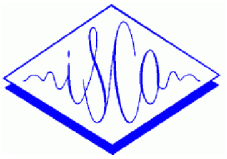
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Audiovisual speech perception in children with autism spectrum disorders and typical controls

Julia R. Irwin[1,2] *and Lawrence Brancazio*[1,2]

[1]Haskins Laboratories, New Haven, CT, USA
[2]Southern Connecticut State University, New Haven, CT, USA

julia.irwin@haskins.yale.edu, brancazio@haskins.yale.edu

Abstract

This paper presents data comparing children with autism spectrum disorders (ASD) to those with typical development (TD) on auditory, visual and audiovisual speech perception. Using eye tracking methodology, we assessed group differences in visual influence on heard speech and pattern of gaze to speaking faces. There were no differences in perception of auditory syllables /ma/ and /na/ in clear listening conditions or in the presence of noise. In addition, there were no differences in perception of a non-speech, non-face control. However, children with ASD were significantly less visually influenced than TD controls in mismatched AV and speech reading conditions, and showed less visual gain (AV speech in the presence of auditory noise). Further, to examine whether differential patterns of gaze may underlie these findings, we examined participant gaze to the speaking faces. The children with ASD looked significantly less to the face of the speaker overall. When children with ASD looked at a speaker's face, they looked less at the mouth of the speaker and more to non-focal areas of the face during the speech reading and AV speech in noise conditions. No group differences were observed for pattern of gaze to non-face, non-speech controls.

Index Terms: audiovisual speech perception, autism spectrum disorders, eye tracking.

1. Introduction

Autism spectrum disorders (ASD) refer to neurodevelopmental disorders along a continuum of severity that are generally characterized by marked deficits in social and communicative functioning[1]. One characteristic feature of individuals with ASD is poor modulation of eye-to-eye gaze with others[2] . This poor modulation of gaze is significant because social, affective and visible articulatory information all reside on the face. Visible speech information influences what typically developing listeners hear and is known to facilitate language processing[3]. Previous literature suggests that children with ASD show reduced influence of visual information on heard speech[4], which could hamper their ability to perceive a speaker's message. However, interpretation of these results is complicated by the tendency of children with ASD to avoid gazing at faces: The reduction in visual influence could reflect a deficit in processing visual speech information or in audiovisual integration, but might simply reflect a failure to view the talker's face, In the present study, we used eye-tracking to examine audiovisual perception in children with ASD and typically developing (TD) controls on trials when we could confirm that the participant fixated on the face. We also compared the gaze patterns on the face for children with ASD compared to TD controls to explore whether differences in attention to specific face regions might contribute to perceptual differences.

2. Method

2.1 **Participants**

Diagnostic criteria for the ASD group: in addition to a clinical diagnosis of an autism spectrum disorder, participants with ASD were assessed with the Autism Diagnostic Observation Schedule Generic[5], a semi-structured standardized assessment of communication, social interaction, and play/imaginative use of materials for individuals suspected of having ASD. Further, caregivers of the children with ASD were interviewed with the Autism Diagnostic Interview-Revised[6]. The ADI-R is a standardized, semi-structured interview for caregivers of individuals with ASD. All participants with ASD met criteria for an autism spectrum disorder on both the ADOS-G and the ADI-R. The TD controls had no history of developmental delays or speech or language problems by parent report.

Experiment 1: Given the paucity of gaze to the face of the speaker in children with ASD, we used eye-tracking methodology to examine responses from a group of children with ASD (13 children, 9 boys, mean age 9.08 years, age range 5-15 years) and age, sex, and verbal mental age matched TD controls (13 children, 9 boys, mean age 9.16, age range 7-12 years) on a set of audiovisual speech perception tasks *when fixated* on the face of the speaker.

Experiment 2: If children with ASD do not have access to the same visible articulatory information as their typically developing (TD) peers because their gaze patterns differ, this may influence their perception of a speaker's message. To examine whether there are differences in pattern of gaze to a speaking face, we compared a subset of 20 children from the previous experiment (10 children with ASD, 8 boys, mean age 10.2, age range 5-15 years, SD 3.1 years and 10 children with TD, 8 boys mean age 9.6 years, age range 7-12 years, SD 2.4 years). The groups were matched on age, sex, and verbal mental age.

2.2 Materials, Experiments 1 and 2: Stimuli consisted of auditory and visual recordings of the consonant-vowel (CV)

syllables /ma/, /na/, and /ga/. A male, monolingual native speaker of American English produced the stimuli in a recording booth.

