Assistance for Autistic People by Segmenting and Highlighting Cross-Modal Perceptual Information

Lars Schillingmann *[1], Matthias Rolf [1], Shinichiro Kumagaya [2], Satsuki Ayaya [2], Yukie Nagai [1]

[1] Osaka University, [2] The University of Tokyo

1. Introduction

People with Autistic Spectrum Disorder (ASD) have to deal with a wide range of perceptual problems. The problems in autism can be related to difficulties in temporal integration. This can affect perception of stimulus length, impaired temporal coherence [1], and problems in orienting attention as well as shifting attention from one stimulus to another [2]. The deficiencies can affect the processing of language [3], processing of visual information such as faces or visual signs of emotion [4], or acoustic signs of emotion [5].

However, multimodal input can still facilitate autistic persons' communication. For example sign language and speech have been reported to mutually shed lights on relevant parts in the conversation [6]. Such additional multimodal cues can help to overcome integration and attention problems by (*i*) *augmenting* available information and (*ii*) *highlighting* parts of other modalities that are important. Likewise, augmenting and highlighting interactions patterns have been discovered in parents' child-directed communication, which have gained recent interest also for the teaching of robots [7].

Yet, existing assistive systems act as a filter that irrevocably removes information from the perceptual stream. Such filters certainly prevent being overwhelmed by too much information, but do not allow for a mutual promotion of multimodal stimuli, and might also remove important information accidentally. We argue that support systems for people with ASD should instead augment available information and highlight it, in order to guide their attention and to help extracting relevant information (see Figure 1). In this paper we approach the development of a support system that allows for such mutual promotion by capitalizing on previous effort on infant directed communication and robotic perceptual systems. We presented two robotic perceptual systems to people with ASD and analyzed their responses based on a questionnaire and a free discussion. Results agree with our hypothesis by identifying the need for an immersive assistive device that integrates into the communicative situation instead of separating the user from it by removing information.

2. Method

The general scope of this paper is to investigate the use of robotic perceptual systems for the integration of auditory and visual stimuli as assistive systems.

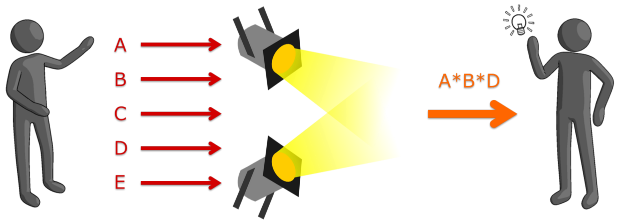

Fig.1 Conceptual diagram of a perceptual assisting system which highlights information instead of restricting the available information.

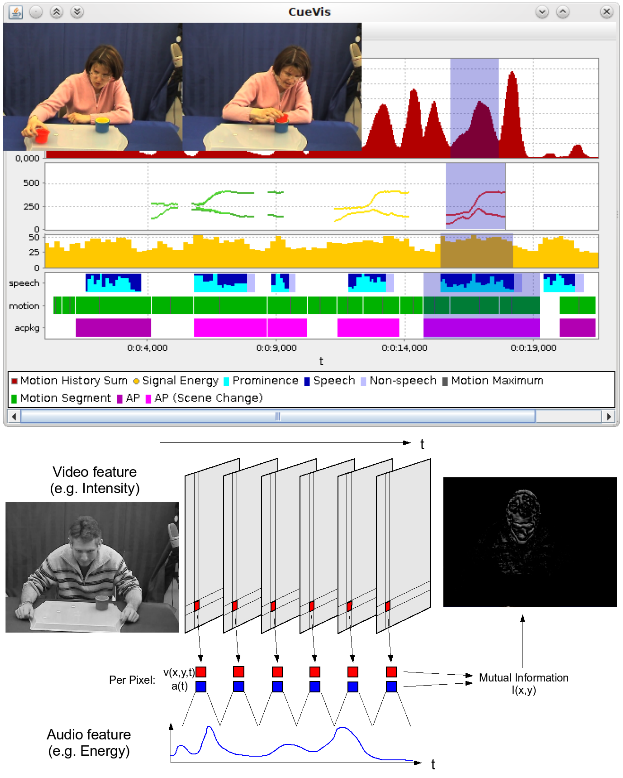

Fig.2 Demonstrated systems: (top) Acoustic packaging; (bottom): Audio-visual synchrony.

