Erik Marchi, Fabien Ringeval and Björn Schuller

8 Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody

Abstract: Autism spectrum conditions (ASC) are neurodevelopmental conditions, characterized by impairments in social interaction, communication (i.e., verbal and non-verbal language), and by restricted interests and repetitive behaviour. The application of robots as a therapy tool has, however, shown promising results, namely because of the robot's ability to improve social engagement by eliciting appropriate social behaviour in children with ASC. Robots can also help clinicians in the diagnosis of ASC, by providing objective measurements of atypical behaviours that are collected during spontaneous interactions between autistic children and automata. In this chapter, we provide a review of real-life examples of voice-enabled assistive robots in the context of ASC, examining the critical role prosody plays in compensating for the lack of robust speech recognition in the population of children with ASC. This is followed by a critical analysis of some of the limitations of speech technology in the use of socially assistive robotics for young persons suffering from ASC.

8.1 Introduction

Autism spectrum conditions (ASC) are neurodevelopmental conditions in which those who suffer from autism experience difficulties with social interaction and communication (both verbal and non-verbal) with others. They also manifest overall behaviour that is generally repetitive and stereotyped. Because of these difficulties, individuals with ASC are thus challenged when using verbal and non-verbal communication for social interaction, lacking a sense of social reciprocity that can result in the failure to develop and maintain appropriate peer relationships (American-Psychiatric-Association, "DSM-IV Diagnostic and Statistical Manual of Mental Disorders", 1994, World-Health-Organization, "ICD-10 – International classification of diseases", 1994).

The social communication deficits, often present in those suffering from ASC, have a pervasive impact on their ability to meet age appropriate developmental

tasks. Such tasks may include everyday negotiation with the schoolteacher or the shopkeeper to the formation of intimate relationships with peers. As a consequence, youngsters with ASC often experience rejection, bullying and isolation (Frith 2003). Overtime, these social communication difficulties hamper the independent functioning of individuals with ASC, including their attainment of occupational and residential goals. Autism plays a significant role in their lives, affecting their ability to find friends, intimate partners and mates, and increases the likelihood of their suffering psychiatric disorders (Howlin 2004). For these reasons, it is imperative to attend to the social communication difficulties of individuals with ASC as early as possible. Indeed, studies of intervention into ASC have shown that the earlier the intervention is provided, the more effective the intervention is in getting the autistic child or young adult on the course where they can lead a relatively functional and autonomous life (Howlin & Rutter 1987).

The ability to attend to socio-emotional cues, interpret them correctly and respond to them with an appropriate expression plays a major role in social development. Three decades of research have shown that children and adults with ASC experience significant difficulties recognizing and expressing emotions and mental states (Hobson 1993; Baron-Cohen 1995). These difficulties are especially apparent when people affected by ASC attempt to recognize emotions from facial expressions (Hobson 1986; Celani, Battacchi & Arcidiacono 1999; Deruelle et al. 2004; Golan, Baron-Cohen & Hill 2006), vocal intonation (Boucher, Lewis & Collis 2000; Golan et al. 2007) as well as gestures and body language (Grèzes et al. 2009; Philip et al. 2010). Such impairments, when taken altogether, lead to difficulties in the integration of multimodal emotional information in context (Yirmiya et al. 1992; Golan, Baron-Cohen & Golan 2008; Silverman et al. 2010).

Limited emotional expressiveness in non-verbal communication is also characteristic of ASC, and studies have demonstrated individuals with ASC have difficulties directing appropriate facial expressions to others (Kasari et al. 1990; Kasari et al. 1993), modulating their vocal intonation appropriately when expressing emotion (Macdonald et al. 1989; Kasari, Chamberlain & Bauminger 2001; Michaud, Duquette & Nadeau 2003; Paul et al. 2005) and using appropriate gestures and body language (Attwood 1998). Integration of these non-verbal communicative cues with speech has for example been shown to be asynchronous (de Marchena & Eigsti 2010). In addition, individuals with ASC have difficulties understanding conversational rules and employing these rules when taking part in a reciprocal conversation (Tager-Flusberg 1992; Chin & Bernard-Opitz 2000; Peterson et al. 2009).

Given the serious communication deficits found in the autistic population, robots have been found to play a significant role. In Section 2, we provide a brief

historical overview of how broad developments in a number of fields, such as Information Communication Technology, Embodied Conversational Agents, and Socially Assistive Robots, have played a significant role in helping children who suffer from ASC. In Section 3, we examine the various voice-controlled robots that have been used to diagnose and treat autistic children. In Section 4, we provide critical analysis of technologies that perform automatic processing of prosody as applied to the socially assistive robot used for helping autistic children. In Section 5, we discuss in detail the limitations of assistive robots and provide a comparative analysis to alternative technology solutions. In Section 6, we provide our conclusion to the chapter.

8.2 Background: the role of information communication technology for diagnosing and treating ASC

The rapid progress in technology, especially in the field of information communication technology (ICT) and robotics, provides new perspectives for innovation in diagnosis and treatment of individuals with ASC. The depicted goals are quite ambitious, since the use of ICT technology focuses on the broad range of communicative problems specific to ASC. Technologic advances in recent years have led to the development of several ICT-enabled solutions for the empowerment of children with ASC (Bölte et al. 2006; Golan, Baron-Cohen & Golan 2008; Schuller et al. 2013a, 2014). For example, there exist ICT programs that aim to teach socio-emotional communication and social problem solving such as I can Problem-Solve (Bernard-Opitz, Sriram & Nakhoda-Sapuan 2001); others aim to teach emotion recognition from pictures of facial expressions and strips of the eye region such as in FEFFA (Schuller et al. 2013a). Emotion Trainer teaches emotion recognition of four emotions from facial expressions (Silver & Oakes 2001); Let's Face It teaches emotion and identity recognition from facial expressions (Tanaka et al. 2010), and Junior Detective program combines ICT with group training in order to teach social skills to children with ASC (Beaumont & Sofronoff 2008).

Embodied conversational agents (ECA) were also proposed to facilitate the collection of socio-emotional data from autistic children, allowing further automatic analysis of these data. The Rachel ECA was proposed to encourage children with ASC to produce affective and social behaviours (Mower et al. 2011b). Speech interactions between autistic children and the Rachel ECA were compared with those obtained during parent-moderated interactions, using both verbal (i.e., analysis of manual transcriptions) and non-verbal features (i.e., pitch, energy

and spectrum coefficients) (Mower et al. 2011a). No significant differences were found on these features between the two types of studied interactions, i.e., with parents or ECA, which means that data collected by using ICT can be representative of the child's abilities in the production of both verbal and non-verbal behaviours. Furthermore, ECA agents were used on children with ASC to show that the amount of social engagement to share enjoyment interactions in speech is related to acoustic patterns occurring before laughter events (Chaspari et al. 2012).

Interesting attempts to support socio-emotional communication in children with ASC also come from the field of socially assistive robotics (SAR). Indeed, children with ASC generally find socially assistive robots more predictable and less intimidating than humans. These robots can therefore be seen as a medium to enable interests for social and affective behaviours for children having ASC. Indeed, social skills such as mutual attention, turn-taking, sharing, and greeting can be practiced through child-robot interaction, even in a triadic interaction, such as the interaction of child, robot, and adult or child, robot and another child (Werry & Dautenhahn 1999; Kozima, Nakagawa & Yasuda 2005; Scassellati 2005b; Duquette, Michaud & Mercier 2008; Stanton et al. 2008; Feil-Seifer & Mataric 2009; Kozima, Michalowski & Nakagawa 2009).

Multiple studies have shown that children with ASC will interact with robots using social behaviours, e.g., by directing speech to the robot (Kozima, Nakagawa & Yasuda 2005; Robins et al. 2005; Duquette, Michaud & Mercier 2008; Stanton et al. 2008; Feil-Seifer & Mataric 2009; Kozima, Michalowski & Nakagawa 2009). Several of these studies have further demonstrated that children with ASC will interact with a parent, caregiver, or another human while engaged with a robot partner (Kozima, Nakagawa & Yasuda 2005; Robins et al. 2005; Kozima, Michalowski & Nakagawa 2009), for instance, by expressing excitement to the robot, and then returning this excitement to a parent (Kozima, Michalowski & Nakagawa 2009). Such results are very interesting as they enable parents for the first time to share affective behaviours with their autistic children, even if it requires the use of an external medium such as a robot. The lack of possibilities for parents to share affective and social behaviours with their children who are affected by ASC is one of the most heart-wrenching issues they must face in addition to the difficulties of seeing their children unable to integrate successfully into society.

Technological advances serve autistic children in yet another way. Robots can be designed to have magnified facial features, with the goal of increasing children's attention to these features. Even if these exaggerated features might not be commonly seen in everyday life interactions, they still represent an important component for enabling socio-emotional communication, thus teaching autistic children how to recognize emotions (Michaud & Théberge-Turmel 2002).

For example, Robins et al. (2005) studied the interaction of four children suffering from ASC with a humanoid robot over a period of 3 months. The authors reported an improvement in the children's imitation, turn-taking and role-switching abilities, as well as improved communicative competence. The use of robots can thus help to develop important social skills that are not originally developed in children with ASC, which can be quite promising.

