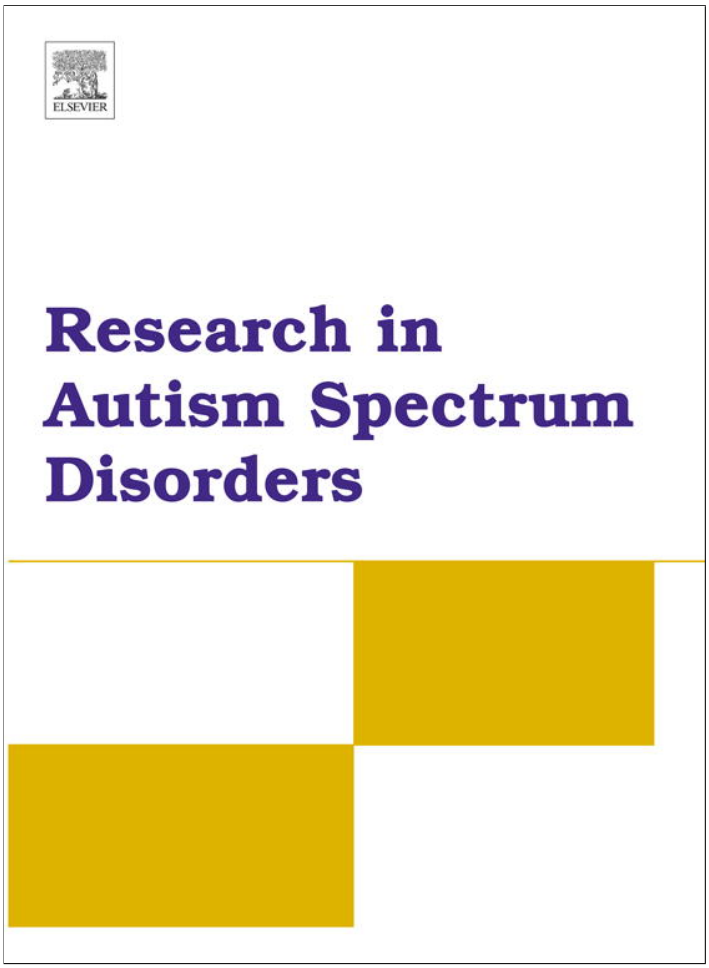


Research in Autism Spectrum Disorders 7 (2013) 741–756

Contents lists available at SciVerse ScienceDirect

Research in Autism Spectrum Disorders

Journal homepage: http://ees.elsevier.com/RASD/default.asp

Assessment of the communicative and coordination skills of children with Autism Spectrum Disorders and typically developing children using social signal processing

Emilie Delaherche [a], Mohamed Chetouani [a], Fabienne Bigouret [b,c], Jean Xavier [c], Monique Plaza [a], David Cohen [a,c,*]

[a] Institute of Intelligent Systems and Robotics, University Pierre and Marie Curie, 75005 Paris, France
[b] University of Paris 8, 93526 Saint-Denis, France
[c] Department of Child and Adolescent Psychiatry, Hôpital de la Pitié-Salpêtrière, University Pierre and Marie Curie, 75013 Paris, France

ARTICLE INFO

Article history:
Received 27 November 2012
Received in revised form 5 February 2013
Accepted 8 February 2013

Keywords:
Social signal processing
Coordination
Imitation
Autism spectrum disorder

ABSTRACT

To cooperate with a partner, it is essential to communicate by sharing information through all available avenues, including hand gestures, gazes, head gestures and naturally, speech. In this paper, we compare the communicative and coordination skills of children with typical development to those of children with Autism Spectrum Disorders (ASDs) in cooperative joint action tasks. Communicative skills were assessed using a pragmatic annotation grid. Coordination skills were assessed based on automatically extracted features that characterize interactive behavior (turn-taking, synchronized gestures). First, we tested the performance of the interactive features when discriminating between the two groups of children (typical vs. ASD). Features characterizing the gestural rhythms of the therapist and the duration of his gestural pauses were particularly accurate at discriminating between the two groups. Second, we tested the ability of these features for the continuous classification problem of predicting the developmental age of the child. The duration of the verbal interventions of the therapist were predictive of the age of the child in all tasks. Furthermore, more features were predictive of the age of the child when the child had to lead the task. We conclude that social signal processing is a promising tool for the study of communication and interaction in children with ASD because we showed that therapists adapt differentially in three different tasks according to age and clinical status.

1. Introduction

Children with Autism Spectrum Disorders (ASD) show impairments in both communication and social interaction. Most of the studies conducted thus far in developmental psychology or child psychiatry have focused on comparing child behaviors to the behavior of another group, behaviors that are responses to specific stimuli or the changes in behaviors over time. However, to cooperate and communicate with a partner, it is essential that information be shared using all available avenues, including hand gestures, gaze, head gestures and naturally, speech. Temporally, the interactive nature of human communication implies that a message a_i produced by A impacts B who, in return, produces message b_i and so on, indicating

* Corresponding author at: Department of Child and Adolescent Psychiatry, Hôpital de la Pitié-Salpêtrière, University Pierre and Marie Curie, 75013 Paris, France. Tel.: +33 01 42 16 23 51; fax: +33 01 42 16 23 53.
E-mail addresses: emilie.delaherche@isir.upmc.fr (E. Delaherche), mohamed.chetouani@upmc.fr (M. Chetouani), fbigouret@yahoo.fr (F. Bigouret), jean.xavier@psl.aphp.fr (J. Xavier), monique.plaza@upmc.fr (M. Plaza), david.cohen@psl.aphp.fr (D. Cohen).

1750-9467/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.rasd.2013.02.003

that some form of coordination occurs between partners A and B (Chaby, Chetouani, Plaza, & Cohen, 2012). In this paper, we compare the communicative and coordination skills of children with typical development and children with ASD in cooperative joint action tasks using social signal processing.