Visual only (speech reading) stimuli

The visual only stimuli were silent versions of the speaker producing /ma/ and /na/, with a total of 20 trials.

Speech in noise stimuli

Auditory-only and audiovisual stimuli were created by adding noise to the 60 dB /ma/ and /na/ tokens to create a range of signal-to-noise levels at 5, 0, -5, -10, -15 and -20 dB, from less to more noisy. The audiovisual stimuli were the same auditory tokens with video of the speaker producing the same CV syllables. For both auditory and audiovisual stimuli, there were 24 trials.

AV match and mismatch (McGurk) stimuli

The mismatch stimuli were dubbed by placing the audio track such that the point of consonant release at the syllable onset for a new auditory token matched the point of release for the original token, at the resolution of a single video frame, for a total of 12 trials. Mismatched stimuli were always a visual /ga/ token paired with an auditory /ma/. Matched stimuli replaced the audio from tokens of the same CV (e.g., a /ma/ visual token paired with a different auditory /ma/), for a total of 16 trials. For the speech in noise and the AV match-mismatch conditions, participants were instructed to watch and listen to the video display. They were then told that they would hear a man saying some sounds that were not words and to say out loud what they heard.

AV non-speech stimuli

The audiovisual non-speech stimuli consisted of a set of figure-eight shapes that increased and decreased in size, paired with sine-wave tones that varied in frequency and amplitude. These stimuli were modeled on the speaker's productions of /ma/ and /na/ to retain the temporal characteristics of speech, but did not look or sound like speech. To create the visual stimulus, we measured the lip aperture in every video frame of the /ma/ and /na/ syllables. We then used the aperture values to drive the size of the figure: when the lips closed the figure was small, upon consonant release into the vowel the figure expanded (insert Figure 1 about here). The auditory stimuli were created by converting the auditory /ma/ and /na/ syllables into sine-wave analogs, which consist of three or four time-varying sinusoids, following the center-frequency and amplitude pattern of the spectral peaks of an utterance[7]. These sine-wave analogs sound like chirps or tones. Thus, the audiovisual non-speech stimuli retained the temporal dynamics of speech, without looking or sounding like a speaking face (see Figure 1). Stimuli were presented in pairs, such that the two stimuli were either modeled on different tokens of the same syllable (both /ma/ or both /na/) or were modeled on different syllables (one /ma/ and the other /na/, with the order counterbalanced across trials). There were 28 trials. .

Figure 1

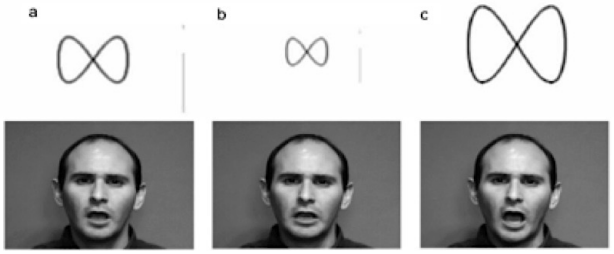


Figure 1. Selected video frames of the non-speech figure driven by lip aperture from a video/na/token.
Note. The images correspond to (a) opening prior to consonantal closure, (b) consonantal closure, and (c) maximum opening for the vowel.

Assessment

Language ability was assessed with the Clinical Evaluation of Language Fundamentals 4th Edition[8]. The CELF-4 provides a core language index (CLI), which quantifies overall language ability. Cognitive ability was assessed using the Differential Ability Scales School Age Cognitive Battery[9] The DAS provides a General Conceptual Ability (GCA) score, which assesses Verbal Ability, Nonverbal Reasoning Ability, and Spatial Ability.

Visual Tracking Methodology

Visual tracking was assessed with an ASL Model 504 pan/tilt remote tracking system. To optimize the accuracy of the pupil coordinates, this model has a magnetic head tracking unit that tracks the position of a small magnetic sensor attached above the left eye of the participant.