All in all, though a number of robots have been created with different appearances, behaviours and target activities they are capable of doing, only a small subset of them can be considered to be voice-enabled, i.e., with the integration of speech-based technologies. The reason for this is that the integration of speech technology in robots is both difficult and challenging, especially when the robots are intended to interact with children.

Potamianos and Narayanan (2007) have examined major differences in children versus adult voices showing how acoustic, lexical and linguistic characteristics of solicited and spontaneous children's speech are correlated with age and gender. These differences are, however, even greater when looking at the population of children affected by ASC. This makes the automatic speech processing tasks much more complex when dealing with the ASC population. When one adds in the background noise of children's homes and doctors' offices, it makes it even harder for automatic speech recognition systems to perform accurately. Yet, in spite of these challenges, we do, however, consider the integration of voice-based technologies in robots as an important component to enable multimodal interaction between children with ASC and robots. The ability to convey socio-affective behaviours through speech is probably the most natural way to engage social interactions, in addition to facial expressions and body gestures.

8.3 Anthropomorphic, non-anthropomorphic, and non-biomimetic assistive robots

A multitude of robots have been used in autism therapy for children across different sites in the world with different level of success (Scassellati, Admoni & Mataric 2012; Cabibihan et al. 2013). A wide variety of physical appearances can be seen in the state-of-the-art assortment of representative systems. Scassellati, Admoni and Mataric (2012) demonstrate how robots can be grouped into three different types of physical appearance according to their resemblance with humans: anthropomorphic, non-anthropomorphic and non-biomimetic.

Anthropomorphic robots can be built to resemble to a child's physical appearance (Kozima & Yano 2012) with either realistic silicon rubber face mask and

minimal-expressive facial features (Pioggia et al. 2007; Dautenhahn et al. 2009), a doll's face with typical, albeit stylized, human appearance (Billard 2003), or a face that resembles a child's physical appearance but with simple and limited expressive abilities (Duquette, Michaud & Mercier 2008; Feil-Seifer & Mataric 2008). The representation techniques used in cartoons are often used to create robots with simple but grossly exaggerated primary features (Matsumoto, Fujii & Okada 2006; Kozima, Nakagawa & Yasuda 2007). Simplified stimuli can also be represented via robots with machine-like bodies and cartoon faces displayed on a screen (Ferrari, Robins & Dautenhahn 2009).

Non-anthropomorphic robots are designed to resemble an animals' appearance, such as the commercial robots AIBO (Stanton et al. 2008) and Pleo (Kim et al. 2012). Such robots appear social but non-intimidating; and in fact they might be more helpful than anthropomorphic robots in eliciting less complex and elementary social interactions. They can also be used as a mean to collect spontaneous data from children with ASC in a non-intrusive way, in that autistic children are often attracted to such robots which they find fascinating and non-threatening (Michaud, Duquette & Nadeau 2003).

Non-biomimetic robots do not match any biological features or appearance. Instead, they have a very simple visual appearance, such as a toy, and are designed to be very easy to use. These robots are generally used to engage children in a task or game with adults and other children (Michaud et al. 2005; Feil-Seifer & Mataric 2009). However, only some of them can perceive or generate vocal messages, such as Roball (Michaud et al. 2007), Tito (Duquette, Michaud & Mercier 2008) and Troy (Goodrich 2012). Even fewer can properly be considered as voice-controlled assistive robots, namely Paro (Marti et al. 2005), Robota (Billard et al. 2007) and Nao (Gillesen et al. 2011).

The frequently used Nao robot (Aldebaran Robotics) is roughly a half a meter tall walking robot, having 25 mechanical degrees of freedom. It is equipped with two digital high definition cameras (for computer vision such as facial and shape recognition), two speakers (for text-to-speech synthesis) and four microphones (for voice recognition and sound localization). It also has different touch sensors and wireless communication capabilities. It can thus engage in interaction through movement, speech, different LEDs in the face and body and in touch (Gillesen et al. 2011). The peculiarity of Nao lies on its design that is intended to look approachable and portray emotions similarly to a two-year old child. Gillesen et al. (2011) linked Nao to a visual programming environment that functions as an interface between robot and trainer. This was used to tailor the behaviour of the robot to the learning objectives and personal characteristics of each unique individual with ASC.

Huskens et al. (2013) investigated the effectiveness of the robot intervention, using Nao, compared to a human-trainer intervention. They reported that the interventions conducted on six children with ASC by the robot and a human trainer were both effective in promoting self-initiated questions. Ismail et al. (2012) estimated the concentration by eye contact measurements in the interaction between the humanoid robot and children with ASC. They conducted an analysis on 12 children with ASC and reported that robot-based intervention could engage more eye contact than human-human interaction. Besides being used to help children with ASC, Nao was also used in nursery schools as an assistive robot for children suffering from attention deficits or hyperactivity to improve their cognitive skills (Fridin & Yaakobi 2011).

The Paro robot was built by Sankyo Aluminium Industry and has the appearance of a baby seal. It is equipped with the four primary senses: sight (light sensor), audition (determination of sound source direction and speech recognition), balance and tactile sense. Its moving parts include vertical and horizontal neck movements, front and rear paddle movements and independent movement of each eyelid, which is important for creating facial expressions. Marti et al. (2005) investigated the use of this artificial pet in the therapeutic treatment of three children with severe cognitive impairment. Kim et al. (2010) proposed and analysed a robot-assisted method to monitor children with ASC during free playing session with the animal-like robot. Regarding the interaction dynamics, it was reported that the robot has permitted to mediate social exchange and stimulate attachment and engagement with ASC children. However, it was not clear which behavioural and physical particularities of the robot have led to these results. Pipitpukdee & Phantachat (2011) conducted a study on 34 children with ASC. They reported that the pet robot can effectively increase communication skills of children with ASC.

The last example of a voice-enabled assistive robot is the small humanoid robot Robota (Billard et al. 2007). It is a doll-shaped versatile robot that can move its arms, legs and head. In addition, it has capabilities for vision, speech recognition (Conversay) and speech synthesis (ELAN). Robota was used within the Aurora project[1] that investigates the potential use of robots as therapeutic or educational "toys" specifically for use by children with ASC. In a preliminary study, Dautenhahn & Billard (2002) tested the interaction of Robota with 14 children with ASC. The children played imitation games with the robot and promising research findings were reported.

---

[1] http://www.aurora-project.com

Although the past decade has seen significant progress in the development of socially aware robots, there are few studies on the clinical evaluation of such technology when used as a medium for the diagnosis or treatment of ASC. For example, the use of socially aware robots as a tool for overcoming the autistic triad (Diehl et al. 2012), a term used in the professional literature and by practitioners to refer to the three main impairments of autism: social and emotional, language and communication and flexibility of though. Indeed, the existing technology may need to be improved further in order to simulate realistic socio-emotional behaviours, so as to enable real-life clinical applications. The integration of emotion recognition and emotion synthesis in robots could be for example a first step in this direction, which will give the robots the ability to perceive affective behaviours produced by individuals with ASC, and respond to them in an appropriate way. This could also allow for the study of how social engagement, e.g., through turn-taking and emotional variations, could be driven for ASC children interacting with such robots. In the next section, we pinpoint the main goals and benchmarks for developing the next generation of socially assistive robots for children with ASC.

8.4 Adding prosody to socially assistive robots: challenges and solutions

Socially assistive robots are being studied as a tool to elicit target behaviours for diagnosis (Scassellati 2005a, b, 2007) and socialization (Werry et al. 2001; Dautenhahn & Werry 2002; Michaud et al. 2005; Kozima, Nakagawa & Yasuda 2005) of children with ASC. With respect to diagnosis, assistive robots can monitor children through long-term analysis of continuous data, or through machine-learning models of normative and diagnostically relevant behaviour. With regard to socialization, robots can be used to model, teach and practise social communication that involves speech, gestures and facial expression. In this scenario the use of speech technology embodied into socially assistive robots provides new perspectives to augment their capabilities when used both for diagnosis and for socialization. However, the technology implemented to date is based on speech recognition abilities, which may not work properly on children's voice, especially when these voices are atypical due to ASC, as mentioned earlier in this chapter.

Another way of integrating speech technologies in assistive robots is through the automatic processing of paralinguistic or non-verbal cues, such as speech prosody. Such an approach has the added advantage of compensating for the lack

of robust speech recognition. That is, by having access to the information transmitted by children through their non-verbal cues, such as prosody, a robot can better understand the autistic child by using low-level descriptors reflecting such non-verbal cues. However, as pointed out by Rodriguez and Lleida (2009), the extraction of prosodic features is also challenging when working with a child's voice when compared to an adult's voice due to the child's voice manifesting specific shapes of their vocal tract, which are not present in an adult.