1.1. Automatic analysis of social interactions

Recently, the automatic analysis of human social interactions has received much attention from the scientific community, and multidisciplinary teams have been formed to share their expertise on human sciences, machine learning and signal processing. These themes are the targets of an emerging research domain, social signal processing (SSP) (Vinciarelli, Pantic, & Bourlard, 2009). Primarily, SSP aims to provide computers, robots or Embodied Conversational Agents (ECAs) with social intelligence to facilitate the acceptance of users. The core idea is to produce adaptive, fluent and expressive interfaces. SSP shows potential in multiple applications in the clinical domain, which include predicting symptoms (Kupper, Ramseyer, Hoffmann, Kalbermatten, & Tschacher, 2010), detecting the experience of pain (Ashraf et al., 2007; Lucey, Cohn, Prkachin, Solomon, & Matthews, 2011) and proposing innovative therapeutic partners for children (Kozima, Michalowski, & Nakagawa, 2009) and elderly patients (Bemelmans, Gelderblom, Jonker, & de Witte, 2012).

Crucial problems addressed by SSP are the detection of non-verbal behavioral cues from interaction data and the inference of meta-signals (e.g., emotions, dominance, and synchrony) from those behavioral cues. Lastly, several studies have attempted to automatically infer high-level information about interactional states from low-level features extracted from speech or gestures. Role recognition (Salamin, Mohammadi, Truong, & Vinciarelli, 2010) and dominance detection (Hung, Huang, Friedland, & Gatica-Perez, 2011; Worgan & Moore, 2011) have been the most targeted problems. The simple features of speaking activity (i.e., who talks when), the adjacency of speaker turns and the duration of speech turns have been demonstrated to be efficient features for those recognition problems. Recently, cohesion (Hung & Gatica-Perez, 2010), conversational patterns (Jayagopi & Gatica-Perez, 2010) and interactive communicativity (Rutkowski, Mandic, & Barros, 2007) have also been considered.

1.2. Evidence of interpersonal coordination

Among social signals, synchrony and coordination have recently been considered (Delaherche et al., 2012b; Ramseyer & Tschacher, 2010). Condon et al. initially proposed a micro-analysis of human behavior (body motion and speech intonation) and provided evidence for the existence of interactional synchrony; i.e., the coordination between listener's and speaker's body movements or between the listener's body movements and the speaker's variations in pitch and stress (Condon & Ogston, 1967). Bernieri et al. define coordination as the "... degree to which the behaviors in an interaction are non-random, patterned or synchronized in both form and timing" (Bernieri, Reznick, & Rosenthal, 1988). Kendon raised fundamental questions about the condition of interactional synchrony arousal and its function in interaction (Kendon, 1970). When synchronizing with the speaker, the listener demonstrates his ability to anticipate what the speaker is going to say; thus, the listener gives feedback to the speaker ensuring a smooth conversation.

The double-video system was designed by Nadel et al. to study the sensitivity of infants to synchrony (Nadel, Carchon, Kervella, Marcelli, & Réserbat-Plantey, 1999). Mothers and infants were situated in two different rooms and filmed with synchronized video cameras. They could see each other through video screens. This setting allowed the timing of exchanges to be manipulated by broadcasting live or pre-recorded videos of the mother to the infant. The infant showed more negative signs (manifestations of anger or distress, cries) in the presence of non-contingent signals. Moreover, when the "live" exchanges were reinstated, positive signals (gazes toward the mother, smiles, etc.) were restored. In these experiments, infants demonstrated expectancies for synchronized and contingent exchanges with the social partner (the mother) beginning at two months old. The key role of synchrony at early ages has also been found in more natural early interactions, such as breast-feeding (Viaux-Savelon et al., 2012), and interaction scenes from the home movies of infants who will subsequently develop autism (Saint-Georges et al., 2011).

As part of the set of social signals, interpersonal coordination is a signal of great importance for evaluating the degree of attention or engagement between two social partners. Interpersonal coordination is often related to the quality of interaction (Chartrand & Bargh, 1999), cooperation (Wiltermuth & Heath, 2009) or the feeling of belonging to a social group ("entitativity") (Lakens, 2010). Finally, the assessment of interpersonal coordination constitutes the first step in the process of equipping a social robot with the ability to anticipate the reactions of a human partner and enter into synchrony with that partner (Michalowski, Simmons, & Kozima, 2009; Prepin & Gaussier, 2010).

1.3. Language and coordination in joint action

In this paper, we focused on cooperative joint action tasks in which two partners must build a 3D jigsaw puzzle by alternatively imitating or giving instructions to a partner. Sebanz et al. define joint action as "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" (Sebanz, Bekkering, & Knoblich, 2006). According to these authors, the quality and the effectiveness of two partners performing joint action rely on their abilities to share representations, predict actions, integrate the predicted effects of one's own and other's actions, and communicate. Indeed, joint action and language are inextricably linked. Two forms of

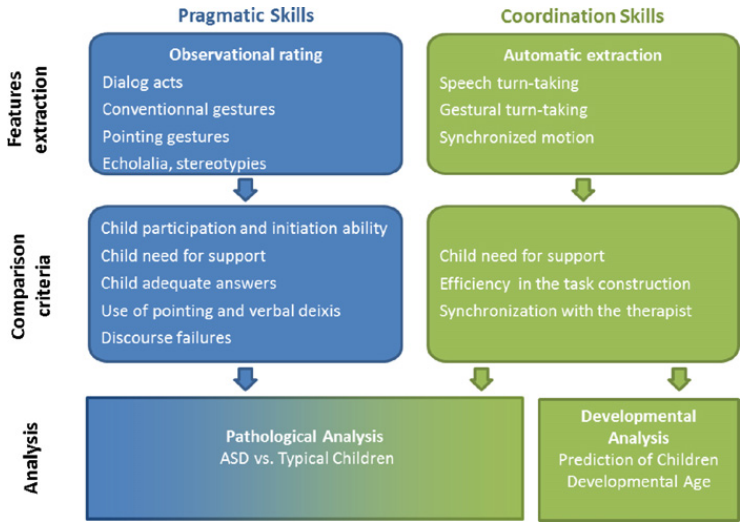


Fig. 1. Synopsis.

coordination are important. Coordination of content relates to what the participant intend to do, here "build a clown with a partner". Coordination of process relates to the "physical and mental states the partners recruit" to perform the joint action, adapt their rhythms to stay at the same step of the assembly, chose the same pieces of puzzle and give instructions to the partner when he is attentive. According to Clark's theory, to achieve coordination of process, the partners need to constantly update their common ground via the available mediums of communication. This process is called grounding and requires more than just sending a message to the partner. Grounding requires "assure[ing] ourselves that it [the message] has been understood as we intended to be."

