factors. One alternative might relate to puberty and potential increases in social interest. Although this is pure speculation, it seems possible that as ASD children enter puberty with the attendant hormonal changes and drives, there is pressure to operate more effectively in social contexts, and that this results in a period of intensified efforts to interact with peers. We are unaware of any studies that have yet explored the development of social interactions across this age band in ASD. However, that we have shown that multisensory speech integration abilities continue to develop into early adolescence in neurotypical children shows that the system continues to have “plasticity” at this relatively late stage of development (Ross et al. 2011), so it is entirely conceivable that a late multisensory learning spurt in ASD would be possible with the appropriate level of practice and motivation. In this regard, it is of significant interest that children above the age of 12 years with ASD show a significant improvement in their abilities to recognize emotions from the upper aspects of facial stimuli (Kuusikko et al. 2009).

**The Present Data and Prior Speech-in-Noise Investigations in ASD**

As mentioned in the Introduction section, there have been only 2 previous studies of multisensory speech integration in ASD, that we are aware of, where background noise levels were expressly manipulated (Smith and Bennetto 2007; Irwin et al. 2011). Our findings in adolescents are somewhat at odds with the first of these, where the abilities of 18 ASD adolescents between 12.4 and 19.5 years to identify keywords in short sentences-in-noise were assessed. Note that this is nearly exactly the age range where entirely typical levels of multisensory speech performance are found herein, and yet Smith and Bennetto found deficits. While they did not specifically examine the extent of the multisensory deficit as a function of age, and it seems likely that they would also have seen improvements in their older adolescents, it remains the case that even their youngest children were already in or approaching the age range at which we see normalization of functioning here. As mentioned previously, one possible source of the discrepancy between theirs and our study may relate to the issue of fixation since they did not expressly track participants’ gaze patterns. While we did track eyes here in the majority of our participants, we actually found that ASD participants generally maintained very good fixations throughout the experiment, so it could be argued that gaze issues are not a particular problem in studies of this nature. A possibility that deserves consideration, however, is that the very act of measuring gaze, with the attendant calibration sessions and the presence of the eye-tracking camera on the surface in front of the participants, may well have had a significant impact on the compliance of our participants. It is important to note that gaze patterns were closely monitored on a computer screen from outside the testing booth throughout the experiment, and children were immediately coached by the experimenter if any lapsing of fixation was noted. Another obvious difference between their study and ours is that we used only single monosyllabic words, whereas they used short but complete sentences. Again here, fixation differences may be implicated, since there are
considerably greater fixation demands during the protracted durations of short sentences than that for singular punctate word utterances.

Another possible difference is that their stimuli contained contextual cues such that one might predict that recognition of any 1 of the 3 keywords within a given sentence might provide cues as to the identity of the other target words. It is plausible that the TD children were better able to use contextual cues to decipher the sentences than the ASD children. In their study, a response was only marked as “correct” if all 3 keywords in the 5–7 word sentences were recognized. Of course, the same conditions pertained in their auditory-alone condition, so while contextual cues may have played a role, they would have had to do so differentially in the multisensory setting to explain the differences between theirs and our study. An additional consideration is the format of their manipulation of background noise relative to ours. Here, we used simple pink noise, whereas in their study, they created background noise by combining the soundtracks of 4 additional speakers reading from story books after low-pass filtering the speech streams to remove articulatory sounds. Despite this low-pass filtering, one could argue that this version of multispeaker background might have presented more of a challenge to their ASD participants. In any case, their results suggest that deficits in multisensory speech recognition continue into adolescence in ASD, a finding that is not supported by the present study.