Yet, in spite of these sorts of challenges with extraction of prosodic features from a child's voice (i.e., rhythm, stress, intonation and expressivity), prosody remains critical to human-robotic interaction for children suffering from ASC. And it is for this reason that there is a growing interest over the past two decades in investigating voice and language impairment in the ASC child population by looking at prosody (Van Lancker, Cornelius & Kreiman 1989; McCann & Peppé 2003; Paul et al. 2005; Russo, Larson & Kraus 2008; Bonneh et al. 2011; Demouy et al. 2011). In fact, atypical prosody has been identified as a core feature of individuals with ASC (Kanner 1943). The observed differences between autistic children and the typically developing (TD) population is that the former show, among other things, monotonic or machine-like intonation, aberrant stress patterns, deficits in pitch, intensity control and voice quality. Before performing a detailed analysis of the role of prosody in helping children with ASC, we first outline prosody in general, taking a look at the role it serves in non-verbal communication.

Prosody (intonation, intensity, and speed in the acoustics of the speech signal) is a supra-segmental phenomenon known to modulate and enhance the meaning of the spoken content through expressiveness at several communication levels, i.e., "grammatical", "pragmatic", and "affective" (Paul et al. 2008). Whereas prosody by itself is neither grammatical, pragmatic nor affective, these terms describe the function prosody takes on in spoken interactions. For example, grammatical prosody is used to signal syntactic information (Warren 1996). As such, acoustic stress is used to signal whether a token is being used as a noun (consider, e.g., "**con**vict") or a verb ("con**vict**").[2] Pitch contours signal the end of utterances and denote whether they are, for example, questions (e.g., by a rising pitch or in rare cases, such as the "Belfast Down", a falling pitch towards the end of the word or word phrase) or statements (e.g., by a steady or slightly falling pitch). Pragmatic prosody on the other hand conveys the speaker's intentions or the hierarchy of information within the utterance (Paul et al. 2008) which results in optional changes in the way an utterance is expressed (Van Lancker, Canter &

---

2 This grammatical difference between verb and noun in the way the word is pronounced is valid for English and does not apply to all languages.

Terbeek 1981). Thus, it carries social information beyond that conveyed by the syntax of the sentence.

Lastly, affective prosody serves a more global function than those served by the prior two forms. In so doing, it conveys a speaker's general emotional state, basically how they feel at that given moment (Winner 1988), and includes associated changes in register when talking to different listeners, e.g., peers, young children or people of higher social status (Paul et al. 2008). Because prosodic deficits contribute to language, communication and social interaction disorders and lead to social isolation, the atypical prosody in individuals with communication difficulties has become a very important research topic. Undoubtedly, prosodic awareness is integral to language skills; consequently, a deficiency in prosody may affect both language development and social interaction.

Nonetheless, it has been very difficult to characterize prosodic production differences between ASC and TD children, using manual procedures (Martínez-Castilla & Peppé 2008; Diehl & Paul 2012), even though there are marked differences in prosody between these two populations. However, some recent studies have proposed automatic systems to assess prosody production (van Santen, Prud'hommeaux & Black 2009) or speech atypicalities (Maier et al. 2009) in children.[3] Such automatic procedures may overcome the difficulties created by categorizing the evaluations (Martínez-Castilla & Peppé 2008) and by the human judging bias. Indeed, the acoustic correlates of prosody are perceptually much too complex to be fully categorized into items by humans, whom have furthermore subjective opinions (Kent 1996), and for which inter-judge variability is also problematic. However, multiple challenges have to be faced by automated systems in characterizing the prosodic variability of language atypicalities in children.

As outlined in the previous paragraph, speech prosody concerns many perceptual features such as pitch, loudness, and rhythm, which are all found in the acoustic speech waveform. Moreover, these acoustic correlates of prosody present high variability due to a set of contextual variables (e.g., disturbances caused by the recording environment) and speaker's idiosyncratic variables, such as affect (Lee & Narayanan 2005) and speaking style (Laan 1997). Yet, prosodic variations due to affective and speaking style are considered as the mean to automatically recognise the non-verbal behaviours communicated by children, rather than disturbances that compromise robustness of automatic speech recognition.

Systems based on speech prosody can, for example, be used to assess the performance of a child on a given task, e.g., producing specific prosodic contours

---

3 Automatic systems have also been used to assess early literacy in children (Black et al. 2009).

to convey sentence modality or emotions. In this case, the system is tuned for each group of children, e.g., TD and ASC, to recognise their sentence modality or emotions, and performance can be compared between the groups to provide cues regarding the observed atypicalities of ASC. Prosody-based systems can also be directly used to perform an automatic diagnosis, by comparing the children's groups. A system is, in this case, tuned to search for differences in speech production between each group of children, which can also be a mean to identify the particularities of ASC, by looking at the features retained by the system when performing the automatic recognition of typical vs. atypical speech.

8.4.1 Automatic recognition of intonation contour in atypical children's voice using static and dynamic machine learning algorithms

A recent study addressed the feasibility of designing a system that automatically assesses a child's grammatical prosodic skills through intonation contours imitation (Ringeval et al. 2011). This task, which is usually administered by speech therapists, was performed automatically using both static (k-nearest neighbours (kNN)) and dynamic (Hidden Markov Models (HMM)) machine-learning algorithms. Using the child pathological speech database (CPSD) that contains prompted imitation of 26 sentences, representing four types of intonation contour (raising, falling, descending and floating) produced in French by children with ASC (10 male and 2 female at the age of 6 to 18 years), pervasive developmental disorders non-otherwise specified (PDD-NOS; 9 male and 1 female at the age of 7 to 14 years), dysphasia (DYS; 10 male and 3 female at the age of 6 to 18 years) and TD children (52 male and 12 female at the age of 6 to 19 years), it was shown that TD children do not use the same strategy as pathologic children (PC) to convey grammatical prosodic information. Instead, PC subjects use more prosodic contour transitions (i.e., variations of pitch and energy over time) than statistically specific features (e.g., mean/standard-deviation of pitch and energy on the whole imitated sentence) to convey the modality.

These findings can be illustrated by the better performance obtained with a dynamic classifier (i.e., HMM) compared to a static classifier (i.e., kNN) in the automatic recognition of the prosodic contours imitated by the PC subjects, whereas the opposite has been observed for TD children, i.e., the static classifier performed better than the dynamic classifier, see Fig. 8.1. According to the used machine-learning algorithm, 6 low-level descriptors (LLDs) were used for the dynamic approach (i.e., pitch, energy and their first and second order derivatives), whereas 162 features were used for the static approach (i.e., the combination of the 6 LLDs with a set of 27 statistical measures), cf. table 1 in Ringeval et al. (2011).
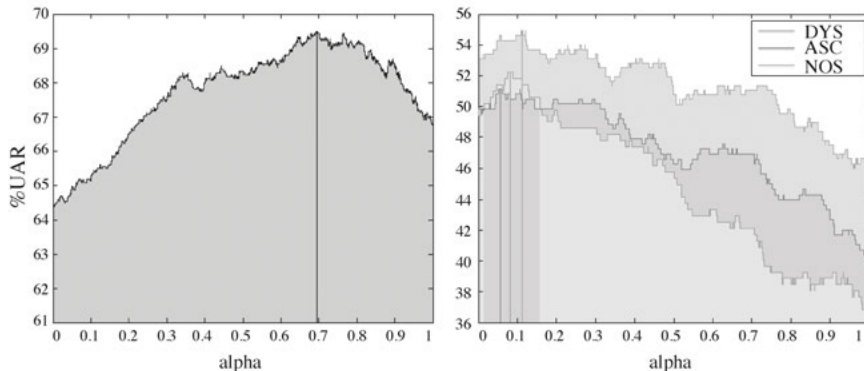
Fig. 8.1: Unweighted average recall of intonation contours using linearly weighted (by a weight-factor alpha) combination of static (alpha=1 equals only static) and dynamic (alpha=0 resembles the "other end of the scale", i.e., dynamic only) classifiers; left: results on typically developing children, right: results on pathologic children; DYS: dysphasia; ASC: autism spectrum conditions; NOS: pervasive developmental disorders not-otherwise specified (Ringeval et al. 2011).

Details of the performance for the fusion intonation recognition system are given in Tab. 8.1. The measure of performance is unweighted average recall (UAR), which takes into account the unbalanced distribution of instances over the categories of intonation contour. The score obtained for all groups of pathology were close to those of TD children and similar between each pathologic group for the "descending" intonation, such as statements, while all other intonations were significantly different ($p < 0.05$) between TD children and PC. However, the system had very high recognition rates for the "rising" intonation for DYS and TD children whereas it performed significantly worse for both ASC and PDD-NOS ($p < 0.05$). This result is consistent with studies that showed that autistic children have more difficulties at imitating questions than statements (Fosnot & Jun 1999) as well as with imitating both short and long prosodic items (McCann et al. 2007; Paul et al. 2008). As pragmatic prosody was strongly conveyed by the "rising" intonation due to the short questions, it is not surprising that such intonation recognition differences were found between DYS children and ASC children.

Indeed, both ASC and PDD-NOS children show pragmatic deficits in communication, whereas DYS children only show pure language impairments. Moreover, Snow (1998) hypothesized that rising pitch requires more effort in physiological speech production than falling tones and that some assumptions could be made regarding the child's ability or intention to match the adult's speech. Because the "rising" intonation included very short sentences (half the duration) compared with others, which involves low working memory load, DYS

Tab. 8.1: Performance of automatic recognition of intonation contours reproduced by four groups of children, using fusion of static and dynamic classifiers (Ringeval et al. 2011).

