

## Chapter 1

# **The Emotional Hearing Aid: An Assistive tool for Children with Asperger's Syndrome**

---

R. El Kaliouby and P. Robinson

### 1.1 Introduction

Children diagnosed with autism, and its milder cousin- Asperger's Syndrome- often have difficulties operating in the highly complex social environment in which we live and are, for the most part, unable to read or understand other people's emotions (e.g. Baron-Cohen, 1995; O'Connell, 1998). Consequently, they need to be taught explicitly how to read other people's minds from nonverbal communication channels such as the face. This paper reports work in progress on the emotional hearing aid, a portable assistive computer designed to help children with Asperger's Syndrome read and react to facial expressions of the people they interact with.

The paper starts with a survey of key therapeutic technologies used for autism, and draws attention to the lack of assistive tools for this disorder. Motivated by the need for this type of technology, the emotional hearing aid draws inspiration from the "emotional indexing" method, an approach for teaching children with autism how to read and respond to emotions. This teaching method is introduced in section 1.1.2. An overview of the tool, including typical use-case scenarios are discussed in Section 1.2, while section 1.3 details the architecture and design of each of the modules. Section 1.4 presents results obtained so far, before section 1.5 concludes this chapter.

#### 1.1.1 Therapeutic Tools for Autism

An increasing number of studies show that computer-aided learning and therapy are well accepted by individuals with Autistic Spectrum Disorders (Moore *et al.*, 2000). Consequently, computing technology is increasingly being used in

therapeutic contexts of autism. *Mind reading* (Baron-Cohen and Tead, 2003) is an interactive guide to learning about emotions, which provides children with a library of over 400 videos and games to test their progress on reading those emotions. *Kidtalk* (Cheng *et al.*, 2003) is a therapist-moderated online chatting environment, where children work through common social situations, such as going to the movies by chatting online. The virtual sand box (Hirose, 1997) and the environment developed by Strickland (1996) enables children to interact in a virtual setting modelled around real-life social scenarios. The AURORA project (Werry *et al.*, 2001; Dautenhahn and Billard, 2002) utilizes an autonomous robot as an interactive toy that can engage children in a therapeutically relevant environment. The toy is meant to encourage pro-active social behaviours towards the robot, elicit robot-child eye contact, and teach the child the basics of turn-taking and interaction games. Different embodiments of this toy have been investigated including a doll, and a four-wheeled vehicle-like toy. The affective social quotient project (Blocher, 1999) consists of short digital videos that embody one of several basic emotions and a set of physical "dolls" linked by infrared to the system. The system knows which dolls correspond to which clips, so that the child can explore emotional situations by picking up dolls with certain emotions, or the system can prompt the child to pick up dolls that go with certain clips. Finally, Kozima and Yano (2001) investigate the possibilities of using humanoid robots in therapy.

Those technologies are mostly remedial tools aimed at providing a learning environment to teach children the fundamentals of social behaviour. They do not provide assistance to individuals with autism beyond that gained through teaching. In addition, as they do not operate in a natural human-human interaction environment, they risk failing to generalize (Howlin *et al.*, 1999).

In contrast to existing work, our proposed portable assistive device is designed to assist people diagnosed with autism in real life situations. In a sense, our tool is analogous to a hearing aid, which allows people with hearing problems to communicate with the rest of the world.

### 1.1.2 Emotional Indexing in Autism

A number of approaches to teaching emotion understanding to children with autism exist. The methods may differ in the amount of structure involved (highly-structured methods use carefully planned teaching material deployed in a relatively controlled environment), the setting in which the teaching takes place (ranges from being hypothetical to being natural), and whether it is interactive or not. One approach, especially suitable for use with children, involves emotional indexing of the child's surrounding environment (Fling, 2000). Typically the child's carer indexes the emotional content of situations as they arise, and suggests possible actions that can be taken by the child. For example "Oh, Mary got hurt. She is crying. Can you tell Mary, 'I am sorry'?" This approach to teaching emotions has been shown to improve the social competence of some children (Fling, 2000; Howlin *et al.*, 1999). In contrary to most other teaching approaches, social

indexing works in the child's natural interaction environment reinforcing appropriate social behaviour in a spontaneous setting.