For instance, Shockley et al. studied the coordination of postural sway between interactional partners performing a cooperation task (Shockley, Santana, & Fowler, 2003) and found that the participants' postural sways were more coordinated when they performed the task (and communicated) together than when they performed the task with a confederate. Moreover, the coordination did not depend on whether the participants could see each other. They hypothesized that coordination operated via language.

1.4. Related works in Autism Spectrum Disorder

This paper is a part of the recent attempts in SSP to provide automatic tools to investigate interaction data in clinical environments. For instance, the USC CARE Corpus was recently proposed to study children with autism in spontaneous and standardized interactions and develop analytical tools to enhance the manual rating tools of psychologists (Black et al., 2011). Tartaro et al. proposed the design of virtual peers to help children acquire communicative skills (Tartaro & Cassell, 2008). These researchers studied the production of contingent discourse in children with ASD in a collaborative task with a virtual peer. The virtual peer was controlled with the Wizard Of Oz methodology and incorporated facilitating features such as yes/no questions or conceptually simple questions to elicit responses from the child. They observed that, compared to an interaction with a human peer, ASD children produced more contingent responses with the virtual peer. Furthermore, over the course of the interaction, the production of contingent responses increased. Recently, Mower et al. (2011) compared the communication patterns of children with autism when they were addressing their parent or an ECA. They found that the communicative patterns did not differ widely in the two conditions, suggesting that the ECA environment could be used to elicit natural interactions with the child.

1.5. Approach

In this paper, we compare the communicative and coordination skills of children with typical development and children with Autism Spectrum Disorder (ASD) in cooperative joint action tasks (Fig. 1). Communicative skills were assessed with a pragmatic annotation grid. Coordination skills were assessed via automatically extracted features characterizing interactive behavior (gestural rhythms, turn-taking, and synchronized gestures).

Several hypothesis were formulated regarding the communicative strategies of ASD children compared to typical children:

- (H1.1) When addressing the therapist, children with ASD make more frequent use of inadequate dialog acts or gestures.
- (H1.2) When addressing the therapist, children with ASD need more support from the therapist.
- (H1.3) Time latencies are longer for children with ASD.

Table 1
Participants characteristics.

n	Sex	Chronological age (years.months)	Developmental age[a] (years.months)	ADI-R scores[b]				ICD-10 diagnosis[c]
				Social	Comm verbal/ non-verb	Stereotypies	First signs < 3 years	
ASD (*N* = 7)								
1	M	9	8.6	13	13/6	3	5	AD
2	F	10.3	5	25	20/10	9	4	AD
3	M	9	4.6	23	23/11	9	4	AD
4	M	8.6	5	18	9/8	1	3	PDD-NOS
5	M	7.11	7.6	13	6/3	1	2	PDD-NOS
6	M	10.1	8	10	5/3	1	3	PDD-NOS
7	M	11.3	7.6	13	8/3	4	5	AD
Typically developing controls (*N* = 14)								
			Sex			Age		
Matched for age and sex (2/1)			2F,12M			Mean (SD[d]) [range]: 7.1 (1.11) [4.4–8.9]		

[a] Assessed with Vineland Developmental Score (Sparrow et al., 2005) or the PsychoEducational Profile-Revised.
[b] ADI-R, autism diagnostic interview-revised.
[c] AD, autistic disorder; PDD-NOS, pervasive developmental disorder-not otherwise specified.
[d] SD = standard deviation.

Regarding coordination, we formulated the hypothesis that ASD children would differ from children with typical development in the following ways:

- (H2.1) Children with ASD would need more intervention from the therapist
- (H2.2) Children with ASD would lack rhythm and require more gestures and more time to perform the task
- (H2.3) Children with ASD would be less synchronized with the therapist

Furthermore, due to the large difference in developmental ages (4–9 years) across the children, we hypothesized that developmental factors would impact performance on the task. Thus, we tested the impact of developmental age on the same hypotheses:

- (H3.1) Younger children would need more intervention from the therapist.
- (H3.2) Younger children would lack rhythm and would require more gestures and more time to perform the task.
- (H3.3) Younger children would be less synchronized with the therapist.

2. Methods

2.1. Participants

The protocol was approved by the local ethical committee. All parents received information on the experiment and gave written consent before participation of their child. Twenty-one children participated to the study (developmental ages = 4–9 years); 7 of the children were followed in the day-care hospital la Pitié-Salpétrière for Autism Spectrum Disorders. Those children suffered from various social impairments including language disabilities, poor communicative skills, and gestural impairments. Fourteen typically developing children were recruited from the first (or primary) school Marie Noël of Montigny Le Bretonneux. Controls met the following inclusion criteria: no verbal communication impairment, no mental retardation, and no motor, sensory or neurological disorders. Controls were matched to the children with ASD for developmental age and gender (2 typical children for 1 ASD child). The developmental ages of the ASD children were assessed with the Vineland Developmental Score (Sparrow, Cicchetti, & Balla, 2005) or the PsychoEducational Profile-Revised. For the control group, the developmental and chronological age were considered to be the same. Characteristics of the participants are summarized in Table 1.

2.2. Procedure

The children were asked to perform three different construction tasks: the ‘‘Imitation’’ task, the ‘‘Child Follows Instructions’’ task and the ‘‘Child Gives Instructions’’ task. These tasks increased in difficulty in terms of communication, but they were simple in terms of motor and cognitive abilities. The ‘‘Imitation’’ and ‘‘Child Gives Instructions’’ tasks consisted of building a clown with 7 polystyrene elements (2 hands, 2 legs, body, head and hat). The ‘‘Child Follows Instructions’’ task

(a) Task 1 : "Imitation" (b) Task 2 : "Child Follows Instructions" (c) Task 3 : "Child Gives Instructions"

Fig. 2. Experimental setup.

consisted of building a frog with 6 polystyrene elements (2 legs, body, head and 2 eyes). The child sat across from the therapist, and the polystyrene elements were arranged on a table in front of them (Fig. 2).