[%]	TD	ASC	NOS	DYS
Descending	64	64	70	63
Falling	55	35[*T]	45[*T]	39[*T]
Floating	72	48[*T]	40[*T]	31[*T]
Rising	**95**	57[*T,D]	48[*T,D]	**81**[*T,A,N]
All	70	56[*T]	53[*T]	58[*T]

Performance is given in UAR; * = $p < 0.05$: alternative hypothesis is true when comparing data from child groups, i.e., T, A, N and D; T: typically developing; A: autism spectrum conditions; N: pervasive developmental disorders non-otherwise specified; D: dysphasia.

children were not disadvantaged compared to ASC children as was found in (Wells & Peppé 2003).

Whereas some significant differences were found in the PC's groups with the "rising" intonation, the global mean recognition scores did not show any dissimilarity between children. All PC subjects showed similar difficulties in the administered intonation imitation task as compared to TD children, whereas differences between DYS and PDDs only appeared on the "rising" intonation; the latter is probably linked to deficits in the pragmatic prosody abilities of PDD and PDD-NOS. The automatic approach used to assess PC prosodic skills in an intonation imitation task confirms the clinical descriptions of the subjects' communication impairments (Demouy et al. 2011). This is a very promising result when aiming at automatically evaluating atypicality in children's voice with ASC that perform a specific task such as intonation contours imitation in the described case. The integration of such an automatic approach in voice-enabled socially assistive robots could provide an interesting support for the assessment of prosodic skills during clinical evaluations. Additionally, the long-term monitoring of prosodic skills of children suffering from ASC in everyday life interaction could be made possible by having a robot present in non-clinical and uncontrolled environment, e.g., at school or at home. The data collected in such long-term interaction could thus be analysed to assess progress of children with ASC in specific tasks, but also identify which kind of context can foster progress in social engagement.

8.4.2 Automatic recognition of emotions in atypical children's voice

To our best knowledge, only few studies exist, which deal with automatic emotion analysis in speech of autistic children. A preliminary study has recently focused

on the recognition of emotional vocal expressions by comparing performance of a few prosodic features against large sets of acoustic, spectral and cepstral features (Marchi et al. 2012b). The study was conducted on the ASC-DB database (Marchi et al. 2012a) that contains prototypical emotions (the "big six" emotions as defined by Ekman (1999), except disgust, plus four mental states: ashamed, calm, proud and "neutral") uttered in Hebrew by 9 children suffering from ASC (8 male and 1 female; age 6 to 12) and 11 TD children (5 female and 6 male; age 6 to 9). Overall, it includes 529 utterances of emotional speech: 178 utterances of children with ASC (focus group), and 351 utterances of TD children (control group).

Three emotion recognition tasks were performed separately on the data collected from both TD and ASC children: one task was devoted to the recognition of one emotion out of the nine emotion categories, a second task focused on the classification of high and low arousal, and the last task on the classification of positive and negative valence. Support vector machines (SVMs) were used for the automatic classification task with a linear kernel. Leave-one-speaker-out cross-validation was used to ensure speaker independence during the automatic evaluation. Two feature sets were used for the analysis of the extent to which specific prosodic features are relevant for the recognition of a child's emotional state: a large features set (termed here "IS12"), stemming from the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al. 2012), that contains 6128 acoustic features including spectral features, voice quality features and prosodic features; and a reduced feature set (termed here "PROS"), that consists of four statistical functionals (mean, standard deviation, maximum and minimum values) computed on few prosodic descriptors: energy such as root-means-square signal frame energy; fundamental frequency (F0); and duration of the F0 contours.

Table 8.2 shows the results as reported for the optimal configuration by Marchi et al. (2012b). As one may expect, the nine-class task is the most challenging and a large decrease of performances is observed when only prosodic features (i.e., "PROS") are used in the cases where valence aspects are included, whereas the arousal task seems to be comparably well modelled by prosodic features exclusively, i.e., without as high a loss in performance. In fact, this may also stem from the commonly agreed fact that arousal is easier assessed by acoustics than valence is.

These empirical studies show that the analyses of prosodic and spectral features allow a reliable automatic recognition of emotions in atypical children's voice. Therefore, such systems could be integrated into voice-enabled socially assistive robots, which will provide the ability to know the emotional state of the child and drastically improve the quality of the child-robot interaction. Besides the automatic recognition of emotions in the voice of autistic children, another promising novel task that could be integrated into voice-enabled socially assistive

Tab. 8.2: Performance of automatic recognition of 9-class emotion and 2-class arousal and valence from speech of ASC children (focus group) and TD children (control group) for two different features sets (Marchi et al. 2012b).

UAR [%]	IS12	PROS
Focus group subset		
9-class Emotion	42.6	28.9
2-class Arousal	84.9	78.8
2-class Valence	82.1	55.1
Control group subset		
9-class Emotion	55.9	18.8
2-class Arousal	89.0	77.5
2-class Valence	81.8	52.4

Unweighted Average Recall (UAR) for a nine emotions task and for binary arousal/valence tasks on a focus group subset and on a control group subset. Shown are (the best) performances obtained with speaker z-normalization for two feature sets (IS12, PROS).

robots deals with the recognition of ASC by their acoustics, and will be discussed in the next section.

8.4.3 Automatic diagnosis of atypical children's voice

The relatively novel task of the automatic diagnosis of children with ASC based on their acoustic features has been addressed more broadly in the context of an open research competition at the recent INTERSPEECH 2013 Computational Paralinguistic Challenge (ComParE 2013) by Schuller et al. (2013b). The autism sub-challenge was based upon the CPSD database that was proposed by Ringeval et al. (2011), which was previously described in Section 4.1 above. As a reminder, speech data were collected by the imitation of prosodic contours by four groups of children (TD, ASC, PDD-NOS and DYS). For the purpose of the computational paralinguistic challenge, the organisers divided the data into speaker disjoint subsets for training, development and testing (Schuller et al. 2013b). The subject ID (anonymous code) of the children was made available to participants of the challenge only on training and development partitions, and was blinded on the test partition; participants were permitted to submit their predictions on the test dataset up to five times. Two speaker independent evaluation tasks have been defined for the challenge: a binary "typicality" task (i.e., typically vs. atypically developing children) by clustering the three non-control group children into one

group, and a "full four-way" "diagnosis" task, i.e., classifying into all four above named groups of (a-)typical development.

The baseline approach that was used for these two tasks used a large set of acoustic features as reported by Schuller et al. (2013b); this set is actually a slight four percent extension of the features that were described in the previous Section 4.2 and contains 6373 features. Static classification was used for the baseline of the organisers: the "typicality" and "diagnosis" tasks were assessed by SVMs with linear kernel. Table 8.3 shows the results for the two tasks. The binary "typicality" task can alternatively be solved by mapping from the four-way task to the two-way decision leading to a high 90.7% UAR on the test set. The four-way "diagnosis" task led to a significant decrease in performance, with only 67.1% UAR on the test set.

The performance of this baseline system was, however, slightly improved by participants of ComParE 2013: the best system reached 93.5% UAR and 69.4% UAR on the test partition, for the "typicality" and "diagnosis" tasks, respectively (Asgari, Bayestehtashk & Shafran 2013). The improvement was made possible by adding voice quality features to the baseline feature set, and using a combination of both SVM based regression and classification. The results of this challenge show that the recognition of atypical voice between TD and three groups of PDD, including ASC, can be carried out in an automatic way with a performance that was by far higher than the chance level (25% UAR for four classes). The automatic diagnosis is, however, demanding in terms of accuracy and robustness of the automatically extracted prosodic features.

Marchi et al. (2012b) addressed another study on the automatic recognition of atypical speech between TD children and children suffering from ASC. The evaluations were based upon the ASC-DB database of prototypical emotional utterances as described shortly above in Section 4.2. The emotional speech of children with ASC comprises 178 utterances of which, 90 and 88 are performed, respectively, by children with Asperger syndrome (AS) and high-functioning (HF)

Tab. 8.3: Performance for the automatic recognition of children's (a-)typicality from the voice (imitated intonation contours; baseline and best participant result of the ComParE 2013 autism sub-challenge) (Schuller et al. 2013b).

UAR [%]	Baseline	Best
2-class Typicality	90.7	93.5
4-class Diagnosis	67.1	69.4

UAR for typicality and diagnosis tasks; baseline and winning team (best) on the test set, by training on the training and development sets.

Tab. 8.4: Performance of automatic recognition of atypical speech (Marchi et al. 2012b).

UAR [%]	IS12	PROS
Full data set		
2-class Typicality	80.0	55.5
Focus group subset		
2-class Diagnosis	82.6	59.3

UAR for typicality and diagnosis tasks, respectively on the entire dataset and on the focus group subset. Typicality classes: typically developing children vs. children with ASC. Diagnosis classes: Asperger Syndrome, high-functioning.

diagnosis. The experimental set-up of the experiments is identical to the one described in Section 4.2.

The recognition of atypical speech was evaluated by the authors with two tasks: the "typicality" task concerns the classification of typically developing children versus children with ASC; the "diagnosis" task aims to distinguish between Asperger syndrome and high-functioning diagnosis. The "typicality" task was performed on the full data set, whereas the "diagnosis" task was evaluated on the focus group only.