2.3 Procedure

After parental consent and child assent were obtained, participants completed the experimental tasks in the eye-tracker laboratory at Haskins Laboratories. Calibration of fixation points in the eye-tracker was completed first. Prior to stimulus presentation, directions appeared on the monitor and were read aloud by a researcher to ensure that the child understood the task. The instructions for each task were as follows: for Visual-only, to report what the man was saying; for Speech in noise and AV match/mismatch, to watch and listen to the video display and report what they heard; for AV non-speech, to judge whether the two shapes opened and closed in the same way or in different ways. Each task began with two practice items with the researcher present to confirm that the child understood and could complete the task. After every five trials, participants saw a video of animated shapes, to maintain attention to the task. Tasks were blocked, with stimuli presented in random order within a block. The inter-stimulus interval for all trials within the blocks was 3 seconds. The blocks were presented in a pseudo-random order; all participants were presented with the auditory-only stimuli first to ensure reliable discrimination between /ma/ and /na/. Audio stimuli were presented at a comfortable listening level (60 dB) from a centrally located speaker under the eye-tracker.

3. Results

3.1 Experiment 1:

Results reported for the AV in noise, visual only (speech reading), and match and mismatch (McGurk) trials include only those trials where the participant was fixated on the face of the speaker within a time window crucial for phonetic judgment with these stimuli: the transition into the consonantal closure, during closure and through to the beginning of the release.

Speech reading condition: There were significantly more trials that had to be dropped for the ASD than the TD group because of lack of fixation on the face of the speaker during consonantal closure, t (24)=2.17, p<05 (ASD: M=8.2, SD=3.8, 41.0% of trials; M=5.7, SD=1.7, 28.0%).

When participants were fixated on the face of the speaker, participants with ASD were significantly less accurate in correctly identifying the visually presented syllable than TD controls, t (24)=2.50, p<.02 (ASD: M=87.9% correct place of articulation, SD=13.3; TD: M=97.6%, SD=3.9), Cohen's d = .98. Notably, the performance for both groups was relatively good, suggesting that there may be even larger differences between the two groups for a more difficult speech reading task.

AV speech in noise: There was a significantly greater number of dropped trials for the ASD than the TD group in the AV speech in noise condition because of lack of fixation on the face of the speaker during consonantal closure, t (24)=-2.15, p<05 (ASD: M=5.4, SD=3.6, 22.5% of trials; TD: M=, 3.2, SD 3.6, 13% of trials).

For auditory-only speech in noise, there were no significant group differences in the percentage of syllables with the place of articulation correctly identified, indicating that both children with ASD and their TD peers were able to identify syllables in the context of auditory noise to a similar degree, *t(24)=*.52, ns, (ASD: M=56.8% correct place of articulation, SD=25.2; TD: M=61.3% correct, SD=18.8). There was also no group difference when all noise levels were included. (Note that the dependent measure was accurate identification of place of articulation, not the actual syllable, so that a /b/ response was scored as correct for /m/. This allowed us to focus our analysis on the extent to which the visual information, which specifies place of articulation (/m/ vs. /n/) but not manner (/m/ vs. /b/), improved perceptual accuracy on its relevant dimension.)

The AV speech in noise condition allows us to measure an increase in identification of the CV syllable in the presence of the face scaled to performance with auditory alone. To remove ceiling effects in the auditory condition, we only included data from the three highest levels of noise **(**-10, -15 and -20 S/N ratio). To increase statistical power, we calculated mean accuracy of place of articulation across the noise levels. We calculated AV gain as the improvement in accuracy from A to AV relative to the maximum possible gain using the formula [(AV-A)/(100-A)]. Importantly, for trials in which children fixated on the face of the speaker, children with ASD showed significantly less visual gain compared to the TD controls. A group comparison revealed a significant difference in visual gain, t (24)= 2.71 p=.01 (ASD: M=57.5%, SD=32.9; TD M=88.9%, SD=25.8), Cohen's d =1.06. This suggests that even when visible articulatory information is available and they are fixated on it, children with ASD do not benefit from this information as much as the TD controls.

AV matched and mismatched

As in the speech in noise and speech reading conditions, significantly more trials were dropped for the ASD than the TD group for lack of fixation on the face of the speaker during consonantal closure for the match-mismatch AV condition, t (24)=-5.88, p<.001 (ASD: M=5.35, SD=2.7, 19.1% of trials; TD: M=.92, SD=.27, 3.2% of trials). For the matched AV syllables, both groups were close to ceiling in percentage of trials with the place of articulation correctly identified, and there is no between-group difference t (24)=1.3, ns (ASD: M=95.3, SD=11.4;TD:M=99.5, SD=1.73). In the mismatched auditory and visual condition (auditory /ma/ and visual /ga/), we compared the groups on percent of visually influenced responses (that is, different from the auditory syllable). Children with ASD were significantly less visually influenced for the mismatched condition, even when fixating on the face, t (24)=2.74, p<.01 (ASD: M=55.7%, SD=33.5; TD: M=87.6%, SD=24.8), Cohen's d = 1.0.