In the following we give a short introduction to these systems and elaborate on our experimental session including an interactive session, in which autistic people interacted with the systems, a questionnaire and a open discussion.

System 1: Acoustic Packaging The first system uses parallel audio and video streams to segment events in both modalities and associate them. Therefore it generates "packages" comprising speech from the audio signal and movements or gestures visible in the video stream [8]. Audio-band energy and motion-history image norms are considered as measures of activity within each modality in order to detect and segment events in the stream. The degree of overlap between events in both modalities is then used to group them, whereas one package always has exactly one acoustic event to which multiple visual events can be associated (see Fig. 2, top). A prominence detection module [9] is finally used to find syllables within packages with the strongest vocal emphasis, which typically reflects the semantically most significant parts of human utterances. Hence, the system is able to discover segments of emphasized speech and associated visual events like gestures. This information is used for to assistance strategies: Firstly, the ongoing segmentation of events and their association is displayed on a computer display. Secondly, the system repeats the most prominent syllables acoustically via loudspeakers which can act as summary of the most important information in a conversation.

System 2: Audio-Visual Synchrony The second system detects short-term synchrony between incoming auditory and visual streams at signal level. Thereby the energy (e.g. loudness) of the acoustic signal is temporally correlated with the change of the video data on a per-pixel basis [10]. Since each pixel in a camera image is correlated individually to the acoustic signal, the result is another image stream indicating how synchronous different areas of the camera image are with the current acoustic sensation (see Fig. 2, bottom). This synchrony is naturally sensitive to sound sources: e.g. a mouth produces temporally synchronous visual and acoustic stimuli during speech. However, also non-sound-sources can be synchronous on purpose in a communicative situation. For instance gesturing synchronously to certain words or expressions is frequently used to highlight them [10], which is similar to the mutual promotion of stimuli during sign language as discussed in the introduction. The feedback strategy used for this system was to integrate the synchrony information back into the original stream of images: areas of the camera image that expose a high degree of synchrony with the audio stream are displayed with slightly increased saturation of colors. Areas with little or no synchrony to the audio stream were still visible, but were displayed with reduced color saturation and slightly spatially smoothed. Therefore, synchronous stimuli create a pop-out effect that easily allows to focus on them. The display of this information was done fully online with an only minimal time-lag.

ID	Question asked (translated)
Q1	Can you accept wearing glasses that display an image?
Q2	Do you believe you would be distracted by glasses that display an image?
Q3	Can you accept a side screen when you talk to someone?
Q4	Do you want to use the system that you experienced today in daily life?
Q5	If you answered yes to the previous question (wanting to use system in daily life), which system do you want to use?
Q6	Comments for improvement of the system experienced today
Q7	Would you accept using a system which communicates information using vibration?
Q8	Which modality do you prefer?
Q9	Do you think it is convenient if there is a tool which can replay important words during conversation?
Q10	Do you think you will be distracted if such word will be replayed during the conversation?
Q11	Do you think that it is useful to have a sound that will emphasize the important information during the conversation?
Q12	Do you think you will be distracted when such a sound is played during the conversation?
Q13	Problems in your daily life?
Q14	What kind of tools would be useful?

Table 1 Questions asked

Procedure We conducted an interactive experimental session with 24 Japanese participants with disorders of various intensity within the autistic spectrum. In the first 45 minutes an introduction was given to them explaining the purpose of the session, and giving a short introduction to the aforementioned systems that were also shortly demonstrated. After that the participants had another 45 minutes time to freely interact with both systems and to test the acoustic package detection and prominent syllable playback, as well as the highlighting based on audio-visual signal-level synchrony. During this interaction session the participants could (*i*) test how the systems react to themselves, as well as (*ii*) how the systems augments and highlights signals from other persons in order to check whether the systems' feedback is useful to interpret and understand them.

After the interaction session we asked participants to fill a questionnaire with the questions shown in Table 1. The main purpose was to find out their attitude towards various ways of technological embedding (e.g. glasses, screens) of assistive systems, as well as towards specific feedback strategies demonstrated in the interaction session (e.g. playback). Finally, each participant was asked to give comments on his experience with the systems as well as general comments not covered by the questionnaire. These comments were collected within a 45 minutes open discussion in which all participants could hear the others' comments.