Table 8.4 shows the results obtained with the "large" feature set (IS12) and the prosodic feature set (PROS) as detailed above in Section 3.2. With the high dimensional feature set a UAR of 80.0% and 82.6% is obtained for typicality and diagnosis, respectively. Both tasks visibly rely significantly on spectral and voice quality features, thus using only prosodic features was observed by the authors to lead to a severe decrease in performance.

The inclusion of automatic diagnosis of autistic children by using prosodic and spectral features could lead to an increased flexibility of the voice-enabled socially assistive robots. In fact, personalised models could be automatically loaded according to the inferred diagnosis. This could enable robots to be used in group scenarios where the interactions could include typically developing children and children suffering from ASC.

8.4.4 The acoustics of eye contact

A further aspect that could open new perspectives for socially assistive robots employed in therapy with autistic children is the use of acoustics to detect visual focus of attention from conversational audio cues. Indeed, an important aspect of social interactions in short dialogues is the attention paid to others as is usually

manifested by specific patterns in gaze behaviour between subjects. The ability to detect visual attention only based on speech data could be a mean to integrate such information without using dedicated algorithms based on visual information processing and without ever-present "camera-observation". If cameras are used, adding acoustic analysis could also help to improve performance of according systems. More importantly, however, analysis of acoustic properties of ASC children that have eye contact with their conversational partners could verify if the voice naturally matches the situation.

Eyben et al. (2013) have provided a first analysis whether such visual attention has an impact on the acoustic properties of a speaker's voice. The analysis was conducted on the multi-modal GRAS² corpus, which was recorded for analysing attention in human-to-human interactions of short daily-life communication with strangers in public places. Recordings of four test subjects interacting with several strangers while equipped with eye tracking glasses, three audio recording devices, and motion sensors are contained in the corpus. This study finds significant correlations between the acoustics of the voice and the distance between the point of view and the eye region of the dialogue partner. Further, it shows that automatic classification of binary decision of eye-contact vs. no eye-contact from acoustic features alone is feasible with a UAR of up to 70%.

This result reveals that the automatic detection of eye-contact during dyadic interaction can be estimated from speech features with a performance significantly higher than chance. A robot could, for example, use such information to provide a stimulus to children with ASC when eye-contact between their conversational partners is assumed, with the goal of increasing their interest in exchanging socio-affective interactions with others.

8.5 Limitations

Diehl et al. (2012) have conducted a study to understand the current status of empirically-based evidence on the clinical applications of robots in the diagnosis and treatment of ASC. They found that most of the findings are exploratory and have methodological limitations that make it difficult to draw firm conclusions about the clinical utility of robots. This observation concords with the fact that the majority of human-robot interaction currently occurs in research laboratories where systems are specifically engineered for one environment and for a pre-determined prototypic user population. As SAR become more widespread in homes, schools, and hospitals, the question of scalability and adaptability arises. Besides this aspect of controlled environments that calls for more robust

integration of signal processing-based technology in SAR, for SAR to be effectively used in various conditions there are still several remaining issues regarding the development of the technology itself. Despite the fact that some recent development appears promising for diagnosis and therapy of children with ASC, there exist some important limitations that need to be overcome, especially regarding the integration of speech based technology. The most crucial ones are outlined below.

Speech recognition and synthesis – The communication parameters play a relevant role in the way a robot can effectively interact with a user. Avramides et al. (2012) have reviewed these characteristics and have shown that the naturalness of the interactions is related to the type of voice the robot uses, which can be either a synthetized or recorded voice; a female, male or artificial voice which may or may not contain emotion.

Since socially assistive systems must provide those suffering from ASC a way to learn social skills that can be used practically in social interactions, SAR need to recreate a real-life conversation scenario. However, speech recognition of children is a difficult problem in itself, but it is even greater when children have ASC (Gerosa et al. 2009). For example, while spontaneous speech often contains disfluencies that significantly perturb the reliability of automatic speech recognition systems (Yildrim & Narayanan 2009), study findings suggest that ASC children show a significantly higher amount of disfluencies than typically developing children (Koegel et al. 1998; Scott et al. 2013). Considering the limitations of actual ASR systems (ten Bosch 2003), the majority of interactive systems for children with ASC that enable speech input are actually prompted by a human user (Tartaro & Cassel 2010; Milne et al. 2010). The other major area which plays an important role for socially assistive systems is speech synthesis. However, it is known that the production of speech synthesis for truly natural, emotional or child speech still presents massive difficulties (Tartaro & Cassel 2010; Watts et al. 2010).

Emotion recognition and speech corpora – Whereas many studies exist that have investigated the ability of autistic children to recognize and mimic facial emotion expressions, few studies deal with children's vocal emotion recognition and expression abilities (Loveland et al. 1997; Boucher, Lewis & Collis 2000). Furthermore and as mentioned in the beginning of this chapter, there are also few studies that deal with automatic emotion analysis of speech of ASC children (Marchi et al. 2012a, b). Boucher, Lewis & Collis (2000) indicate that autistic children show differences in control of articulation and intonation when compared to regular children. Thus, when developing automatic emotion analysis systems for children suffering from ASC, many parameters of current systems must be re-evaluated under these conditions.

Many recent studies dealing with naturalistic emotions, however, deal with adult speech. The reason for that is that many commercially interesting applications of emotion recognition technology, such as for detecting customer frustration with call center agents and IVRs, road rage among motor vehicle drivers, and perturbations in those participating in high stakes computer gaming are primarily intended for adults. Unfortunately at present there are only few child speech corpora with emotion labels, which can be used for research regarding children's emotional speech. Likely the most widely used and known one is the FAU Aibo Emotion corpus (Steidl 2009), which was used for the first INTERSPEECH 2009 Emotion Challenge (Schuller et al. 2009). A recognition rate of 44% for a 5-class task is the current state of the art, which was obtained by fusing the decisions of the best challenge submissions.

These results indicate the great challenges of naturalistic emotions in conjunction with children's speech (Schuller et al. 2011). Indeed, as children's speech differs largely from adult speech due to some of the variables we outlined above such as different vocal tract sizes, immature pronunciation, simpler grammar and vocabulary, methods and models tuned to traditional tasks of performing adult speech and emotion recognition must not only be adapted to the domain of children's speech, but must be revisited ab initio so that we can learn how such models can be adapted to accommodate the unique construction of children's voices.

Additionally, the availability of speech corpora is positively correlated with typicality: the more typical the population is, the easier it generally is to collect enough data for building relevant models. The less typical the envisaged population is, the more difficult it is to obtain sufficient amounts of data. For example, children with ASC are a population that is atypical in several respects: they are a limited age group, they might have problems with an experimental setting where their speech should be recorded, and they belong to a specific subgroup of children. Recruiting children for scientific studies is also often more difficult than it is for adults, because the consent of the parent is needed for the child to participate in the study. Several ethical issues also need to be carefully addressed when recording data, especially with children. As a consequence, current databases of children with ASC rarely contain more than 10 subjects, which can only provide indicative pointers rather than strong markers of their corresponding deficiencies. Given the described limitations, it is clear that speech emotion recognition in children is error-prone.

In this context, it seems noteworthy to mention a recent ICT-enabled solution, namely the ASC-Inclusion project (Schuller et al. 2013a, 2014). This project deals with children's vocal emotion recognition among other modalities. Its goal is to

create an internet-based platform that will assist children with ASC to improve their socio-emotional communication skills, attending to the recognition and expression of socio-emotional cues and to the understanding and practice of conversational skills. It does so by combining several technologies in one game environment, including further analysis of users' gestures and facial expressions.

8.6 Conclusions

In this chapter, we discussed the perspectives and limitations of speech technology applied to socially assistive robotics for individuals with ASC. We first gave examples of voice-enabled assistive robots, for which there is empirically-based evidence in the professional literature on the clinical applications of such robots in the diagnosis and treatment of ASC. We subsequently explored how the use of speech technology embodied in socially assistive robots provides new perspectives to augment the capabilities of robots when used for both diagnosis and socialization. More specifically, we showed how speech prosody could be seen as a promising avenue to improve real-life systems that are used for the automatic recognition of atypicalities in ASC children's voice, as a natural extension of typical ASR systems that encounter massive problems with analyzing children's voices in general (Steidl et al. 2010; Wöllmer et al. 2011).

Based on this reflection of the state-of-the-art and the latest results in the field that we provided in this chapter, we find that new research paradigms are very much needed to address this important topic. Such paradigms require nothing less than a multi-disciplinary approach, closely uniting computer and industrial engineers with clinicians and others working in related fields. This will ensure that the development of socio-affective based technology will find its way out of the laboratories so that it can be made an integral part of the design of socially assistive robots that can help in the everyday lives of children with ASC.

Acknowledgments

References

American Psychiatric Association (1994) *DSM-IV Diagnostic and Statistical Manual of Mental Disorders.* 4th Edition. American Psychiatric Association: Washington DC.

Asgari, M., Bayestehtashk, A. & Shafran, I. (2013) Robust and accurate features for detecting and diagnosing autism spectrum disorder. Interspeech 2013, *14th Annual Conference of the International Speech Communication Association (ISCA), 25–29 August, Lyon, France*, pp. 191–194.