Unfortunately, this method is not always available for the child, as it requires the physical presence of the carer, which in some cases (e.g. school) might be impractical. Also, unlike highly structured approaches, with this method it is almost impossible to recreate events once they have occurred.

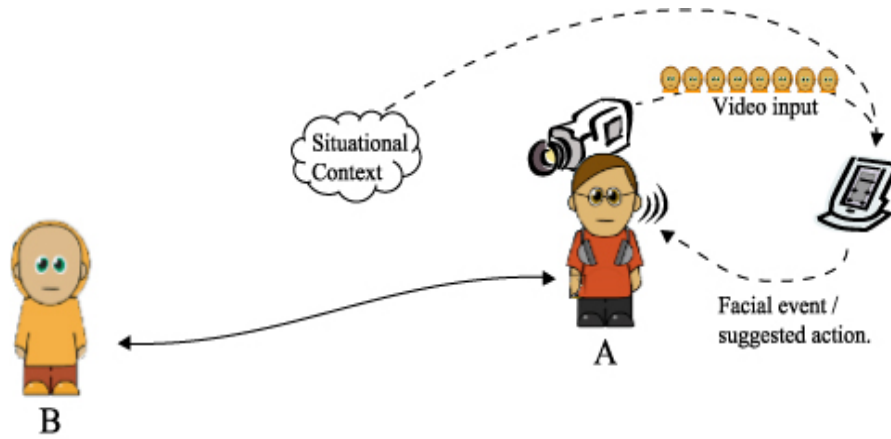
## 1.2 Overview

The emotional hearing aid aims at providing real time assistance with reading facial expressions of other people, and reacting to it in a child's natural social environment. It is designed as a portable assistive device, which consists of a digital camcorder, a personal digital assistant (PDA), and an earpiece speaker.

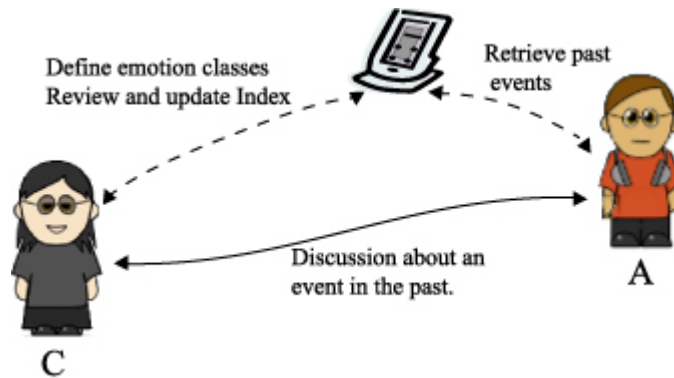
Figure 1.1 illustrates how the emotional hearing aid provides assistance in a typical interaction scenario between a child with Asperger's Syndrome (character A in the figure) and another person (shown as B). Video sequences of B are sent to the PDA. The PDA is responsible for analysing the incoming video, and any available context cues for mental state information. It also indexes this event for further retrieval, and uses it, along with a repertoire of situations to suggest a course of action. This advice is sent back to the wearer in real time, but continuous feedback is avoided to minimise the number of distractions. Also depending on the level of engagement, the output can be visual or audio, and varies in the degree of detail presented.

Humans make considerable use of the contexts in which expressions occur to assist interpretation (Bruce and Young, 1998; Edwards, 1998), including situational context. Howlin *et al.* (1999) define situation-based emotions as those that involve inferring a person's emotional state from a particular sequence of events. We thus define several profiles (very much like those used on mobile phones) to indicate the various situations the child is in. Profile information is used along with the input video to boost the reliability of the inference process. Currently, only simple profiles, such as "in school", or "in playground" are supported, and are explicitly selected. As the tool gets more sophisticated, more detailed profile information would be deduced automatically.