3. Survey Results and Discussion

The systems were exposed to 24 participants which subsequently filled a questionnaire asking them

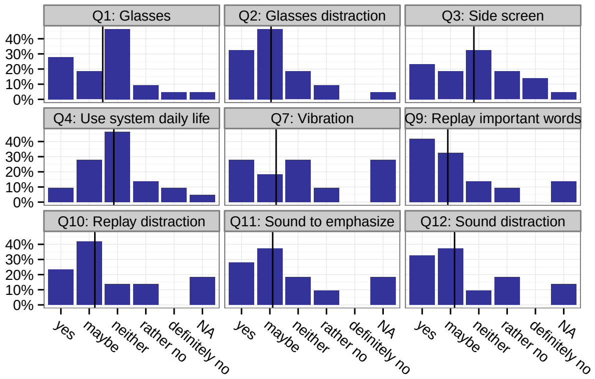

Fig.3 Survey results. Thin black bars indicate the mean response.

Frequency	Comment Category
2	Accuracy was low
2	System might be expensive
2	System output should be adaptable
2	Multi-talker problem
3	Contrast in display was low
6	System for mobile / iphone use

Table 2 Q6: Comments for improving the system?

Frequency	Comment Category
2	Problems with attention
2	Tired fast
3	Emotional perception
3	Problem getting semantics
4	Multi-talker problem
5	Organizational problem

Table 3 Q13: Problems in your daily life?

Frequency	Comment Category
2	Escape the current situation
2	Mobile device
2	Multi-talker problem
3	Emotion perception

Table 4 Q14: What kind of tools would be useful?

about their experiences with the system, improvements, suggestions and their own problems in daily live. In the following the responses of the participants to the questionnaire are discussed. Firstly, we will address questions related to each potential device and its possible risk for distracting the user. On the topic of using glasses which include a screen displaying additional information (Questions Q1 and Q2, Figure 3) people responded mainly positively. They either find it possible to accept wearing glasses (41.7%) or do not have a strong opinion whether to accept or reject glasses (41.7%). However, they also see a relatively high risk of being distracted by them. A group of 70.8% believes they will likely be distracted by such glasses. Concerning a side screen which displays additional information, the opinions are more diversified (Question Q3, Figure 3) which is reflected in a higher standard deviation of responses ($sd = 1.3$) compared to the previous question ($sd = 1.1$). Using a vibration device (Q7) is also acceptable for 41.7% of the participants. However, 25% do not answered this questions, which might be due to lack of personal experience with such devices. Concerning the acoustic modality 66.7% of the participants support the idea of a device which replays important words during the conversation (Q9), which is a strong support for this approach compared to the previous responses. However, also 58.3% believe they might be distracted (Q10), which is lower compared to Q1 but still significant. The same result 58.3% is observed for the question about a sound that emphasizes relevant information (Q11). Here, 62.5% assume they will be distracted (Q12), which is consistent with Q11. Concerning the two *particular* systems that were presented there is a slight favor of the acoustic packaging system (29.2%) which emphasized the important word in the conversation over the audio-visual synchrony system (20.8%) which emphasizes regions in the visual stream (Q5). However, 50% of the participants do not have a clear opinion here, which indicates a low significance of these answers. In particular, participants had an opinion inconsistent to Q5 in Q8, asking for the *general* preference for a modality in which feedback should be provided. A group of 58.3% prefers feedback as visual input, while only 20.8 prefer tactile input and a minority of 4.2% seems to favor acoustic input.

An important pattern in these results is the strong contrast between the acceptance of a device which assists in processing perceptual information and the concern that such a system poses an additional distraction factor. This pattern can be consistently observed for both the acoustic and visual modality. This result supports our hypothesis that any assistive device is required to be immersive and should not diminish any positive effects by causing distraction to the user.

To better understand the specific needs of persons with ASD concerning an assistive system the open questions in the survey were classified into categories and the frequencies of these categories were counted (see Table 2, 3, and 4). Frequencies below 2 occurrences have been removed to focus on frequent aspects. A very strong demand seems to be the portability of an assistive device (see Table 2). Also due to organizational problems, having to carry around multiple items does not seem to be feasible for people with ASD. Furthermore, the results reflect typical problem domains in ASD: linguistic processing in noisy conditions and situations with multiple talking people are mentioned multiple times (see "Multi-Talker" problem in all tables). Also problems with emotion perception are shown (see Table 3 and 4), as well as the wish for support for that issue.