Attwood, T. (1998) *Asperger's Syndrome: A Guide for Parents and Professionals*. Jessica Kingsley Publishers: London.

Avramides, K., Bernardini, S., Foster, M. E., Frauenberger, C., Kossyvaki, L. & Mademtzi, M. (2012) 'State-of-the-art in TEL to support social communication skill development in children with autism: a multi-disciplinary review', *International Journal of Technology Enhanced Learning (IJTEL),* 4(5–6):359–372.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press/ Bradford Book: Boston.

Beaumont, R. & Sofronoff, K. (2008) 'A multi-component social skills intervention for children with Asperger syndrome: The junior detective training program', *Journal of Child Psychology and Psychiatry*, 49(7):743–753.

Bernard-Opitz, V., Sriram, N. & Nakhoda-Sapuan, S. (2001) 'Enhancing social problem solving in children with autism and normal children through computer-assisted instruction', *Journal of Autism and Developmental Disorders*, 31(4):377–384.

Billard, A. (2003) 'Robota: clever toy and educational tool', *Robotics and Autonomous Systems: Socially Interactive Robots*, 42(3–4):259–269.

Billard, A., Robins, B., Nadel, J. & Dautenhahn, K. (2007) 'Building Robota, a mini-humanoid robot for the rehabilitation of children with autism', *Assistive Technology: The Official Journal of RESNA*, 19(1):37–49.

Black, M., Tepperman, J., Kazemzadeh, A., Lee, S. & Narayanan, S. (2009) Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children. *IEEE 34th International Conference on Acoustic, Speech and Signal Processing (ICASSP), 19–24 April, Taipei, Taiwan,* pp. 4861–4864.

Bölte, S., Hubl, D., Feineis-Matthews, S., Prvulovic, D., Dierks, D. & Poustka, F. (2006) 'Facial affect recognition training in autism: can we animate the fusiform gyrus?', *Behavioral Neuroscience*, 120(1):211–216.

Bonneh, Y. S., Levanon, Y., Dean-Pardo, O., Lossos, L. & Adini, Y. (2011) 'Abnormal speech spectrum and increased pitch variability in young autistic children', *Frontiers in Human Neuroscience*, 4(237).

ten Bosch, L. (2003) 'Emotions, speech and the ASR framework', *Speech Communication,* 40(1–2):213–225.

Boucher, J., Lewis, V. & Collis, G.M. (2000) 'Voice processing abilities in children with autism, children with specific language impairments, and young typically developing children', *Journal of Child Psychology and Psychiatry*, 41(7):847–857.

Cabibihan, J. J., Javed, H., Ang Jr., M. & Aljunied, S.M. (2013) 'Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism', *International Journal of Social Robotics*, 5(4):593–618.

Celani, G., Battacchi, M. W. & Arcidiacono, L. (1999) 'The understanding of the emotional meaning of facial expressions in people with autism', *Journal of Autism and Developmental Disorders*, 29(1):57–66.

Chaspari, T., Provost, E. M., Katsamanis, A. & Narayanan, S. (2012) An acoustic analysis of shared enjoyment in ECA interactions of children with autism. *IEEE 37th International Conference on Acoustic, Speech and Signal Processing (ICASSP), 25–30 March, Kyoto, Japan*, pp. 4485–4488.

Chin, H. Y. & Bernard-Opitz, V. (2000) 'Teaching conversational skills to children with autism: effect on the development of a theory of mind', *Journal of Autism and Developmental Disorders*, 30(6):569–583.

Dautenhahn, K. & Werry, I. (2002) *A quantitative technique for analysing robot-human interactions*. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 30 September–4 October, Lausanne, Switzerland, 2:1132–1138.

Dautenhahn, K. & Billard, A. (2002) Games children with autism can play with Robota, a humanoid robotic doll. *Universal Access and Assistive Technology*. Springer: London, part 2, pp. 179–190.

Dautenhahn, K., Nehaniv, C., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A. & Blow, M. (2009) 'KASPAR – a minimally expressive humanoid robot for human-robot interaction research', *Applied Bionics and Biomechanics*, 6(3–4):369–397.

Demouy, J., Plaza, M., Xavier, J., Ringeval, F., Chetouani, M., Perisse, D., Chauvin, D., Viaux, S., Golse, B., Cohen, D. & Robel, L. (2011) 'Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment', *Research in Autism Spectrum Disorders*, 5(4):1402–1412.

Deruelle, C., Rondan, C., Gepner, B. & Tardif, C. (2004) 'Spatial frequency and face processing in children with autism and Asperger syndrome', *Journal of Autism and Developmental Disorders*, 34(2):199–210.

Diehl, J. J., Schmitt, L. M., Villano, M. & Crowell, C. R. (2012) 'The clinical use of robots for individuals with autism spectrum disorders: A critical review', *Research in Autism Spectrum Disorders*, 6(1):249–262.

Diehl, J. J. & Paul, R. (2012) 'Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders', *Research on Autism Spectrum Disorders*, 6(1):123–134.

Duquette, A., Michaud, F. & Mercier, H. (2008) 'Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism', *Autonomous Robots*, 24(2):147–157.

Ekman, P. (1999) *Basic emotions.* In *Handbook of Cognition and Emotion,* T. Dalgeish and M. Power (eds.), John Wiley & Sons Ltd., 98:45–60.

Eyben, F., Weninger, F., Paletta, L. & Schuller, B. (2013) The acoustics of eye contact – detecting visual attention from conversational audio cues. *6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction (GAZEIN), Satellite Event of the 15th International Conference on Multimodal Interaction (ICMI), 13 December, Syndey, Australia, ACM New York*, pp. 7–12.

Feil-Seifer, D. & Mataric, M. J. (2008) *B³IA: a control architecture for autonomous robot-assisted behaviour intervention for children with autism spectrum disorders*. 17th IEEE International Symposium on Robot Human Interactive Communication (RO-MAN 2008), 1–3 August, Munich, Germany, pp. 328–333.

Feil-Seifer, D. & Mataric, M. J. (2009) *Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders*. 11th International Symposium on Experimental Robotics, Khatib, O., Kumar, V. & Pappas, G. J. (eds.), *Springer Tracts in Advanced Robotics,* Springer: Berlin, Heidelberg, 54:201–210.

Ferrari, E., Robins, B. & Dautenhahn, K. (2009) Therapeutic and educational objectives in robot assisted play for children with autism. *18th IEEE International Symposium on Robot Human Interactive Communication (RO-MAN 2009), 27 September–2 October, Toyama, Japan,* pp. 108–114.

Fosnot, S. M. & Jun, S. (1999) Prosodic characteristics in children with stuttering or autism during reading and imitation. *14th International Congress of Phonetic Sciences (ICPhS), 1–7 August, San Francisco,* CA, pp. 1925–1928.

Fridin, M. & Yaakobi, Y. (2011) Educational robot for children with ADHD/ADD, architecture design. *Paper presented at the International Conference on Computational Vision and Robotics (ICCVR), 13–14 August, Bhubaneswar, India.*

Frith, U. (2003) *Autism: Explaining the Enigma*, Blackwell Publisher: Oxford.

Gerosa, M., Giuliani, D., Narayanan, S. & Potamianos, A. (2009) A review of ASR technologies for children's speech. *2nd Workshop on Child, Computer and Interaction (WOCCI), Satellite Event of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI), 5 November, Cambridge, MA*, pp. 71–78.

Gillesen, J., Barakova, E., Huskens, B. & Feijs, L. (2011) From training to robot behavior: Towards custom scenarios for robotics in training programs for ASD. *IEEE International Conference on Rehabilitation Robotics (ICORR), 29 June–1 July, Zurich, Switzerland,* pp. 1–7.

Golan, O., Baron-Cohen, S. & Hill, J. J. (2006) 'The Cambridge mindreading (CAM) face-voice battery: testing complex emotion recognition in adults with and without Asperger syndrome', *Journal of Autism and Developmental Disorders*, 36(2):169–183.

Golan, O., Baron-Cohen, S., Hill, J. J. & Rutherford, M. D. (2007) 'The 'reading the mind in the voice' test - revised: A study of complex emotion recognition in adults with and without autism spectrum conditions', *Journal of Autism and Developmental Disorders*, 37(6):1096–1106.

Golan, O., Baron-Cohen, S. & Golan, Y. (2008) 'The 'reading the mind in films' task [child version]: complex emotion and mental state recognition in children with and without autism spectrum conditions', *Journal of Autism and Developmental Disorders*, 38(8):1534–1541.

Goodrich, M. A., Colton, M., Brinton, B., Fujiki, M., Alan Atherton, J., Robinson, L., Ricks, D., Hansen Maxfield, M. & Acerson, A. (2012) 'Incorporating a robot into an autism therapy team', *IEEE Intelligent Systems Magazine*, 27(2):52–59.

Grèzes, J., Wicker, B., Berthoz, S. & de Gelder, B. (2009) 'A failure to grasp the affective meaning of actions in autism spectrum disorder subjects', *Neuropsychologia*, 47(8–9):1816–1825.