The emotional hearing aid is also designed to work in tandem with the child's carer as shown in Figure 1.2. The carer can define the emotional states the tool can address, can review and update the archived events, and can engage in discussions about past events with the child.



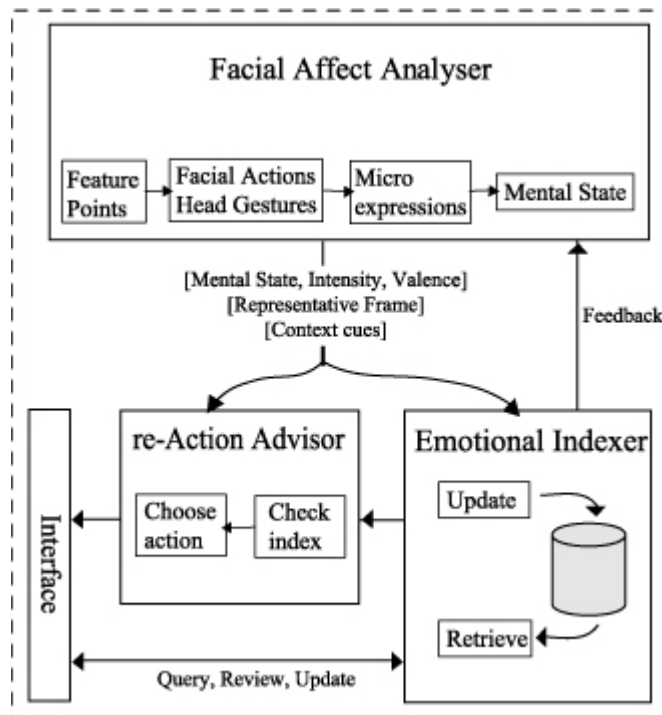
**Figure 1.1.** Child A (diagnosed with Asperger’s Syndrome) is using the emotional hearing aid in an interaction with person B. Video sequences of B and situational context cues are sent to the PDA for analysis, and suggested reactions. Depending on the mode of interaction, the output can be visual or audio, and can vary in the degree of detail presented.



**Figure 1.2.** Another interaction scenario where A (the child), and C (the carer) can query the index for past events. In addition, C is able to define the emotion classes and update the index through the interface.

### 1.3 Architecture and Detailed Design

The main modules of the emotional hearing aid (shown in Figure 1.3) parallel that of the child’s carer in emotional indexing. The facial affect analyser identifies a facial event in real time, picks representative frames of that expression and annotates it with an emotional label. The re-action advisor appraises the current situation, and suggests appropriate courses of actions to take when a similar one arises. In addition, an emotional indexer module, responsible for keeping an



**Figure 1.3.** The main modules defining the emotional hearing aid. For every incoming video frame, the facial affect analyser identifies facial feature points, and uses their displacements to infer facial actions and head gestures. Those displays are then combined temporally to form micro expressions, a sequence of which portrays a mental state. A representative frame of the mental state, its label, intensity, valence (whether is a positive or negative state), and accompanying context cues are input to the emotional indexer and re-action advisor. The former appends that event to an index, while the latter utilises the information to suggest an appropriate re-action. The interface manages the communication between the user and the other modules, namely the action advisor and emotional indexer.

archive of past events, is integrated into the design. In doing so, we extend the emotional indexing approach to allow events to be re-played. Finally, an interface layer is responsible for managing the communication between the user and the rest of the modules.

### 1.3.1 The Facial Affect Analyser

For every frame extracted from the incoming video sequence, twenty-two points are identified on the face using prior knowledge of face shapes. The points include pose estimation ones (such as nose tips, nose root, and nostrils), and feature points used for emotion classification (such as inner and outer eyebrow, upper and lower lips, and pupils). Feature point displacements are calculated between consecutive

frames. In addition, two head translation parameters indicate the head's movement and three rotation ones are calculated to indicate the head's rotation along three axes: horizontal, vertical and planar.

