Growing Up ‘In Sync’: Connecting a Bridge to an Autistic Mind’s World

Adelina Corina Cozma
Bayview Secondary School, Richmond Hill, Ontario

This study investigates the benefits of using an innovative digital system to screen for auditory-visual integration deficits by determining the temporal integration windows of bimodal speech perception in autistic and typically developing children, and to evaluate whether speech intelligibility and prosodic processing can be improved through audio-video noise reduction in two dimensional (2D) and stereoscopic three dimensional (S3D) display environments. Eleven high-functioning autistic children and eleven age-matched controls (age range: 12-14 years) completed auditory-visual integration tasks incorporated in two innovatively developed software programs presented in 2D and S3D, testing: 1) onset synchrony/asynchrony detection of conversational speech associated with signal lead recognition and 2) speech intelligibility and prosodic processing of unimodal and bimodal emotional and non-emotional stimuli at low-, mid- and high-frequencies.

Accuracy and response times were compared, revealing that autistic children, characterized by wider temporal integration windows, did not integrate stimuli as effectively as controls. Everyone improved bimodal speech perception in S3D. Speech intelligibility and prosodic processing decreased for stimuli presented with noise, especially at low-frequencies. For stimuli without noise at mid-frequencies, all subjects, especially the autistic, exhibited enhanced performances in S3D, for bimodal rather than unimodal inputs, and recognized approach-related emotions more quickly and accurately than withdrawal-related emotions. By personalizing and streaming content in medical, educational and personal environments, the latest digital video technology can improve the lives of typically developing individuals and those with autism and other learning disabilities.
enhances autistic individuals’ ability to integrate audio-visual stimuli. Hence, it is further suggested that such stereoscopic systems be implemented when designing interactive software for individuals with autism in order to improve their overall quality of life.

The experiments presented by the author are elegantly designed, and the data presented holds the potential for significant impact in improving the lives of individuals with autism.

Manuscript edited by Karren Yang and Shama Bhatia

L’intelligibilité de la parole et la prosodie ont diminué pour les stimuli en présence de bruit, en particulier à basses fréquences. Pour les stimuli sans bruit à fréquence moyenne, tous les sujets, et particulièrement les sujets autistes, ont démontré des performances améliorées en S3D, pour des entrées bimodale plutôt que les entrées unimodales, et bimodale dues à un stimulus émotionnel et non émotionnel à basse, moyenne et haute fréquences. La précision et le temps de réponse furent comparés, révélant que les enfants autistes caractérisés par des fenêtres plus larges d’intégration temporelle n’ont pas intégré les stimuli avec autant d’efficacités que les témoins. Une amélioration globale a été constatée dans la perception de la parole bimodale en S3D.

Auditory-visual integration depends on the spatial and temporal relationships of stimuli. Individuals have different temporal alignments of auditory and visual processing, possibly because either one needs to be delayed for the stimuli to be perceived in-sync. Variations in individuals’ auditory-visual synchrony windows correspond to variations in their auditory-visual speech perception. In typically developed adults, auditory-visual asynchronies up to 45ms (audio preceding video) and 200ms (video preceding audio) are not reliably perceived [4]. These findings are not known, however, to apply to either typically developed or autistic children; most research on autism has utilized tasks involving non-word syllables rather than prosodic speech [8]. Moreover, the observations have mostly been based within a 2D digital environment rather than the latest stereoscopic three-dimensional display (S3D) environment. S3D has the potential to improve depth and space perception, thus enhancing the understanding of complex processes such as speech [8].
Recent developments in digital technology have provided researchers with new tools for investigating auditory-visual information processing. This paper focuses on using time-shifting technology to synchronize or delay the auditory-visual input stimuli, and on noise reduction digital technology to filter the noise from perceived stimuli, as new tools to assess auditory-visual integration. These tools will ultimately lead to the development of new strategies to help both TD and autistic individuals with integration deficits, allowing them to recognize basic speech elements and prosodic speech with greater accuracy and intelligibility.