Hobson, R. P. (1986) 'The autistic child's appraisal of expressions of emotion', *Journal of Child Psychology and Psychiatry*, 27(3):321–342.

Hobson, R. P. (1993) *Autism and the Development of Mind*. Psychology Press, Lawrence Erlbaum Associates Publishers: Hove (UK), Hillsdale (USA).

Howlin, P. & Rutter, M. (1987) *Treatment of Autistic Children*. John Wiley & Sons.

Howlin, P. (2004) *Autism and Asperger Syndrome. Preparing for Adulthood*, 2nd Edition, Routledge, Taylor & Francis Group: London and New York.

Huskens, B., Verschuur, R., Gillesen, J., Didden, R. & Barakova, E. (2013) 'Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot

intervention compared to a human-trainer intervention', *Developmental Neurorehabilitation*, 16(5):345–356.

Ismail, L. I., Shamsudin, S., Yussof, Hanafiah, Y., Hanapiah, F. A. & Zahari, N. I. (2012) 'Estimation of concentration by eye contact measurement in robot-based intervention program with autistic children', *Procedia Engineering*, *International Symposium on Robotics and Intelligent Sensors (IRIS)*, 41:1548–1552.

Kanner, L. (1943) 'Autistic disturbances of affective contact', *Nervous Child*, 2:217–250.

Kasari, C., Sigman, M., Mundy, P. & Yirmiya, N. (1990) 'Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children', *Journal of Autism and Developmental Disorders*, Springer, 20(1):87–100.

Kasari, C., Sigman, M., Yirmiya, N. & Mundy, P. (1993) Affective Development and Communication in Young Children with Autism. In *Enhancing Children's Communication: Research Foundations for Intervention, Communication and Language Intervention Series*, Kaiser, A. P. & Gray, D. B. (Eds.), Paul H. Brookes Publishing, 2:201–22.

Kasari, C., Chamberlain, B., Bauminger, N. (2001) Social emotions and social relationships: Can children with autism compensate? *The Development of Autism: Perspectives from Theory and Research*, Burack, J. A., Charman, T., Yirmiya, N. & Zelazo, P.R., (Eds.), Lawrence Erlbaum Associates Publishers, pp. 309–323.

Kent, R. D. (1996) 'Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders', *American Journal of Speech-Language Pathology,* 5(3):7–23.

Kim, E. S., Paul, R., Shic, F. & Scassellati, B. (2012) 'Bridging the research gap: making HRI useful to individuals with autism', *Journal of Human-Robot Interaction*, 1(1):26–54.

Kim, Y. D., Hong, J. W., Kang, W. S., Baek, S. S., Lee, H. S. & An, J. (2010) Design of robot assisted observation system for therapy and education of children with autism. *Social Robotics, 2nd International Conference on Social Robotics (ICSR), Lecture Notes in Computer Science (LNCS)*, Springer: Berlin, Heidelberg, 23–24 November, Singapore, pp. 222–231.

Koegel, R. L., Camarata, S., Koegel, L. K., Ben-Tall, A. & Smith, A. E. (1998) 'Increasing speech intelligibility in children with autism', *Journal of Autism and Developmental Disorders*, 28(3):241–251.

Kozima, H., Nakagawa, C. & Yasuda, Y. (2005) *Interactive robots for communication-care: A case-study in autism therapy. 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN), 13–15 August, Nashville, TN*, pp. 341–346.

Kozima, H., Nakagawa, C. & Yasuda, Y. (2007) 'Children-robot interaction: a pilot study in autism therapy', *Progress in Brain Research*, *From Action to Cognition*, 164:385–400.

Kozima, H., Michalowski, M. P. & Nakagawa, C. (2009) 'Keepon. A playful robot for research, therapy, and entertainment', *International Journal of Social Robotics*, 1(1):3–18.

Kozima, H. & Yano, H. (2012) A robot that learns to communicate with human caregivers. *1st International Workshop on Epigenetic Robotics, 17–18 September, Lund, Sweden,* pp. 47–52.

Laan, G. P. M. (1997) 'The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and read speaking style', *Speech Communication,* 22(1):43–65.

Lee, C. M. & Narayanan, S. (2005) 'Toward detecting emotions in spoken dialogs', *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.

Loveland, K. A., Tunali-Kotoski, B., Chen, Y. R., Ortegon, J., Pearson, D. A., Brelsford K. A. & Gibbs, M. C. (1997) 'Emotion recognition in autism: Verbal and nonverbal information', *Development and Psychopathology*, 9(3):579–593.

Macdonald, H., Rutter, M., Howlin, P., Rios, P., Le Conteur, A. & Evered, C. (1989) 'Recognition and expression of emotional cues by autistic and normal adults', *Journal of Child Psychology and Psychiatry*, 30(6):865–877.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M. & Nöth, E. (2009) 'PEAKS – A system for the automatic evaluation of voice and speech disorder', *Speech Communication,* 51(5):425–437.

de Marchena, A. & Eigsti, I. M. (2010) 'Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency', *Autism Research*, 3(6):311–322.

Marchi, E., Batliner, A., Schuller, B., Fridenzon, S., Tal, S. & Golan, O. (2012a) Speech, emotion, age, language, task and typicality: Trying to disentangle performance and feature relevance. *1st International Workshop on Wide Spectrum Social Signal Processing (WS³P), Satellite Event of the 4th ASE/IEEE International Conference on Social Computing (SocialCom), 3 September, Amsterdam, Netherlands,* pp. 961–968.

Marchi, E., Schuller, B., Batliner, A., Fridenzon, S., Tal, S. & Golan, O. (2012b) Emotion in the speech of children with autism spectrum conditions: Prosody and everything else. *3rd Workshop on Child, Computer and Interaction (WOCCI), Satellite Event of INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association (ISCA), 14 September, Portland*, OR.

Marti, P., Pollini, A., Rullo, A. & Shibata, T. (2005) Engaging with artificial pets. *Annual Conference on European Association of Cognitive Ergonomics (EACE), 28 September – 1 October, Creta, Greece,* pp. 99–106.

Martínez-Castilla, P. & Peppé, S. (2008) 'Developing a test of prosodic ability for speakers of Iberian-Spanish', *Speech Communication*, 50(11–12):900–915.

Matsumoto, N., Fujii, H. & Okada, M. (2006) 'Minimal design for human-agent communication', *Artificial Life and Robotics*, 10(1):49–54.

McCann, J. & Peppé, S. (2003) 'Prosody in autism spectrum disorders: A critical review', *International Journal of Language & Communication Disorders*, 38(4):325–350.

McCann, J., Peppé, S., Gibbon, F., O'Hare, A. & Rutherford, M. (2007) 'Prosody and its relationship to language in school-aged children with high functioning autism', *International Journal of Language & Communication Disorders*, 42(6):682–702.

Michaud, F. & Théberge-Turmel, C. (2002) 'Mobile robotic toys and autism', *Socially Intelligent Agents*, *Multiagent Systems, Artificial Societies, and Simulated Organizations*, 3:125–132.

Michaud, F., Duquette, A. & Nadeau, I. (2003) Characteristics of mobile robotic toys for children with pervasive developmental disorders. *IEEE International Conference on Systems, Man and Cybernetics (SMC), 5–8 October, Washington*, DC, 3:2938–2943.

Michaud, F., Laplante, J. F., Larouche, H., Duquette, A., Caron, S., Letourneau, D. & Masson, P. (2005) 'Autonomous spherical mobile robot for child-development studies', *IEEE Transactions on Systems, Man, and Cybernetics*, *Part A: Systems and Humans*, 35(4):471–480.

Michaud, F., Larouche, H., Larose, F., Salter, T., Duquette, A., Mercier, H. & Lauria, M. (2007) Mobile robots engaging children in learning. *30th Canadian Medical and Biological Engineering Conference (CMBEC), 16–19 June, Toronto, Canada.*

Milne, M., Luerssen, M., Lewis, T., Leibbrandt, R. & Powers, D. M. W. (2010) Development of a virtual agent based social tutor for children with autism spectrum disorders. *International Joint Conference on Neural Networks (IJCNN), 18–23 July, Barcelona, Spain*, pp. 1–9.

Mower, E., Black, M., Flores, E., Flores, E., Williams, M. & Narayanan, S. (2011a) Rachel: Design of an emotionally targeted interactive agent for children with autism. *International Conference on Multimedia and Expo (ICME), 11–15 July, Barcelona, Spain*.

Mower, E., Lee, C. C., Gibson, J., Chaspari, T., Williams, M. & Narayanan, S. (2011b) Analyzing the nature of ECA interactions in children with autism. *Interspeech 2011, 12th Annual Conference of the International Speech Communication Association (ISCA), 28–31 August, Florence, Italy*, pp. 2989–2993

Paul, R., Augustyn, A., Klin, A. & Volkmar, F. R. (2005) 'Perception and production of prosody by speakers with autism spectrum disorders', *Journal of Autism and Developmental Disorders*, 35(2):205–220.

Paul, R., Bianchi, N., Augustyn, A., Klin, A. & Volkmar, F. R. (2008) 'Production of syllable stress in speakers with autism spectrum disorders', *Research in Autism Spectrum Disorders*, 2(1):110–124.