Feature point displacements and head motion parameters (calculated every hundredth of a second) are tracked over time to identify head and facial gestures. Those in turn typically last between one to two seconds, emphasizing the fact that facial expressions are inherently dynamic processes. Head gestures in particular follow some pattern of temporal regularity (Davis and Vaks, 2001). For instance, a head nod is a series of vertical up and down movement of the head used to show agreement, excitement, or comprehension while listening. Head gestures vary in duration as well as in intensity. Such variations often signify different user intents. For example, a quick and strong nod tends to indicate more agreement than a weaker and slower one. In El Kaliouby and Robinson (2003), we define a number of parameters that are used to estimate the strength of a head gesture, and present a state machine that processes accumulated head motions (such as move-up or move-down) to recognize a number of head gestures in real time.

Upper (eyes and eyebrows) and lower (mouth) facial action analysis is performed simultaneously with head gesture recognition. We extend the approach used in Tian *et al.* (2000), which combines both shape (geometric) and colour information to determine the different states of the mouth. As shown in Figure 1.4, the mouth area is represented by eight feature points. In addition, an origin is defined as the intersection between the lines joining the two lip corners, and the two central ones. The distance between the two corners is used to determine the width of the mouth, while the distance from the upper centre point to the origin, and lower centre point to origin depicts two height parameters. The angle of orientation of the mouth is also calculated. A total of four hyperbolic arcs approximate the lip contours. The two arcs representing the outer contours pass through the upper and lower mouth feature points, whereas the inner two require further colour information to approximate. Colour information is mainly used to determine if teeth are present or not (aperture) indicated by the luminance of the pixels. Once shape and colour information are calculated, mouth states (open, open with teeth, closed, tightly closed) are determined. Mouth states are tracked over time to infer the underlying lower facial action. For example, if the mouth progresses from being closed, to being open, with lip corners pulled outward, and teeth present, then this would indicate a smile.

A similar approach is used in determining the states of the eye (open, shut, wide open, squinting). With regards to the eyebrows, the displacements of the feature points are tracked over time to find out if the eyebrows have been raised, and if the two inner points have moved closer (as in a frown).

Finally, concurrent facial and head gestures combine to form a micro-expression, a sequence of which constitutes an emotion (Edwards, 1998). This typically lasts between six to eight seconds. The analyser also extracts a number of parameters on the mental state level such as intensity and valence (positive or negative), and a representative clip of that mental state. This is all then sent to the emotional indexer and the re-action advisor.



**Figure 1.4.** The facial affect analyser extracts facial feature points on the face, and uses that along with colour information to determine facial actions. A close up of the mouth feature points, the inner and outer lip contours, and additional colour information (teeth, aperture) is shown on the right. The pose estimation points (e.g. nose tip and nose root) are tracked to recognise head gestures (Video courtesy of the Autism Research Centre, Cambridge).

### 1.3.2 The Emotional Indexer

The primary function of the emotional indexer is to archive facial events as they occur. The archive is accessible to the facial affect analyser and re-action advisor modules to improve inference results. The archive is also made available (through the interface layer) to the child and carer for discussion, learning and reviewing purposes. The latter is particularly important if the stored events are going to affect future inference results. Every event is stored as a tuple in the index containing the representative frames of the facial expression, a label, any additional parameters and context cues available.

### 1.3.3 The re-Action Advisor

This module's primary function is to appraise the current facial input, making use of profile information and archived events, in order to suggest a number of possible courses of action to take. We identify several issues that are fundamental to the design of this module. The most obvious one is associated with the timing of a suggestion: when and how often should a reaction be suggested, and how soon after a change of emotional state is detected should an action be suggested. Needless to say, while the analyser module emits an inference every 6 seconds on average, it would only cognitively overload (and indeed frustrate!) the child if an action is suggested with every inference.