In summary the results exhibit a strong wish for a system that provides semantic information about ongoing events. Especially on topics where people with ASD typically have problems such as speech perception and emotion perception. This result is consistent with the positive responses to question 9 and 11 that refer to a first semantic interpretation of the input. Other comments which are not reflected in the above tables due to lower frequencies indicate not only a wish for interpreting the input, but also for systems capable of improving the own communicative output. Participants suggested that they would like to evaluate their own communication for training purposes. Another request described a the system that assists people with ASD in communicating their inner state to the communication partner.

4. Conclusion

We presented two robotic perceptual systems to people with ASD and analyzed their responses based on a survey and a free discussion. The results show that people with ASD would accept a system which augments input. Additionally, the results clearly confirm our hypothesis, that an assistive system should be immersive and not distract the user. This result is especially supported by the contrast between acceptance for such devices which coincides with a concern regarding distraction it may cause. Therefore, a key aspect in developing on assisitive systems for people with ASD is that they do create additional distraction for the user. Furthermore comments indicate that users would not like to be separated from their environment by the system. Another result of our analysis is that the participants strongly demand support for semantic interpretation. This means from the user viewpoint an ideal assisitive device should not just enhance information, as for example a directional microphone enhances the signal to noise ratio to make a specific speech source more understandable. Instead interpreted information such as subtitles for speech, digests of past talks, face recognition, emotion recognition is requested. Another requirement seems the systems adaptability to users specific needs. Different ASD characteristics might not only require different modalities to be enhanced but also to inject enhancements into a specific output modality as well as specific fine tuning. Furthermore, an assisitive system should not rely only on processing input and presenting information to the user. Some users also expressed their wish in support for expressing their inner state such as feelings to the communication partner.

In summary the results show that developing an assisitive system for ASD requires careful consideration of the target users and their specific characteristics of ASD. Our results on the topic of distraction show that even negative effects might be created if the device is not immersively integrated into the users environment.

5. Acknowledgements

This work has been supported in part by MEXT KAKENHI "Constructive Developmental Science" (24119003).

References

[1] J. Brock, C. C. Brown, J. Boucher, and G. Rippon, "The temporal binding deficit hypothesis of autism.," *Development and psychopathology*, vol. 14, pp. 209–24, Jan. 2002.

[2] R. A. Coulter, "Understanding the Visual Symptoms of Individuals with Autism Spectrum Disorder (ASD)," *Optometry & Vision Development*, vol. 40, no. 3, p. 164, 2009.

[3] J. Ricketts, C. R. G. Jones, F. Happé, and T. Charman, "Reading comprehension in autism spectrum disorders: the role of oral language and social functioning.," *Journal of autism and developmental disorders*, vol. 43, pp. 807–16, Apr. 2013.

[4] J. C. McPartland, S. J. Webb, B. Keehn, and G. Dawson, "Patterns of visual attention to faces and objects in autism spectrum disorder.," *Journal of autism and developmental disorders*, vol. 41, pp. 148–57, Feb. 2011.

[5] P. F. Heaton, L. Reichenbacher, D. Sauter, R. Allen, S. K. Scott, and E. L. Hill, "Measuring the effects of alexithymia on perception of emotional vocalisations in Autistic Spectrum Disorder and typical development," Feb. 2012.

[6] Ayaya and Kumagaya, "Hattatsu Syougai Toujisha Kenkyu (in Japanese)," in *Igaku Shoin*, 2008.

[7] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?," *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.

[8] L. Schillingmann, B. Wrede, and K. J. Rohlfing, "A Computational Model of Acoustic Packaging," *IEEE Transactions on Autonomous Mental Development*, vol. 1, pp. 226–237, Dec. 2009.

[9] L. Schillingmann, P. Wagner, C. Munier, B. Wrede, and K. Rohlfing, "Using Prominence Detection to Generate Acoustic Feedback in Tutoring Scenarios," in *Interspeech 2011*, Aug. 2011.

[10] M. Rolf, M. Hanheide, and K. Rohlfing, "Attention via Synchrony: Making Use of Multimodal Cues in Social Learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, pp. 55–67, Apr. 2009.