The second study to address multisensory speech-in-noise performance was that of Irwin et al. (2011), who tested a younger ASD cohort (5–15 years). In line with the present findings, after carefully controlling for fixations, they uncovered deficits in multisensory phoneme recognition, but they did not explore any potential age relationship and since there were only 13 participants, they would not have had adequate representation in the over 13 brackets to make meaningful distinctions. It is noteworthy that their TD and ASD cohorts showed no differences in performance in the auditory-alone condition. This is consistent with our results where only a very modest 4.2% difference in performance in auditory-alone performance was uncovered. In their study, they also revisited the issue of sensitivity to the McGurk illusion, again finding that ASD individuals showed less susceptibility to illusory fusions. Additionally, they found that ASD individuals were worse at speechreading; although at 89% accuracy in the ASD group relative to 97% accuracy in TD children (due to the limited number of phonemes that constituted their stimulus set), both populations performed at levels of accuracy that bear very little resemblance to speechreading performance under naturalistic speech settings with unconstrained stimulus sets.

More realistic speechreading abilities are represented in the visual-alone condition of the current study where word recognition performance amounted to an average of 10.28% in TD children, but only 4.62% in ASD. In Figure 2C, it is evident that a large number of ASD children were thoroughly unable to accurately report even a single word in the stimulus set (50% scored zero) during pure speechreading, whereas a considerably smaller number of TD children showed this inability (19.7%). As with a number of previous studies, it seems very likely that speechreading deficits contribute substantially to observed deficits in multisensory speech integration. However, follow-up analyses, where we compared ASD and TD children with comparable speechreading abilities, showed that speechreading alone did not explain the differences in audiovisual gain between groups. In this regard, the reader might also note that the older ASD children, who show apparent recovery of their multisensory speech integration abilities, do so despite persistent deficits in speechreading abilities. This older ASD cohort shows equivalent audiovisual performance to neurotypicals (57.6% vs. 60.5%) despite both auditory-alone deficits (34.6% vs. 37.8%) and speechreading deficits (6.2% vs. 13%). Indeed, this latter result would suggest that ASD children in this older age bracket may, in fact, benefit more from multisensory integration than neurotypicals; an intriguing suggestion that will certainly bear replication in future studies.

It has also been suggested that poor speechreading in ASD might point to a unisensory visual deficit, rather than a multisensory origin for the deficits seen in ASD (Williams et al. 2004). Such an account seems entirely too simplistic to us. A key question that arises is whether one could ever truly consider speechreading to be a unisensory function, since it is an ability that arises through cross-sensory learning in all individuals who have intact auditory and visual systems. To posit a purely unisensory account, one would have to propose that ASD individuals are unable to process the basic biological motion information in the visual signal, and that this leaves them to rely heavily, or even completely, on the auditory input. While there is some evidence for deficits in biological motion processing in ASD (Blake et al. 2003; Nackaerts et al. 2012), in our view, the weight of current evidence points to relatively intact basic abilities in this domain (Jones et al. 2011). We will return to the issue of biological motion below. That we find complete recovery of multisensory speech abilities in early adolescence would also seem to argue against a fundamental deficit in visual biological motion processing. The alternative proposal, and one that seems considerably more consistent with the current dataset, is that the deficit lies in delay of the ability to form implicit cross-sensory associations between heard and seen inputs during early development. These associations can only be formed through multiple repeated exposures to speakers over obviously protracted periods of learning. After all, the stimulus set to be learned (i.e. the lexicon) must surely represent one of the greatest learning tasks faced by the developing human. Thus, a simple extent-of-exposure account seems highly plausible. If ASD infants and children are considerably less inclined to fixate appropriately on a speaker (Kikuchi et al. 2011; Noris et al. 2011), then they may simply receive much less cross-sensory learning experience over their early years, and in turn, the crucial cross-sensory correspondences will not be appropriately or deeply encoded. As can be seen for the TD children, there is significant correlation between speechreading abilities and age, so even older children continue to learn to encode these correspondences. Looking at the same function in ASD children, they show a wholly similar rate of improvement in speechreading abilities, but the group as a whole lags the TD group.