Deciding on the function of a suggested reaction is also important. Reactions to facial expression can be for the purposes of feedback, empathy (Surakka and Hietanen, 1998), or communicative (e.g. signal turn-taking).

Finally, determining the intensity of a reaction also needs to be considered. Intensity is a function of the facial event, its valence and intensity, the current situation and the degree of approachability of the other person.

### 1.3.4 The Interface Layer

The interface layer manages communication between the child (or carer) and the other modules of the emotional hearing aid. To start with, the interface informs the wearer whenever a facial event occurs, and returns the possible courses of actions suggested by the advisor module. The interface also decides on the modality and format of the output depending on the active profile and level of engagement of the child with the current social scenario. A summary-mode is adopted when the user is actively taking part in an interaction, and only needs assistance with the suggested course of action. In this case, the output is unobtrusive to avoid interrupting the interaction, and can be visual (coded-coded characters) or audio (ambient sounds). In the detailed output mode, all available information pertaining to the event is presented to the user visually.

In query mode, the interface allows the index to be queried by both the child and carer using a number of different parameters such as emotion labels, intensity, valence, and date of event. In the update mode, the carer is able to retrieve event entries, and update them. Finally, the carer can also define the emotion classes the analyser deals with, activating and de-activating classes as needed.

## 1.4 Preliminary Results

The prototype currently in place, utilizes a commercial digital camcorder connected to a standard PC. Our system operates in real time (30 fps) and does not impose a frontal position on the user. In contrast to other existing systems, which operate on short clips (1-3 seconds), we place no constraints on the duration of the input because the facial affect analyser keeps a rotating buffer of the input.

Preliminary tests were carried out on subjects from within the Computer Laboratory at the University of Cambridge, asking them to act out basic emotions (e.g. happy, disgusted) and various head gestures (e.g. head nods, head shakes). In addition, further tests were carried out using videos from the *mind reading* DVD (Baron-Cohen and Tead, 2003). Videos and emotional labels from that database were verified by a panel of judges to ensure that they correctly match. So far, we are able to distinguish between a number of prominent head gestures: head nod, headshakes, and head tilts, and facial actions: smile, mouth open, mouth closed, eyebrow raise, and frown. Table 1.1 lists the mental states that our system supports so far, and the characteristics of each state exploited by the analyser.

Finally, the only action advice supported so far is in the form of real time feedback, acknowledging a facial event, its class and intensity. The feedback also provides information as to when the emotional state started and ended.

**Table 1.1.** List of mental states supported, their characteristic head gesture and facial actions

<b>Mental State</b>	<b>Head Gesture</b>	<b>Lower actions</b>	<b>Upper actions</b>
<b>Agreeing</b>	Nod	<i>(Nothing specific)</i>	<i>(Nothing specific)</i>
<b>Comprehending</b>	Nod	<i>(Nothing specific)</i>	Drawn inwards
<b>Confused</b>	Tilt in one direction	Tightly closed	Frown
<b>Convinced</b>	Medium nod	Subtle smile	Eyebrow raise
<b>Decided</b>	Strong, quick nod	<i>(Nothing specific)</i>	<i>(Nothing specific)</i>
<b>Delighted</b>	<i>(Nothing specific)</i>	Smile	Eyebrow raise
<b>Disagreeing</b>	Headshake	<i>(Nothing specific)</i>	Frown
<b>Disbelieving</b>	Quick headshake	Open	Drawn inwards
<b>Grieving</b>	Slow headshake	Tightly closed	<i>(Nothing specific)</i>
<b>Undecided</b>	Alternate head tilts	Pulled sideways	Squint

## 1.5 Conclusion

This paper reports work in progress on the emotional hearing aid, a portable assistive computer intended to help children diagnosed with Asperger's Syndrome read, understand and react to facial expressions in a socially-appropriate way. The architecture and detailed design of the modules were presented, drawing inspiration from the "emotional indexing" approach to teaching emotions to children with autism. We believe that such a tool offers children with Asperger's Syndrome more opportunities to engage in natural social interactions, beyond the hypothetical scenarios used in a teaching environment. The emotional hearing aid provides assistance even when the child's carer is not available, whereas the indexer ensures that events are accessible even after their occurrence for discussion and learning purposes. Future work includes completing the prototype and deploying the tool in a number of user studies to gain feedback on usability.