Purpose
This project used an innovative digital system to examine and screen for deficits in auditory-visual integration. The first goal was to analyze and compare the auditory-visual synchrony windows of age-matched TD and high-functioning autistic children, as determined by their performances on various language-related tasks. The second goal was to evaluate whether audio-video noise reduction in two-dimensional and S3D display environments can improve speech intelligibility and prosodic processing.

Materials and Methods

Recording and Processing Speech
The researcher was recorded in 2D and S3D environments by a Fujifilm Finepix Real 3D W3 10.0 MP digital camera. Short, entertaining stories composed of complete sentences were presented in five emotional states – happiness, surprise, neutral, sadness and anger – at a consistent speech rate. The camera captured the researcher’s face and most of her neck, maximizing the visual presentation of her face, in particular the mouth, while eliminating communicative body language. The frequencies of the clips fell within the fundamental frequency range for the respective emotional states: anger (210 to 250 Hz), sadness (145 to 165 Hz), neutral (190 to 210 Hz), surprise (215 to 250 Hz), happiness (200 to 270 Hz) [7].

Stereoscopic 3D Audio-Visual Environment
To implement the S3D digital environment, a monitor with a frequency of 120 Hz – double that of a regular monitor – and a pair of liquid crystal active shutter glasses were used. The left and right eye images are presented in rapid succession on the monitor as the lenses on the glasses alternate between transparent and opaque to simulate 3D perception.

The S3D medical test was performed on all participants prior to experimentation to verify their abilities to see stereoscopic images.

Auditory-Visual Speech Perception Test
The synchronicity of the audio and visual channels in the captured 2D and S3D material was verified using Goldwave 5.5, Audacity 1.3 and Sencore Audio/Video CMA1820 3.4.1.0. The original recorded digital clips were then transcoded and pre-processed by DRC Stream software. Time shifting was enabled to create various out-of-sync slip rates between the audio and video content. Specifically, two out-of-sync versions of each recording (i.e. 6, 8, 10, 12, 14 and 16 frames shifted in either direction) as well as 2 in-sync clips were created, each approximately 10 seconds long. The sentences were presented according to the successive events of the story, but the order of slip rates and directions were randomized. The participants were asked to indicate as quickly as possible whether the clip was in-sync, and if not, whether the audio or the video was leading (Figure 1).

Prosody Perception and Speech Intelligibility Test
The originally recorded digital clips were similarly processed by DRC Stream software as in the auditory-visual speech perception test, except audio-video low/high pass frequency plug-ins were enabled to create audio or visual content with sparse spectral representations at low or high frequencies. The clips with or without filtering were presented in both 2D and S3D environments. Participants were instructed to indicate as quickly as possible which emotion was expressed in each of the audio and/or video clips and to answer some speech intelligibility questions.

Testing Procedure
The audio and/or video clips were embedded in two interactive, animated software programs (Figures 1 and 2). 11 verbal high-functioning autistic children and 11 age-matched typically developing children (age range: 12-14 years) without auditory or visual impairments were evaluated on speech integration, prosody perception and speech intelligibility. Participants were seated in front of a 2D/S3D laptop in which the stimuli were presented in 2D and then in S3D environment. The tests were carefully explained to the participants; they completed tutorials and went through several sample trials to ensure that they completely understood the instructions. Each participant’s accuracy of responses and response...
Figure 1. Auditory-Speech Perception Software

Step 1.
Look and listen to the audio-video clips of the researcher presenting several sentences. Pay close attention to the speech sounds and their corresponding mouth movements. The sentences will be presented either out-of-sync or in-sync, with or without delays between speech sounds and their corresponding mouth movements.

Step 2.
After each sentence presentation, indicate how you perceive the content: in-sync or out-of-sync by clicking the YES or NO button.

Whenever perceiving the content out-of-sync, indicate whether the audio or the video leads the other by clicking the Audio Lead the Video button or the Video Lead the Audio button.

Figure 2. Auditory/Visual Prosody Perception and Speech Intelligibility Software

Step 1.
Look and/or listen to the audio and/or video clips of the researcher. Pay close attention to the emotional content that can be seen and/or heard.