**Implications for the Connectivity Theory of Autism**

At the level of underlying neural mechanisms, perhaps the most often implicated pathophysiological substrate in ASD is aberrant connectivity across neuronal circuits at multiple spatial scales (Just et al. 2004, 2007; Courchesne and Pierce 2005; Vissers et al. 2012). Considerable work from both neuroimaging and postmortem studies now points to the

当然，跨模态整合过程必然依赖于这些长距系统的稳定性和完整性，因为这些途径可能在ASD中发生改变。在另一方面，信号的合理恢复在早期 childhood developmental trajectory for long-range structural connections could play a role in the pattern of results observed here. A third possibility cannot be discounted. That is, perhaps the protracted development of AV speech arises because ASD individuals must engage an entirely different and presumably lesser than optimal circuit to integrate seen and heard speech, a circuit that allows them, albeit delayed, to overcome limitations imposed by reduced long-range connectivity. Functional imaging and electrophysiological studies will shed light on this possibility in the future.

**Other Findings Regarding Multisensory Processing in ASD**

Not all work on multisensory integration in ASD has pointed to deficits. For example, in the so-called beep-flash illusion, a single visual stimulus is often seen to flash twice when it is accompanied by a pair of short successive tone bursts, highlighting the fact that highly acute temporal processing in the auditory system can have profound cross-sensory effects on visual perception (Shams et al. 2000, 2002). In a sample of 15 high-functioning adults with ASD, it was found that they were equally susceptible to this illusion, suggesting preserved cross-sensory temporal processing effects (van der Smagt et al. 2007). That multisensory deficits may be constrained to childhood in ASD is also supported by the work of Keane and colleagues in adult ASD participants. In addition to finding no evidence for a lack of sensitivity to McGurk fusions in ASD adults, in the other 2 experiments conducted as part of that study (an audiovisual motion capture paradigm and a variant of the beep-flash illusion), their ASD cohort also performed comparably with control participants, suggesting that at a variety of levels of multisensory processing, from basic combinations to more complex speech inputs, adults with ASD show little to no evidence of a multisensory integration deficit.

One could argue from these results that multisensory processing in ASD may be specifically affected in the speech domain, but this has not been borne out by more recent electrophysiological findings, where deficits in multisensory integration were observed for basic bisensory pairings in ASD children (Russo et al. 2010; Brandwein et al. 2013). Brandwein and colleagues showed that simple response speed facilitation, which is typically seen under audiovisual stimulation (Molholm et al. 2002, 2006), was not observed in older children with ASD. This lack of behavioral facilitation was accompanied by clear electrophysiological evidence for less robust early neural integration of audiovisual inputs. It will be of great future interest to assess whether there is a strong relationship between the extent of deficits in integration of fundamental sensory inputs (i.e. simple tones and flashes) and that in higher-level speech integration in these children.

**The Possible Roots of the Multisensory Speech Deficit in ASD**

What are the roots of this multisensory speech deficit? Some insights may be forthcoming from a study in somewhat younger ASD children than were tested here (4.6–6.1 year olds: n = 16) by Bebko et al. (2006). These authors used the preferential looking method to assess whether ASD children would be as sensitive as their neurotypical peers to large temporal asynchronies between visual and auditory speech streams (3 s offsets). The experimental setup involved 2 flanking video screens. The authors constructed 3 levels of audiovisual stimulation, from a basic setup where a simple plastic ball fell through a plastic maze with the attendant collision and dropping sounds, to a woman simply counting upwards from one at a slow regular pace, and finally, the same female narrating a more complex story. On one screen, a synchronous clip was displayed, while on the other, the asynchronous version was shown. For the basic ball-in-the-maze videos, both ASD and control children showed preferential looking toward the synchronous screen, although it should be noted that this preference, albeit significantly different, is often not very strong (i.e.
in the region of 60–65% of total looking time). In any case, when the children observed the linguistic inputs, the control children maintained a modest bias to look at the synchronous inputs, whereas the ASD sample went to chance levels (i.e., 50–50 viewing). While the authors interpreted these results as evidence for a deficit in the detection of temporal synchrony in ASD, the results of our study might suggest that their results reflect a deficit in speech integration more generally. That their ASD children did detect the temporal synchrony of the nonlinguistic stimuli supports this view. Clearly, an area for further study will be to examine whether the severity of speech integration deficits in ASD are, indeed, related to issues with sensitivity to cross-sensory temporal synchrony. Work by the group of Mark Wallace has shown that the audiovisual temporal integration window appears to be extended in ASD children (Foss-Feig et al. 2010; Kwakye et al. 2011), although this extended window of integration would be an unlikely substrate for the recognition of asynchronies on the order of 3 s.