In the "Imitation" task, the therapist led the task and showed the child each step of the assembly. Speech interventions from the therapist were limited to the times at which the child encountered difficulties. The children were asked to "do as the therapist" to encourage imitation. The goal of this task was to test the ability to imitate a partner.

During the "Child Follows Instructions" and "Child Gives Instructions" task, a folding screen was introduced between the child and the therapist to prevent the partners from seeing each others' gestures; however, they could still gaze at and see each others' faces. In the "Child Follows Instructions" task, the therapist led the task and explained to the child how to construct the frog. The therapist answered the possible questions of the child but did not answer to visual solicitations from the child (for example, attempts to show the frog above the folding screen). This task aimed to test the ability of the child to follow verbal instructions without visual content.

In the "Child Gives Instructions" task, the child led the interaction. The child explained to the therapist how to construct the clown. The therapist only intervened to support and elicit instructions from the child ("What shall I do next?", "Where do I put the blue piece?", etc.). As the same construction had already been performed during the "Imitation" task, the goal was to test the ability of the child to address injunctions to a third party.

The interactions were recorded using a single camera placed above the participants. Audio recordings were collected at 48 kHz and the video recordings at 25 fps. The participants were equipped with color bracelets to facilitate tracking of their hands. The duration of the tasks were as follows (mean, standard deviation, total duration for all participants): Task1 (1:53, 1:09, 40:00 mn), Task2 (1:57, 1:05, 41:00 mn) and Task3 (2:03, 1:51, 53:00 mn).

2.3. Pragmatic skills analysis

Audio data were annotated with the Anvil annotation tool (Kipp, 2008) by two speech therapists to segment the speakers' speech turns and annotate the dialog acts. Therapists' and children's utterances were labeled according to the following categories: initiative assertions, questions, retorts, answers, orders/requests, expressive assertion. Moreover, the children's answers were labeled according to their adequacy; thus, these categories also took into account the unanswered questions. Conventional gestures (head nods or head shakes) and pointing gestures were also annotated. Characteristic expressions of autism such as echolalia (the automatic repetition of vocalizations made by the dialog partner) and stereotypies (a repetitive or ritualistic movement) were added to the grid (Appendix Table A1). In total, 742 speech turns for the children, 1715 speech turns for the therapist, and 139 conventional or pointing gestures of the children were annotated.

Cohen's kappa (Cohen, 1960) between the two annotators was calculated for each dyad, each task and each item of the grid (Appendix Table A2). For all items, the kappa values were between 0.74 and 1; kappa values between 0.61 and 0.81 are considered to indicate substantial agreement, kappa values above 0.81 indicate almost perfect agreement. We present the confusion matrices for the children's and therapists' utterances for all tasks in Appendix Fig. A1. For the therapist, the main confusion concerned "Initiative Assertions" and "Follow-up Assertions"; the annotators disagreed on whether the topic of the assertion was new or previously introduced. For the children, the main source of confusions concerned the adequacy of some answers containing verbal deixis when the folding screen was present. Deixis refer to words or phrases whose meanings require contextual information to be understood (e.g., here, there). The annotators disagreed on whether these answers were induced and unexpected or inadequate. Finally, the segments for which the annotators disagreed were discarded, and only the intersection between the two annotations was kept for further analysis.

From the annotation, five criteria were deduced to evaluate and compare the children's abilities to cooperate and communicate with the partner. These criteria corresponded to the consolidation of selected items from the annotation:

- The child's participation and initiation ability: the number of initiative assertions, requests and questions from the child.
- The need for support: the number of questions from the therapist.
- Adequate answers: the number of adequate answers (verbal or conventional head gestures) given by the child.
- The use of pointing and verbal deixis.
- Discourse failure: the number of digressions, echolalias, unanswered questions, and inadequate answers from the child.

2.4. Coordination automatic assessment

We also sought to extract the following low-level automatic features to characterize the child-therapist interactive behavior during the tasks: speech turn-taking, gestural turn-taking and synchronized movements.

2.4.1. Speech turn-taking features

Based on the manual segmentation of the speakers' turns with Anvil, we extracted several features to describe the alternance of speech turns during the task. First, we extracted all the continuous time segments when the child or the therapist was speaking and all the time segments when neither the child nor the therapist was speaking. We calculated standard statistics (mean, median, standard deviation, range, minimum and maximum) on the durations of the speaking segments ($\text{Child}_{Statname}$ and $\text{Therapist}_{Statname}$) and on the duration of the pause segments ($\text{Pause}_{Statname}$) to gather information on the duration of the participants' utterances and their variations. Was there a majority of backchannels or an alternance of backchannels with more complex utterances (explanations, requests, orders, etc.)? We also measured the percentages of interactional time in which neither participant was talking (Pause_{Ratio}), one of the participants was talking (Therapist_{Ratio} and Child_{Ratio}) and when both participants were talking at the same time (Ovlp_{Ratio}).

Finally, we added several features to evaluate whether there was intermodal synchrony between the partners' vocal features and their gestures. $\text{Child}^{IntraSync}$ and $\text{Therapist}^{IntraSync}$ measure the percentage of interactional time when the child or the therapist was gesturing and speaking at the same time. We also measured whether the child or the therapist tended to gesture while the other partner was speaking ($\text{InterSync}^{GestTherapist}_{SpeechChild}$ and $\text{InterSync}^{GestChild}_{SpeechTherapist}$).

2.4.2. Gestural turn-taking features

In our three tasks, we should observe alternations in gestural turns between the therapist and the child. In the "Imitation" task, the child needed to look at the demonstrator and then reproduce the same actions. Then, the demonstrator waited until the child finished before showing him the next stage of the assembly. In the "Child Follows Instructions" task, the therapist described what he was assembling and then waited for the child to perform the same assembly before moving on the next step. The "Child Gives Instructions" task was similar to the second task except that the roles of the child and therapist were reversed. We assumed that constant and smooth alternations of turns (constant gestural sequences, shorter pauses) would accompany more coordinated dyads.