Step 2.
After each audio and/or video presentation, indicate which emotion was expressed.

Step 3.
Indicate which answer is correct for the question displayed. Below is an example.
times were automatically saved. Statistical analysis was performed using hierarchical linear modeling software.

Results

Autistic Children Exhibit Reduced Ability to Recognize Auditory-Visual Asynchronies Compared with Typically Developing Children

For typically developed children, the threshold of recognition of asynchronous stimuli at 100% accuracy was 360ms (2D) and 300ms (S3D) when the audio led the video, and 420ms (2D) and 360ms (S3D) when the video led the audio (Figure 3). In contrast, autistic children did not exhibit a threshold of recognition at 100% accuracy for any slip rate tested (Figure 4).

S3D Digital Environment Greatly Enhances Asynchrony Detection in Autistic Children

While the S3D digital environment enhanced asynchrony detection in both groups, high-functioning autistic children showed more significant improvement than typically developed children for higher slip rates in both directions (Figure 4, $b = 4.70$, SE = 1.51, $t(564) = 3.12, p < .01$). In the 2D environment, typically developed children were better able to recognize slips in both directions (Figure 3, $b = -2.55$, SE = .92, $t(564) = -2.76, p < .01$). For all subjects, accuracy rates were higher in S3D relative to the 2D environment ($b = 5.94$, SE = 1.74, $t(568) = 3.42, p < .01$), and response times were faster in the S3D relative to the 2D environment (graph not shown, $b = -.20$, SE = .05, $t(568) = -4.37, p < .01$).

Autistic Children are More Negatively Impacted by Frequency Filtering than Typically Developed Children

Low frequency auditory and visual stimuli negatively influenced the performances of both groups when

---

**Figure 3.** Asynchrony detection of conversational speech associated with signal lead recognition in 2D and S3D for Typically Developed Participants

**Figure 4.** Asynchrony detection of conversational speech associated with signal lead recognition in 2D and S3D for High-Functioning Autistic Participants
compared to normal and high frequency auditory and visual stimuli. Interestingly, autistic children were significantly slowed by the high or low frequency visual stimuli ($b = -1.9$, $SE = .09$, $t(656) = -2.13$, $p < .05$), which was not the case for the typically developing children (Figures 5 and 6).

**S3D Digital Environment Enhances Speech Intelligibility**

Both typically developing and autistic children exhibited the highest accuracy ($b = 10.00$, $SE = 2.87$, $t(328) = 3.48$, $p < .01$), and the fastest reaction times in speech intelligibility ($b = -1.15$, $SE = .16$, $t(328) = -7.22$, $p < .01$) in the S3D as opposed to the 2D environment (graph not shown).

**Discussion**

Most importantly, the time-shifting technology embedded as an application into the computer-based program proved to be a successful method of screening for auditory-visual integration deficits in the participants. Auditory-visual integration is atypical in children with autism, taking an average of 310 ms instead of 100-200 ms in typically developed children, which may lead to delays and overloads in further processing of stimuli [8]. The results from this study suggest that this impairment can be distinguished based on poor asynchrony detection. Moreover, the observation that both groups exhibited improvements in detecting of lip-sync errors when the trial was conducted in the S3D environment support previous
previous work on how human space and depth perception is influenced by the environment; the S3D digital environment has positive effects on visual processing [9]. The results also suggest that prosodic processing and speech intelligibility decreases when the audio or video stimuli is filtered based on frequency, which is in line with previous reports that people respond differently to emotional stimuli at different frequencies [11].

Conclusions
The research findings led to the development of an innovative digital system that can provide a fast and accurate method to test auditory-visual integration. The system can screen for auditory/visual integration deficits by determining the temporal integration windows of bimodal speech perception, and can improve prosodic processing and speech intelligibility by personalizing content to suit each individual’s specific needs. In this way, the latest digital video technology can improve the lives of typically developing individuals and those with autism and other learning disabilities.