AV non-speech

To compare performance in children with ASD and their TD controls in detecting non-speech cross-modal matching, we employed A', a nonparametric signal detection measure. A "same" response to two AV shapes modeled on the same syllable was coded as a "hit", and a "different" response to two AV shapes, one modeled on /na/, the other on /ma/ was coded as a "correct rejection." The A' measure ranges from 1.0 (perfect performance) to 0 (consistently incorrect) with an A' of .5 corresponding to chance responding The groups did not differ on ability to detect whether the non-speech AV tokens were same or different t (24)=.52, ns (mean A' ASD: .67, SD= .27;TD: M=.72, SD=.19). A comparison of the A' value to .05 (chance) responding indicated significant differences for both groups by comparing the A' value to .5 or chance responding, with t (12)=2.37, p<.03 for the ASD group and t (12)=4.13, p<.001 for the TD group. Thus, the groups did not differ in sensitivity to AV non-speech tasks modeled on the dynamics of speech.

3.2 Experiment 2:

Participant gaze to the speaker's face was examined by group for the AV speech in noise, visual only (speech reading) and non-speech trials. For all analyses, separate analyses were conducted for fixations within different regions at different time samples over the course of the trial. Time was collapsed into 8 ms bins and the analyses were conducted on bins at 96 ms intervals. The dependent variable was the percentage of trials with a fixation in the region within the time bin. The first set of analyses included the two speech tasks and examined whether there were group differences in the percentage of trials with fixations anywhere on the face of the speaker. A series of independent 2 (condition: noisy speech, visual only) x 2 (group: ASD, TD) analyses of variance (ANOVAs) at each time sample indicated a clear pattern of significant differences in percentage of time gazing at the face of the speaker throughout the trial. The children with ASD made significantly fewer fixations on the face of the

speaker than the TD children at nearly every time sample, (F 1, 18 ranged from 8.6-13.5, p<.05). Across time bins and tasks, the mean percentage of bins with a fixation on the face was 63.3% for the ASD group and 82.3% for the TD group. These mean differences reflected large effect size estimates[10].

The second analysis examined whether there were group differences in gaze to specific regions on the face during the speech tasks. A series of independent ANOVAs were run on the percent of trials with fixations in a given region of interest at different time samples. The labeled face regions included: forehead, jaw, cheeks, ears, eyes, mouth, nose, and other face (this included everything that wasn't in one of the other regions, primarily the space between the eye and the ear, between the nose and cheek, beneath the nose and between the eyes). The regions of interest were the mouth and non-focal regions. Non-focal regions on the face were defined as the ear, the cheek, the forehead, and the spaces between the eye and ear, between the nose and cheek, between the eyes, and beneath the nose. The series of 2 (listening condition) x 2 (group) ANOVAs on each region indicated a significant main effect for the mouth, with the children with ASD making fewer fixations to the mouth (across time bins and tasks, M = 25.2% of bins) of the speaker than the TD control group (M = 52.2%) (F 1, 18 ranged from 7.3-20.2, p<.05). They also indicated a significant main effect at the non-focal areas of the face, with the children with ASD (M = 16.8%) making more fixations in those areas than the TD control group (7.2%) (F 1, 18 ranged from 2.8-6.6 p<.05). These group differences in gaze to the mouth indicated large effect sizes, with differences in non-focal areas yielding moderate to large effect sizes[10].

Finally, to assess whether there were group differences in gaze to the non-speech stimuli, a series of independent ANOVAs were run on fixations to the figure eight shapes at different time samples. We defined two regions of interest: a broad region encompassing an area around the outline of the figure eight shape at its largest point, and a narrow region encompassing the area around the outline of the shape at its smallest point. We analyzed percentage of trials with fixations in each region at time samples that incorporated the shape's transition from a small outline to a large one. There were no significant differences between the ASD and TD groups for either region at any of the time samples.

4. Discussion

Even when fixated on the face of the speaker, children with ASD were less visually influenced than TD controls for tasks that involved phonetic processing of visual speech. Children with ASD were significantly weaker at speech reading than TD controls and showed reduced visual influence for the mismatched auditory and visual (McGurk) and AV speech in noise stimuli, where they reported auditory-only percepts significantly more often than the TD controls. Children with ASD exhibited particular difficulty with processing of AV *phonetic* information, including speech reading, AV speech in noise and AV matched and mismatched speech.