We tracked the participants' hand trajectories with the coupled Camshift algorithm (Bradski, 1998). The participants were equipped with salient color bracelets to facilitate following of their gestures. We then compressed the x and y Cartesian coordinates to the polar coordinate $r = \sqrt{x^2 + y^2}$ and derived r to obtain the velocities of the hands.

Based on the hand velocities, we extracted a binary feature that was set to 1 if the hand velocity was above a threshold (participant assembling) and 0 otherwise (participant still). We performed morphological post-processing to remove isolated 1 (cleaning) and connect continuous gestures (dilatation).

From the binary features of the child and the therapist, we extracted four features to depict gestural turn-taking:

- $^{ChildOn}_{TherapistOff}$ ratio: the percentage of time in which the child was gesturing and the therapist was still.
- $^{ChildOff}_{TherapistOn}$ ratio: the percentage of time in which the therapist was gesturing and the child was still.
- $^{ChildOn}_{TherapistOn}$ ratio: the percentage of time in which the therapist and child were gesturing at the same time.
- $^{ChildOff}_{TherapistOff}$ ratio: the percentage of time in which the therapist and child were still at the same time.

Moreover, the tasks were particularly repetitive (successive assembling and observing phases), so we can assume that a rhythm was established during the task. All pause segments should last approximately the same duration as should the gesturing segments. A deviation of the duration of the segments could be perceived as a disruption of the fluency of the interaction. Thus, we extracted several statistics on the durations of the gestural and pause segments:

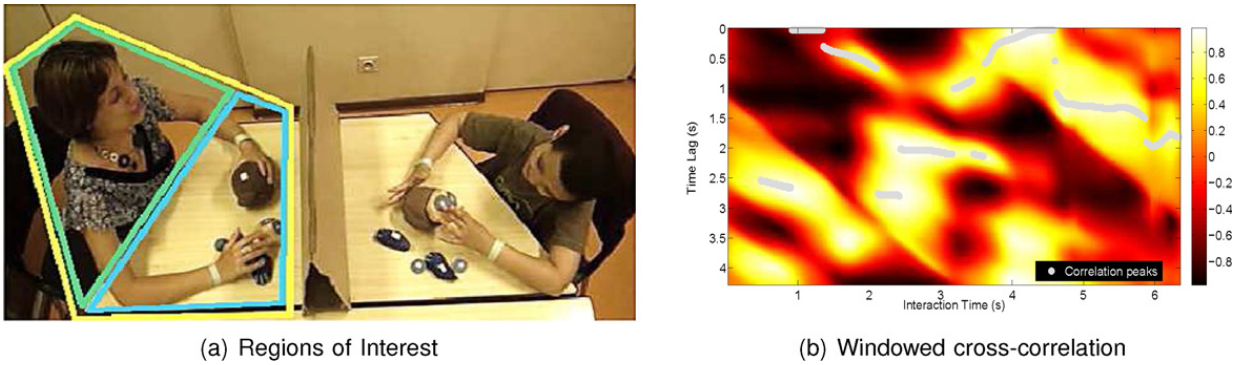
- Gestural pause ($\text{Therapist}^{Pause}_{Statname}$ or $\text{Child}^{Pause}_{Statname}$) and segment ($\text{Therapist}^{Gest}_{Statname}$ or $\text{Child}^{Gest}_{Statname}$) durations. From the pause and gestural segments, we calculated standard statistics on the durations of the segments (mean, median, standard deviation, range, minimum and maximum).
- Pause ratio ($\text{Therapist}^{Pause}_{Ratio}$ or $\text{Child}^{Pause}_{Ratio}$). We measured the percentage of interaction time in which the participant was still.

Lastly, we simply counted the number of continuous sequences of movements for each participant ($\text{Therapist}^{Gest}_{Nb}$ and Child^{Gest}_{Nb}) and the ratio between the number of sequences for each participant ($\text{TherapistChild}^{Gest}_{Ratio}$).

2.4.3. Synchronized motion features

We also sought to measure the coordination between the child and the therapist as the degree of similarity between their motion features.

Thus, we extracted the global movement (Global) of each partner, their postural movement (Posture) and the movement of their hands (Hands). For each feature, a distinct Region Of Interest was defined for the child and the therapist, and the



(a) Regions of Interest (b) Windowed cross-correlation

Fig. 3. Regions of interest (a). The yellow line delineates the global movement space, the green line the postural movement space and the blue line delineates the hands movement space. Windowed cross-correlation (b). The grey dots represent the peak found by the peak picking algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

movement in the ROI was estimated with Motion Energy Image. An example ROI is given in Fig. 3(a). We also took into account the hand velocities obtained with the tracking algorithm (RHand and LHand).

Then, we computed similarity measures (windowed cross-correlation) between the features of the therapist and the child in windows of 1.5 s with a maximum lag of 5 s. The lags were always applied between the partner who led the task and the one who followed. The matrices from the windowed cross-correlation analysis were submitted to a peak-picking algorithm to calculate peak correlations nearest a lag of zero and their respective time lags (Boker, Xu, Rotondo, & King, 2002). We fixed the size of the search region to force a succession of smaller values on each side of the peaks, thus ensuring that a peak was not a local maximum. The size of the search region was fixed at 0.75 s. An example of this peak-picking is shown in Fig. 3(b), which shows three different peaks for time lags of 0.74 s ($r_{peak} = 0.5$), 1.8 s ($r_{peak} = 0.69$) and 2.8 s ($r_{peak} = 0.81$) at time $t = 2.2$ s. The search region helped to find the larger maximum located at a time lag of 2.8 s.

To measure the average strength and the variation of synchronized sequences, we extracted the means ($\text{Feature}^{Mean}_{Peak}$) and standard deviations ($\text{Feature}^{Peak}_{Std}$) of the peak correlations for each task and each dyad. To take into account the delay between partners, we also measured the mean ($\text{Feature}^{Lag}_{Mean}$) and standard deviation ($\text{Feature}^{Lag}_{Std}$) of the lags associated with the peak correlation.

2.5. Statistical analysis and classification computing

To compare typically developing versus ASD children in the five criteria obtained from the pragmatics annotation grid, we used Mann–Whitney non-parametric tests and a significance threshold of $p \leq 0.05$.