Acknowledgements
I would like to thank the participants for being a part of my experiment, the parents and teachers who have been supportive of my study by giving me their required permissions, and the Autism Ontario Enews – iContact Community who has helped me recruit participants. I would also like to thank the Digital Rapids and Ulestudio companies, who have provided me with their hardware and software products for my research purpose and with whom I have collaborated to develop the software applications used in experiment and the innovative digital system. Last but not least, I want to thank my parents and my mentors from the University of Toronto and York University. Without everyone’s contributions, this project would not have met its goal.

Keywords
autism; auditory-visual integration; technology; education; autisme; intégration auditif-visuel; la technologie; éducation.

References
Review of *Growing Up ‘In Sync’: Connecting a Bridge to an Autistic Mind’s World*

The purpose of the present study was to evaluate the auditory-visual temporal integration abilities of both normally developing (ND) children and children diagnosed with autistic spectrum disorders (ASD). In addition, the impact of stereoscopic 3D visual information on integration abilities was measured. Eleven ASD and eleven age-matched ND children were given two auditory-visual integration tasks. First, the asynchrony between verbal auditory and visual (both 2D and stereoscopic 3D) speech signals was varied and detection thresholds were measured. An individual’s threshold was calculated as the offset required for detection of asynchrony of the visual and auditory stimulus. Second, participants were given an intelligibility test with degraded stimuli spoken with different emotions. The task included identifying the emotion a sentence was read with (5 choices of emotion) as well as understanding the gist of the spoken sentences. The data were described too briefly to properly evaluate but the author asserts that the ASD individuals showed worse integration than the ND participants. In addition, the 3D environment was more helpful to the ASD group than the ND group (who benefitted greatly from just the 2D visual information). The author argues that both groups identified certain emotions better than others and the 3D stimuli were identified faster than the 2D. The conclusion was that speech understanding for both ND and ASD children can be enhanced by digital video technology.

The paper is interesting and well-written overall, the hypotheses are spelled out and the relevance to the field is clear. The topic is of interest to a wide audience. There are a number of ways in which this report could have been improved, however. In particular, a great many details of the project need to be clarified. The design of this study was extremely complicated and some of the critical details were missing or underspecified. This is an extremely ambitious project with a great many variables. A simpler design would have been preferable.

The instructions were not really age appropriate. I ran them through a text analyzer and the grade level it returned was about grade 12. These are mainly 8th graders so the text should be no more difficult than a grade 5 level (adjusting for the expected lower reading level of the ASD individuals). It is unclear whether some of the differences between the groups could be attributed to difficulty in understanding or attending to the instructions.

A number of methodological details are missing or are unclear. For example, in the method section it is stated that the task 2 stimuli were “processed at low and high frequencies”. This is not clear - were they low or high-pass filtered? If so, what frequencies were used? In the abstract it is mentioned that noise was added but the statement in the methods seems to indicate that the stimuli were distorted by filtering instead. In addition there is no description of the emotional stimuli. How did she verify that the target emotion was the one experienced? The number and distribution of the 2 and 3D stimuli is not clear—in the text it merely states that the stories were “presented in both 2D and S3D environments”. Also, the dependent variables were not identified in the method. In the results section the reader is told that response times as well as accuracy were measured. It is not obvious how the response times were taken (if it was a mouse click then was there a difference in familiarity or facility with a computer mouse between the groups? Were the participants told to respond as quickly as possible? Were analyzed RTs based just on correct responses?). The lack of clarity in the methods makes it very difficult to evaluate the impact of the study. In scientific reports it is usually required that the methods be spelled out well enough so that another researcher can replicate the study. The level of detail here makes that impossible.

At the very least some indication of mean performance by condition should be presented to make the results more understandable. In addition, the hierarchical modeling was not described adequately. The analyses for task 2 were even more opaque. In summary, this research report is interesting and the results are provocative. However, given the brevity of the report it is very difficult to effectively evaluate the results and thus impossible to assess its scientific impact. The study was enormously complex and that complexity did not lend itself to the page limitations of this forum.

*Aimee Surprenant, Ph.D., Professor, Psychology, Memorial University of Newfoundland, St. John’s, NL*