## 1.6 References

- Baron-Cohen S (1995) *Mindblindness: an essay on autism and theory of mind*. MIT Press
- Baron-Cohen S, Tead THE (2003) *Mind reading: The interactive guide to emotion*. Technical report. Autism Research Centre, Cambridge
- Blocher K (1999) *Affective Social Quotient (ASQ): Teaching Emotion Recognition with Interactive Media and Wireless Expressive Toys*. S.M. thesis, MIT, Cambridge, MA
- Bruce V, Young A (1998) *In the Eye of the Beholder: The Science of Face Perception*. Oxford University Press

- Cheng L, Kimberly G, Orlich F (2003) KidTalk: Online Therapy for Asperger's Syndrome. Technical Report, Social Computing Group, Microsoft Research
- Dautenhahn K, Billard A (2002) Games Children with Autism Can Play with Robota, a Humanoid Robotic Doll. *Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology [CWUAAT]* In: S. Keates, PJ. Clarkson, PM. Langdon and P. Robinson (Eds.) *Universal Access and Assistive Technology*, Springer-Verlag (London), 179-190
- Davis J, Vaks S (2001) A Perceptual User Interface for Recognizing Head Gesture Acknowledgements, in *Workshop on Perceptive User Interfaces*
- Edwards, K (1998) The face of time: Temporal cues in facial expression of emotion. *Psychological Science*, 9, 270-276
- El Kaliouby R, Robinson P (2003) Real time head gesture recognition in affective interfaces. In M. Rauterberg, M. Menozzi and J. Wesson (Eds.) *Human Computer Interaction Interact'03*, IOS Press, 950-953
- Fling E (2000) *Eating an Artichoke: A Mother's Perspective on Asperger Syndrome*. Jessica Kingsley Publishers Ltd
- Hirose M, Kijima R, Shirakawa K, Nihei K (1997) Development of a Virtual Sand Box: An Application of Virtual Environment for Psychological Treatment. In: G. Riva (Eds.) *Virtual Reality in Neuro-Psycho-Physiology: Cognitive, Clinical and Methodological Issues in Assessment and Treatment*, IOS Press
- Howlin P, Baron-Cohen S, Hadwin J (1999) *Teaching Children with autism to mind-read: A Practical Guide for Teachers and Parents*. John Wiley and Sons
- Kozima H, Yano H (2001) Designing a robot for contingency-detection game. Working Notes Workshop *Robotic & Virtual Interactive Systems in Autism Therapy*. University of Hertfordshire, Technical Report No 364
- Moore D, McGrath P, Thorpe J (2000) Computer-Aided Learning for People with Autism-a Framework for Research and Development. *Innovations in Education and Training International*, 37, 3, 218-228
- O'Connell S (1998) *Mindreading: How we learn to love and lie*. Arrow Books
- Strickland D (1996) A virtual reality application with autistic children, *Presence: Teleoperators and Virtual Environments* 5(3), 319-329
- Surakka E, Hietanen JK (1998) Facial and emotional reactions to Duchenne and nonDuchenne smiles. *International Journal of Psychophysiology*, 29(1):23-33
- Tian Y, Kanade T, Cohn J (2000) Robust lip tracking by combining shape, color and motion. In *Proceedings of the 4th Asian Conference on Computer Vision (ACCV'00)*
- Werry I, Dautenhahn K, Ogden B, Harwin W (2001) Can Social Interaction Skills Be Taught by a Social Agent? The Role of a Robotic Mediator in Autism Therapy. *Proceedings CT2001, The Fourth International Conference on Cognitive Technology: Springer Verlag, Lecture Notes in Computer Science, sub series Lecture Notes in Artificial Intelligence*