To compare typically developing versus ASD children in the automatic features obtained in Section 2.4, we used binary classifiers. The classification results were obtained with an SVM classifier (linear kernel) and a Leave-One-Out Cross-Validation approach (i.e., the classifier was trained on all but one dyad and tested on the last dyad). We also estimated the discriminative power of the automatically extracted features to predict which children were in the clinical group (ASD vs. typical) based on each task separately. First, we estimated the SVM discriminant function f from the training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, composed of the vectors of the observations $x_i \in X$ and the corresponding labels $y_i \in Y = \{1, 2\}$. Then, we modeled the distribution of the hidden states y given the value of the discriminant function $f(x)$ with a sigmoid function:

$$P(y = 1 | f(x)) = \frac{1}{1 + \exp(af(x) + b)}$$

The parameters a and b were fit using maximum likelihood estimation from the training set ($f(x_i)$, y_i) (Platt, 1999). To evaluate the discriminative power of each feature, we estimated the a posteriori probability of each observation belonging to its own class.

In Section 3, Table 3(a)–(c) presents the classification accuracy (in %), the mean and standard deviation of these probabilities across all observations from the database according to each task. To limit table size and to focus on the most meaningful results, we only present the features with classification accuracies greater than 75%. A right-tailed t-test was performed to determine whether the means of the features of the control group were lower than those of the ASD group (↗), and a left-sided t-test was performed search for the opposite trend (↘).

Finally, we trained continuous classifiers (SVR) to predict the developmental age of all the children ($N = 21$) based on the features proposed in Section 2.4 regardless of their clinical status. For ASD children, we used the developmental age, and for the typical children, we used chronological age. The ages were normalized between 0 and 1. The performance of the classifier was evaluated with a Leave-One-Out Cross-Validation approach. We would like to distinguish our approach from current works on age recognition. Age recognition aims at inferring the chronological bracket age of a person from utterance-level acoustic features (Ajmera & Burkhardt, 2008; Metze et al., 2007; Wolters, Vipperla, & Renals, 2009) or facial

Table 2
Participants pragmatic differences according to clinical status (ASD vs. Controls) and tasks.

Task	"Imitation"	"Child Follows Instructions"	"Child Gives Instructions"
Participation	=	=	=
Discourse failure	↗[a]	↗	↗
Adequare answers	=	=	↘
Pointing and verbal deixis	NA	=	=
Time latency	↗	↗	↗
Need for support from adults	↗	↗	↗

[a] NA, not appropriate; ↗ means ASD > Controls; ↘ means ASD < Controls; = means ASD = Controls.

images (Fu, Guo, & Huang, 2010). Knowledge of the speaker's age is particularly useful for adapting multimedia communication systems (degree of automation, waiting queue music) to the age of the person. In our approach, we infer the developmental age of the child based on his interactive behavior (turn-taking or coordination behavior). We argue that the child's interactive behavior depends on his/her developmental age. We also argue that the therapist should adapt his communicative and gestural behavior according to the developmental age of the child.

In Section 3, we present the correlation coefficients (CCs) and mean linear errors (MLEs) in Table 4(a)–(c). We only present the features for which the classifier correlation coefficient performance was greater than 0.3.

3. Results

3.1. Pragmatic skills analysis

Table 2 presents the results of the pragmatic annotation grid according to clinical status (ASD vs. controls) and task. We found the following: (1) when addressing the therapist, children with ASD made more frequent use of inadequate dialog acts or gestures; and (2) time latencies were longer for children with ASD. Our hypotheses H1.1 and H1.3 were verified. In all tasks, children with ASD needed more support from the therapist.

3.2. Prediction of children's clinical diagnoses based on single features

The following Table 3(a)–(c) presents the classification accuracies (in %) and discriminative powers of the single features in predicting the clinical status (ASD vs. control) of the children. Table 3(a) presents the classification results for the "Imitation" task. Regarding hypothesis (H2.1), the amount of intervention of the therapist was not a good feature for predicting the child's diagnosis. In contrast, the ratio of pauses and the ratio of child interventions were good predictors of the child's diagnosis. Several features extracted from the children's gestures performed well in discriminating the two groups of children. A larger range of movement durations for a child may indicate a lack of rhythm (as every move of the puzzle takes approximately the same amount of time). The mean duration of the phases of movement of the children performed well in discriminating the two groups of children. This measure tended to be larger for the ASD group, which is in line with our hypothesis H2.2. We also found that the clinician's gestural features helped to predict the child's clinical diagnosis. These results show that the therapist adjusted his behavior to that of the child. Finally, features from synchronized movements did not perform well in discriminating the two groups in this task.

Table 3(a) presents the classification results for the "Child Follows Instructions" task. On this task, the following features based on speech turn-taking performed in discriminating the two groups: the mean and standard deviation of the duration of the pause, the volume of speech of the therapist and the maximum length of the therapist's intervention. These results validate our hypothesis H2.1, which states that the therapist would need to intervene more with children with ASD. We did not find any feature based on child turn-taking that helped to predict the two groups. For the gestural features, only the mean duration of the pauses of the therapist helped to discriminate the two groups. This mean duration was larger for ASD children. Two features based on synchronized movements performed well. These features characterized the variation of the correlation peak value for postural and global movement.

Table 3(b) presents the classification results for the "Child Gives Instructions" task. We did not find any feature based on speech turn-taking that correctly predicted the clinical diagnosis of the child. Regarding gestures, we found that only therapist features were good predictors of the child population group. These features characterized the duration of the gestural phases of the therapist (mean, range standard deviation and max). It is quite interesting to see that the behavior of the therapist seems strongly impacted in this task, possibly according to the instructions of the child. For synchronized movements, we found that the correlation peak's mean value and standard deviation performed well in discriminating the two groups. This finding shows that the children in the typical group tended to be more in sync with the therapist's movement despite the folding screen. Moreover, the variation in the time-lag between the two partners was also a good predictor.

Table 3
Features performance. The following tables present the classification accuracy (in %) and the discriminative power of the features. We only present the features with a classification accuracy greater than 75%. A right-tailed *t*-test is performed to find out if the mean of the features of the control group is lower than the one of the ASD group (↗) and a left-sided *t*-test is performed to seek for an opposite trend (↘). If a significant trend exists, it is indicated in the last columns.