The current study also examined pattern of gaze to a speaking face by children with ASD and a set of TD controls, under conditions that create a strong incentive to attend to the speaker's articulations, namely, audiovisual speech with background noise and visual only speech. We found robust differences in the gaze patterns of children with ASD relative to their TD peers, which may impact their ability to obtain visible articulatory information. The findings indicated that children with ASD were significantly less likely to gaze to a speaking face than the child TD controls, which is consistent with diagnostic criteria for this disorder and findings from previous research[2] Critically, the children with ASD were also significantly less likely to gaze at the speaker's mouth than the TD children (note that a previous study of adults indicated gaze to the mouth of a speaker during extended monologues about half of the time, even in the presence of auditory noise[11]). This contrasts with previous findings of increased gaze to the mouth by individuals with ASD[12]. However, this disparity is likely due to the demands of our task; participants were asked to identify what the speaker said in audiovisual stimuli with noise and in visual-only stimuli, creating stronger incentive to look at the mouth. Our results do appear to confirm that individuals with ASD are more likely to gaze to non-focal areas of the face[13]. Importantly, the non-focal areas, including the ears, cheeks, and forehead, carry little, if any, articulatory information. Finally, there were no significant differences by group in pattern of gaze for the non-speech, non-face control condition. This demonstrates that the differences in gaze patterns between children with ASD and TD do not occur for all AV stimuli, and are consistent with the notion that these differences are specific to speaking faces.

The current results suggest that children with ASD do not spontaneously look to critical areas of a speaking face, even in the presence of background noise. This may be particularly problematic, as auditory noise may be especially disruptive for individuals with ASD in speech perception[14]. Therefore, intervention to train individuals with ASD to look at the mouth of the speaker could provide greater access to visible articulatory information, which is crucial for communicative functioning in the natural listening and speaking environment.

5. Conclusions

The children with ASD used audiovisual information less than their typically developing peers. They showed no differences in comparison to TD children in their sensitivity to non-speech (and non-face) AV stimuli. Thus, the current study reveals a potential mechanism that underlies the speech and language difficulties in children with ASD, a deficit in phonetic processing of AV speech.

Children with ASD were more likely to gaze at the non-focal areas of the face, which contain little to no articulatory information. Since the mouth holds much of the articulatory information available on the face, these findings indicate that children with ASD may not have access to this critical speech information. These results may help account for the language and communication difficulties exhibited by children with ASD and may inform us about the significant developmental consequences of atypical gaze to the face of a speaker.

Beginning early in development, young children with ASD likely look less at a speaking face than their typically developing peers. This behavior could lead to weaker AV speech perception,

which may have cascading effects on language development. In this manner, fundamental differences in attention during social interactions may influence the development of language perception and use.

6. Acknowledgements

This work was supported by NIH grants R03 DC-007339, P01 HD-01994, and R21 DC011342.

7. References

[1] American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders. 4th ed., Text Revision. Washington, DC.

[2] Volkmar, F.R, Sparrow, S.S., Rende, R.D. and Cohen D.J. (1989). Facial perception in autism. Journal of Child Psychology and Psychiatry, 30, 591-598.

[3] Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant Behavior and Development, 13*, 343-354.

[4] 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 Disorders*, *38*(7), 1349-1358.

[5] Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., Pickles, A., & Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism & Developmental Disorders, 30* (3), 205-223.

[6] Lord, C., Rutter, M. & LeCouteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*, 659-685.

[7] Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947-949.

[8] Semel, E., Wiig, E., Secord, W. (2003). Clinical evaluation of language fundamentals (4th ed.): Examiner's manual. San Antonio TX: Harcourt Assessment.

[9] Elliot, C.D. (1991). Differential Ability Scales: Introductory and Technical Handbook. San Antonio, TX: The Psychological Corporation.

[10] Cohen, J. (1973). Eta-squared and partial eta squared in fixed factor ANOVA designs. Educational and Psychological Measurement. 33, 107-113.

[11] Vatikiotis-Bateson, Eigsti, I., Yano, S. & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60(6), 926-940.

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

[13] Pelphrey, K. A., Sasson, N.J., Reznick, J.S., Paul, G., Goldman, B.D. and Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders, 32*, 249-261.

[14] Alcantara J.I., Weisblatt, E.J.L, Moore, B.C.J and Bolton, P.F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *The Journal of Child Psychology and Psychiatry*, 45(6), 1107-1114.