Peterson, C. C., Garnett, M., Kelly, A. & Attwood, T. (2009) 'Everyday social and conversation applications of theory-of-mind understanding by children with autism-spectrum disorders or typical development', *European Child & Adolescent Psychiatry*, 18(2):105–115.

Philip, R. C., Whalley, H. C., Stanfield, A. C., Sprengelmeyer, R., Santos, I. M., Young, A. W., Atkinson, A. P., Calder, A. J., Johnstone, E. C., Lawrie, S. M. & Hall, J. (2010) 'Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders', Psychological Medicine, 40(11):1919–1929.

Pioggia, G., Sica, M. L., Ferro, M., Igliozzi, R., Muratori, F., Ahluwalia, A. & De Rossi, D. (2007) Human-robot interaction in autism: FACE, an android-based social therapy. *16th IEEE International Symposium on Robot Human Interactive Communication (RO-MAN), 26–29 August, Jeju, South Korea,* pp. 605–612.

Pipitpukdee, J. & Phantachat, W. (2011) The study of the pet robot therapy in Thai autistic children. *5th International Conference on Rehabilitation Engineering & Assistive Technology (i-CREATe), 21–23 July, Bangkok, Thailand, article no. 43*.

Potamianos, A. & Narayanan, S. (2007) A review of the acoustic and linguistic properties of children's speech. *9th IEEE Workshop on Multimedia Signal Processing (MMSP), 1–3 October, Chania, Crete, Greece*, pp. 22–25.

Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D. & Plaza, M. (2011) 'Automatic intonation recognition for the prosodic assessment of language impaired children', *IEEE Transactions on Audio, Speech and Language Processing*, 19(5):1328–1342.

Robins, B., Dautenhahn, K., Te Boekhorst, R. & Billard A. (2005) 'Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?', *Universal Access in the Information Society*, 4(2):105–120.

Rodriguez, W. R. & Lleida, E. (2009) Formant estimation in children's speech and its application for a Spanish therapy tool. *Workshop on Speech and Language Technologies in Education (SLaTE), International Speech Communication Association (ISCA) Special Interest Group (SIG), 3–5 September, Wroxall Abbey Estate, UK*.

Russo, N., Larson, C. & Kraus, N. (2008) 'Audio-vocal system regulation in children with autism spectrum disorders', *Experimental Brain Research*, June, 188(1):111–124.

van Santen, J. P. H., Prud'hommeaux, E. T. & Black, L. M. (2009) 'Automated assessment of prosody production', *Speech Communication,* 51(11):1082–1097.

Scassellati, B. (2005a) Using social robots to study abnormal social development. *5th International Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, 22–24 July, Nara, Japan*, pp. 11–14.

Scassellati, B. (2005b) *Quantitative metrics of social response for autism diagnosis*, 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN), 13–15 August, Nashville (TN) USA, pp. 585–590.

Scassellati, B. (2007) 'How social robots will help us diagnose, treat, and understand autism', *Robotics Research*, *Springer Tracts in Advanced Robotics*, 28:552–563.

Scassellati, B., Admoni, H. & Mataric, M. (2012) 'Robots for use in autism research', *Annual Review of Biomedical Engineering*, 14:275–294.

Schuller, B., Batliner, A., Steidl, S. & Seppi, D. (2011) 'Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge', *Speech Communication (SPECOM), Special Issue: Sensing Emotion and Affect – Facing Realism in Speech Processing*, 53(9–10):1062–1087.

Schuller, B., Steidl, S., Batliner, A., Noth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G. & Weiss, B. (2012) The INTERSPEECH 2012 Speaker Trait Challenge. *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association (ISCA), 9–13 September, Portland,* OR.

Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Robinson, P., Davies, I., Golan, O., Friedenson, S., Tal, S., Newman, S., Meir, N., Shillo, R., Camurri, A., Piana, S., Bölte, S., Lundqvist, D., Berggren, S., Baranger, A. & Sullings, N. (2013a) ASC-Inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions. *1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI), Satellite event of the 8th International Conference on the Foundations of Digital Games (FDG), ACM, SASDG Digital Library, 14 May, Chania, Crete, Greece*.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Salamin, H., Polychroniou, A., Valente, F. & Kim, S. (2013b) The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Interspeech 2013, 14th Annual Conference of the International Speech Communication Association (ISCA), 25–29 August, Lyon, France*, pp. 148–152.

Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Golan, O., Fridenson, S., Tal, S., Newman, S., Meir, N., Shillo, R., Camurri, A., Piana, S., Staglianò, A., Bölte, S., Lundqvist, D., Berggren, S., Baranger, A. & Sullings, N. (2014) The state of play of ASC-Inclusion: Integrated Internet-based environment for social inclusion of children with autism spectrum conditions. *2nd International Workshop on Digital Games for Empowerment and Inclusion (IDGEI), Satellite event of the 19th International Conference on Intelligent User Interfaces (IUI), ACM, 24 February, Haifa, Israel*.

Scott, K. S., Tetnowski, J. A., Flaitz, J. R. & Yaruss, J. S. (2014) 'Preliminary study of disfluency in school-aged children with autism', *International Journal of Language and Communication Disorders*, 49(1):75–89.

Silver, M. & Oakes, P. (2001) 'Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others', *Autism*, 5(3):299–316.

Silverman, L. B., Bennetto, L., Campana, E. & Tanenhaus, M. K. (2010) 'Speech-and-gesture integration in high functioning autism', *Cognition*, 115(3):380–393.

Snow, D. (1998) 'Children's imitations of intonation contours: Are rising tones more difficult than falling tones?', *Journal of Speech, Language and Hearing Research*, 41:576–587.

Stanton, C. M., Kahn Jr., P. H., Severson, R. L., Ruckert, J. H. & Gill, B. T. (2008) Robotic animals might aid in the social development of children with autism. *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), 12–15 March, Amsterdam, Netherlands*, pp. 271–278.

Steidl, S. (2009) *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag: Berlin.

Steidl, S., Batliner, A., Seppi, D. & Schuller, B. (2010) 'On the impact of children's emotional speech on acoustic and language models', *EURASIP Journal on Audio, Speech, and Music Processing, Special Issue on Atypical Speech*, 6.

Tager-Flusberg, H. (1992) 'Autistic children's talk about psychological states: deficits in the early acquisition of a theory of mind', *Child Development*, 63(1):161–172.

Tanaka, J. W., Wolf, J. M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., Brown, C., Stahl, S., Kaiser, M. D. & Schultz, R. T. (2010) 'Using computerized games to teach face recognition skills to children with autism spectrum disorder: The lets face it! Program', *Journal of Child Psychology and Psychiatry*, 51(8):944–952.

Tartaro, A. & Cassell, J. (2008) Playing with virtual peers: Bootstrapping contingent discourse in children with autism. *8th International Conference for the Learning Sciences (ICLS), ACM Press, 23–28 June, Utrecht, Netherlands*, 2:382–389.

Van Lancker, D., Canter, G. J. & Terbeek, D. (1981) 'Disambiguation of ditropic sentences: Acoustic and phonetic cues', *Journal of Speech, Language and Hearing Research,* 24(3):330–335.

Van Lancker, D., Cornelius, C. & Kreiman, J. (1989) 'Recognition of emotional prosodic meanings in speech by autistic, schizophrenic, and normal children', *Developmental Neuropsychology*, 5(2–3):207–226.

Warren, P. (1996) 'Parsing and prosody: An introduction', *Language and Cognitive Processes*, 11(1–2):1–16.

Watts, O., Yamagishi, J., King, S. & Berkling, K. (2010) 'Synthesis of child speech with HMM adaptation and voice conversion', *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1005–1016.

Wells, B. & Peppé, S. (2003) 'Intonation abilities of children with speech and language impairments', *Journal of Speech, Language and Hearing Research*, 46(1):5–20.

Werry, I. & Dautenhahn, K. (1999) Applying mobile robot technology to the rehabilitation of autistic children. *7th International Symposium on Intelligent Robotic Systems (SIRS), 20–23 July, Coimbra, Portugal*, pp. 265–272.

Werry, I., Dautenhahn, K., Ogden, B. & Harwin, W. (2001) 'Can social interaction skills be taught by a social agent? The role of a robotics mediator in autism therapy', *Cognitive Technologies: Instruments of Mind*, 2117:57–74.

Winner, E. (1988) *The Point of Words: Children's Understanding of Metaphor and Irony*. Cambridge, MA: Harvard University Press.

Wöllmer, M., Schuller, B., Batliner, A., Steidl, S. & Seppi, D. (2011) 'Tandem decoding of children's speech for keyword detection in a child-robot interaction scenario', *ACM Transactions on Speech and Language Processing, Special Issue: Speech and Language Processing of Children's Speech for Child-machine Interaction Applications*, 7(4): article 12.

World-Health-Organization 1994, *ICD-10 - International classification of diseases*, World Health Organisation.

Yildirim, S. & Narayanan, S. (2009) 'Automatic detection of disfuency boundaries in spontaneous speech of children using audio-visual information', *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):2–12.

Yirmiya, N., Sigman, M. D., Kasari, C. & Mundy, P. (1992) 'Empathy and cognition in high-functioning children with autism', *Child Development*, 63(1):150–160.