Feature	Acc (%)	Discriminative power Mean (Std)	↘↗
(a)"Imitation" task			
Speech			
$Pause_{Ratio}$	85.7	0.62 (0.23)	↘
$Child_{Ratio}$	76.2	0.64 (0.23)	↗
Gestures			
$Child^{Pause}_{Ratio}$	90.5	0.68 (0.22)	↘
$Child^{Gest}_{Median}$	81	0.62 (0.22)	↗
$Child^{Gest}_{Range}$	81	0.72 (0.23)	↗
$Child^{Gest}_{Max}$	81	0.73 (0.23)	↗
$Child^{Gest}_{Mean}$	76.2	0.67 (0.23)	↗
$Child^{Pause}_{Mean}$	76.2	0.65 (0.24)	↗
$Therapist^{Pause}_{Range}$	76.2	0.6 (0.22)	↗
$Therapist^{Pause}_{Max}$	76.2	0.6 (0.22)	↗
(b) "Child Follows Instructions" task			
Speech			
$Therapist_{Ratio}$	95.2	0.75 (0.21)	↗
$Therapist_{Range}$	85.7	0.68 (0.24)	↗
$Pause_{Mean}$	76.2	0.72 (0.22)	↘
$Pause_{Std}$	76.2	0.66 (0.23)	↘
$Pause_{Median}$	76.2	0.72 (0.22)	↘
$Therapist_{Max}$	76.2	0.67 (0.24)	↗
Gestures			
$Therapist^{Pause}_{Mean}$	81	0.59 (0.22)	↗
Synchronized movements			
$Posture^{Peak}_{Std}$	85.7	0.69 (0.23)	↗
$Global^{Peak}_{Std}$	76.2	0.68 (0.23)	↗
(c) "Child Gives Instructions" task			
Gestures			
$Therapist^{Pause}_{Std}$	81	0.75 (0.23)	↗
$Therapist^{Pause}_{Mean}$	76.2	0.6 (0.24)	↗
$Therapist^{Pause}_{Max}$	76.2	0.72 (0.25)	↗
$Therapist^{Pause}_{Range}$	76.2	0.72 (0.25)	↗
$Therapist^{Nb}_{Gest}$	76.2	0.6 (0.21)	↗
Synchronized movements			
$Posture^{Peak}_{Mean}$	95.2	0.79 (0.19)	↘
$Global^{Peak}_{Std}$	90.5	0.74 (0.23)	↗
$Posture^{Peak}_{Std}$	85.7	0.79 (0.19)	↗
$Global^{Peak}_{Mean}$	81	0.73 (0.23)	↘
$Global^{Lag}_{Std}$	81	0.69 (0.23)	↗

In summary, our hypotheses were partially validated. The speech turn-taking cues only performed well in the "Child Follows Instructions" task, showing that the volume of intervention from the therapist can help discriminate ASD and typical children. However, these cues did not perform well in the "Imitation" or the "Child Gives Instructions" tasks. The features based on gestural turn-taking performed interestingly in the "Imitation" and "Child Follows Instructions" tasks. These performances showed that the features that characterized rhythm (variation in the duration of gestures, variation in the duration of pauses), the duration of the movement or the number of gestures were efficient in discriminating ASD and typical children in all tasks. Surprisingly, for the "Child Follows Instructions" and "Child Gives Instructions" tasks, we were able to discriminate between the groups of children based on the therapist's gestural features. This finding shows that the therapists' adapted to the behavior of the children. Finally, features based on synchronized gestures were more efficient in the "Child Follows Instructions" and "Child Gives Instructions" tasks. We hypothesize that the presence of the folding screen increased the complexity of the synchronization between the two partners.

Table 4
Prediction of the child developmental age. The performance of the classifiers was evaluated with a Leave-One-Out Cross-Validation approach. We present the correlation coefficient (CC) and the mean linear error (MLE). We only present in these tables the features with CC $\geq$0.3.

Feature	CC	MLE
(a)"Imitation" task		
Speech		
Therapist_{Min}	0.48	0.27
Pause_{Mean}	0.43	0.29
Therapist_{Ratio}	0.3	0.3
Gestures		
$\text{Therapist}_{Nb}^{Gest}$	0.49	0.27
Child_{Nb}^{Gest}	0.37	0.29
(b) "Child Follows Instructions" task		
Speech		
$\text{Therapist}_{Median}$	0.34	0.27
Gestures		
$\text{Therapist}_{Mean}^{Pause}$	0.38	0.3
$\text{Therapist}_{Std}^{Pause}$	0.35	0.31
$\text{Therapist}_{Max}^{Pause}$	0.33	0.29
$\text{TherapistChild}_{Ratio}^{Gest}$	0.31	0.3
$\text{Therapist}_{Range}^{Pause}$	0.3	0.31
(c) "Child Gives Instructions" task		
Speech		
Therapist_{Std}	0.71	0.2
Therapist_{Range}	0.66	0.22
$\text{Therapist}_{Median}$	0.53	0.26
Therapist_{Max}	0.51	0.26
Child_{Ratio}	0.45	0.27
Therapist_{Ratio}	0.36	0.29
Gestures		
$\text{Therapist}_{Ratio}^{Pause}$	0.56	0.24
$\text{Child}_{Median}^{Pause}$	0.55	0.25
$\text{Therapist}_{Mean}^{Pause}$	0.46	0.27
$_{TherapistOn}^{ChildOff}\text{Ratio}$	0.46	0.27
$\text{Therapist}_{Median}^{Pause}$	0.44	0.28
$\text{Child}_{Mean}^{Pause}$	0.34	0.3
Synchronized movements		
$\text{RHand}_{Mean}^{Peak}$	0.47	0.28
$\text{Hands}_{Mean}^{Peak}$	0.35	0.3

3.3. Prediction of children's developmental ages based on single features

In the last section, we found several features that could discriminating ASD and typical children, showing that this pathological condition affected ASD children during task completion. However, given the developmental age range of the recruited children (4–9 years), we also hypothesized that clinical diagnosis is not the only factor that could impact the completion of the task. Performance of the task was certainly affected by the developmental age of the child in terms of coordination, autonomy and initiative capabilities. Thus, we intended to identify the best features with which to infer the developmental age of the child. For the typical children, developmental age was considered equal to the chronological age. For the ASD children, developmental age was assessed with the Vineland scale. Chronological and developmental ages are reported in Table 1. The following Table 4(a)–(c) presents the correlation coefficients (CCs) and the mean linear errors (MLEs) of the single features in predicting the developmental age of the child. We only present the features for which the classifier correlation coefficient performance was greater than 0.3.