Turning to even younger children, a fascinating study in 2-year-old toddlers with ASD ($n = 21$) assessed whether there was a specific deficit in orienting to multisensory biological motion signals (Klin et al. 2009). Clearly, the movements of the face during speaking can be considered a canonical class of biological motion stimuli. These authors created a set of point-light animations of an actor performing various children’s games while also recording the corresponding audio track of the actor. They also used a preferential looking paradigm, where these animations were played on one side of a computer screen while an upside-down and temporally reversed version was played on the other side. In this way, the audio track only corresponded with one of the animations, and only one of the animations looked biologically plausible. When typical children looked at these displays ($n = 39$), they were highly biased toward looking at the upright animation, showing a clear preference for real biological motion. The ASD toddlers, on the other hand, showed no such bias, spending equal amounts of time looking at both sides of the screen, and showing essentially no evidence that they were aware at all of the differences between the 2 stimulus classes and no obvious sensitivity to the correspondence between one of the video tracks and the concurrent soundtrack. But there was much more to this story. It was not simply the case that the ASD toddlers showed no sensitivity to multisensory synchrony, rather they were simply following a very different class of multisensory correspondences relative to the TD children. That is, the authors realized that there were other audiovisual synchronies wherein the envelope of the soundtrack might correspond better to motion changes in one or the other visual stream in a way that was unrelated to the biological motion correspondence in the veridical video clip. They went back to their stimulus set and computed the motion trajectories of each of the point lights in both video streams (the real one and the upside-down and backwards one), and they derived a composite motion-to-sound envelope regressor. In this way, they could ask if the seemingly random looking patterns of the ASD toddlers might be related to this more fundamental and purely coincidental audiovisual synchrony. Remarkably, they found that ASD looking patterns were, indeed, highly correlated with this measure of multisensory synchrony. The crux of the findings, therefore, was that the ASD toddlers were integrating multisensory inputs, just not the same ones that the TD kids were. They were following a much more basic set of correspondences, whereas the TD children were highly sensitive to the biological motion aspect of the inputs. The authors went on to confirm this by testing prospectively in a confirmatory cohort of 10 new ASD toddlers.

As pointed out earlier, it bears mentioning that results in children and adults with ASD have been somewhat mixed with regard to biological motion processing, although a growing number of studies have now shown little evidence for deficits (Saygin et al. 2010; Rutherford and Troje 2012), although some studies have pointed to differences (Blake et al. 2003; Cook et al. 2009). The most consistent finding in these studies, however, appears to be that ASD participants, while generally sensitive to basic biological motion, show significant deficits in their abilities to interpret the emotional content of these stimuli (Hubert et al. 2007; Parron et al. 2008). Nonetheless, it is remains plausible that initially severe biological motion processing deficits during infancy and early childhood might ameliorate later during development, in the age ranges that have been much more typically studied. The study of Klin and colleagues highlights the need to conduct studies as early as possible during development and the critical need for better early diagnostic tools to allow for this.

**Study Limitations and Future Directions**

A number of obvious directions exist for further investigation. In the current study, participants were only required to decipher inputs from a single speaker, and we used somewhat artificial pink noise as our main manipulation of background SNR. However, it is in multispeaker situations that multisensory speech integration mechanisms are likely of greatest utility. Under such circumstances, a large component of the task is to figure out and effectively orient to the speaker of interest at any given moment, such that the observer must be able to switch between the relevant information streams quickly and effectively. In turn, speech integration processes must interact with selective spatial attention mechanisms, so that inputs from distracting speakers can be readily suppressed (Senkowski et al. 2008). Certainly, there is emerging evidence for anomalies in how visual space is mapped in ASD (Frey et al. 2013) and in how spatial attention is deployed (Robertson et al. 2013) that will need to be considered in the context of the emergence speech integration abilities in this population.

A study by Alcantara et al. (2004), albeit in a small cohort of ASD adults and adolescents ($N = 11$), assessed the abilities of participants to recover speech-in-noise while they varied the nature of the background noise. They used backgrounds such as a single female speaker, and speech-shaped noise that included temporal dips, spectral dips, or a combination of both. Their results were intriguing in that they suggested that ASD individuals showed deficits only under conditions where there were temporal dips in the noise background, suggesting that the fragmented signals that were “glimpsed” in these gaps were more readily integrated and recognized by neurotypical controls ($N = 9$). Clearly, future work will need to consider the nature of the distracting background noise. In a similar vein, DePape et al. (2012) asked ASD adolescents ($n = 27$) to repeat sentences presented to one ear while they tried to ignore competing sentences presented to the other ear. Their ASD participants ($n = 27$) showed worse performance on this task than neurotypicals, suggesting a specific deficit in suppressing competing speaker input. It is noteworthy that the mean age of participants in this study was 14.8 years, within the age range
where we find only very modest deficits in auditory-alone performance and wholly typical multisensory speech recognition. Thus, once again, the nature of the background noise may be an important factor when assessing speech integration in this population. Of course, in this latter study, there is also the distinct possibility that the observed deficits could be attributed to deficits in the deployment of spatial attention, since the task involved attending to inputs to one ear while suppressing inputs to the other side. Thus, a failure to sustain spatial attention could have driven these effects. As above, the interface between spatial attention and speech integration mechanisms in ASD will require vigorous investigation going forward.

Another obvious avenue to pursue would be to relate structural brain measures (i.e. white matter maturation) to the development of multisensory integrative abilities. A targeted but relatively short-term longitudinal study could assess the critical period around 11–14 years of age when the recovery of multisensory speech processing appears to occur, and assess whether this is related to myelination patterns in specific pathways, and potentially to amelioration of some of the social deficits. Such a study should also include explicit estimates of developmental age (e.g., using Tanner’s methods) rather than relying solely on chronological age as was done here (Tanner 1986).

Another limitation of the current study was that we did not use varying noise levels in the visual-alone speechreading condition in the interests of limiting the duration of the testing period as much as was possible for the young children who participated. It is possible that speechreading performance might have been differentially impacted across group as a function of noise level, although given that the noise manipulation had no differential impact on A-alone performance, we doubt that it would impact V-alone performance any differently. Of course, the fact that V-alone performance was so poor in the first place leaves very little room for any impact of such a manipulation.

In closing, it is worth pointing out that, even in our cohort of TD children, the data clearly point to the fact that a considerable amount of multisensory learning remains to be achieved during the later schooling years, and that explicit efforts to accommodate this learning may well be warranted (Ross et al. 2011).

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Authors’ Contributions

J.J.F., L.A.R., and S.M. conceived the project, analyzed the data, and wrote the paper. D.B., N.R., V.D.B., and L.A.R. collected the data. N.R. performed the clinical phenotyping of the initial cohort of children. D.S.A. and L.A.R. constructed the stimulus set, and H.P.F. conducted and analyzed the eye-tracking component of the experiments. All authors discussed the results, commented on the manuscript, and approved the final version.

Funding

Primary support for this work was provided through a grant from the US NIH (MH085322 to J.J.F. and S.M.). Additional support during protocol development was provided by Cure Autism Now (J.J.F.), The Wallace Research Foundation (J.J.F. and S.M.), Fondation du Quebec de Recherche sur la Societe et la Culture (D.S.A. and J.J.F.), and the Canadian Institute of Health Research (D.S.A. and J.J.F.). The Human Clinical Phenotyping Core, where the children enrolled in this study were recruited and clinically evaluated, is a facility of the Rose F. Kennedy Intellectual and Developmental Disabilities Research Center (RFK-IDDRC), which is funded by a center grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH P30 HD071593).

Notes

The authors would like to express their sincere gratitude to Mrs Alice Brandwein, Ms Sarah Ruberman, and Mr Frantzy Acluche, who provided invaluable assistance during data collection. We thank Dr Juliana Bates for her clinical insights and for performing a large proportion of the clinical phenotyping in our ASD cohort. Conflict of Interest: None declared.

References