In the "Imitation" task, we found that the best speech features with which to infer the age of the child were the pause duration, Pause_{Mean}, and the proportion of the therapist utterances, Therapist_{Ratio}. Thus, the quantity of verbal information necessary to perform the task also differed according to the age of the child. The number of gestures necessary to build the puzzle (i.e., $\text{Therapist}_{Nb}^{Gest}$ and Child_{Nb}^{Gest}) were the best predictors of the age of the child.

In the ''Child Follows Instructions'' task, it was essentially the gestural behavior that correctly predicted the ages of the children. In particular, the duration of the therapist pause ($Therapist^{Pause}_{Mean}$) was, on average, larger and varied ($Therapist^{Pause}_{Std}$ and $Therapist^{Pause}_{Range}$) more for younger children. The child/therapist proportion of gestures in building the puzzle also performed well.

In the ''Child Gives Instructions'' task, the statistics (standard deviation, range, median and max) of the duration of the therapist utterances performed well in predicting the age of the child. The ratio of the utterances of the child ($Child_{Ratio}$) and the therapist ($Therapist_{Ratio}$) were also good predictors of the age of the child. Regarding gestural features, the pause durations of the therapist (ratio, mean duration and median duration) were larger when they were interacting with younger children. We also found that the percentage of time that the therapist spent constructing the puzzle while the child stayed still was informative regarding the age of the child. Finally, the performance of the average of the peaks of correlation for hand movements was informative.

In summary, in the ''Imitation'' task, the number of gestures required to complete the task predicted the developmental age of the child. The children of younger developmental age were less fluent (more small manipulations, more attempts). For the ''Child Follows Instructions'' task, the best predictors of the developmental age of the child were those that characterized their gestures; we found that the best features for discriminating between the pathological and typical group were verbal features. For the ''Child Gives Instructions'' task, the developmental age of the child clearly influenced the vocal and gestural behavior of the therapist. The features that predicted the developmental age of the child with $CCs > 0.3$ were clearly more numerous in the ''Child Gives Instructions'' task compared to the ''Imitation'' and ''Child Gives Instructions'' tasks compared to the ''Child Gives Instructions'' task. These results show the increasing difficulty of the three tasks and the wider gap between 4 and 8 year olds on the ''Give Instructions Task''.

Table 5
Synthesis of features performance. Pathological factor and developmental factor.

	''Imitation''	''Child Follows Instructions''	''Child Gives Instructions''
Speech			
Volume of child interventions		ASD > Controls	↘ with decreasing dev[a] age
Volume of therapist interventions	↗ with decreasing dev age	ASD > Controls	↗ with decreasing dev age
Volume of pause	ASD < Controls		
Duration of child interventions			
Duration of therapist interventions	[min]: ↘ with decreasing dev age	[max]: ASD > Controls [med]: ↘ with decreasing dev age	[med, max]:: ↗ with decreasing developmental age
Duration of pause	[mean]:: ↘ with decreasing dev age	[med]: ASD < Controls	
Variation of child utterances duration			
Variation of therapist utterances duration		[range]: ASD > Controls	[range, std]: ↗ with decreasing dev age
Variation of pause duration		[std]: ASD < Controls	
Gestures			
Volume of child gestures	↗ with decreasing dev age		
Volume of therapist gestures	↗ with decreasing dev age		ASD > Controls
Volume of child pauses	ASD < Controls		
Volume of therapist pauses			↗ with decreasing dev age
Duration of child gestures	[med, max, mean]: ASD > Controls		
Duration of therapist gestures			
Duration of child pauses	[mean]: ASD > Controls		[mean]: ↘ with decreasing dev age
Duration of therapist pauses	[max]: ASD > Controls	[mean]: ASD > Controls [mean, max]: ↗ with decreasing dev age	[mean, max]: ASD > Controls [med, mean]: ↗ with decreasing dev age
Variation of child gestures duration	[range]: ASD > Controls		
Variation of therapist gestures duration	[range]: ASD > Controls	[range, std]: ↗ with decreasing dev age	[range, std]: ASD > Controls
Synchronized movements			
Peak strength			[Post, Glob]: ASD < Controls [Rhand, Hands]: ↘ with decreasing dev age
Peak variation		[Post, Glob]: ASD > Controls	[Post, Glob]: ASD > Controls
Lag			
Lag variation		[Glob]: ASD > Controls	

[a] dev, developmental.

4. Discussion

4.1. Synthesis of the current results

We present in Table 5 the synthesis of the results obtained in the current study. Globally, the duration of the therapist pause was a good predictor of the clinical group compared to the control group. The variation in the therapist rhythm (variation of the duration of the pauses) was also a good predictor of the clinical group in the ''Imitation'' and ''Child Gives Instructions'' tasks. The duration of the gestural pauses tended to be larger for the ASD group. Features characterizing synchronized movements were more suitable in the tasks in which the folding screen was present (''Child Follows Instructions'' and ''Child Gives Instructions''). Indeed, it is more difficult to synchronize with a partner without visual feedback. From this table, we can also see that the features that were suitable for discriminating the groups according to clinical status were not necessarily adequate for predicting the developmental age of the child. For instance, the number of gestures necessary to assemble the puzzle was larger for younger children than for older children. This feature was not accurate in discriminating ASD children from controls. Another example is the length of the intervention of the therapist and the variation of the duration of the therapist intervention. These features were not discriminative in terms of clinical status, but they clearly differed according to the developmental age of the child between the ''Imitation'' and ''Child Gives Instructions'' tasks. Younger children appeared to need more support than older ones.

We also note that the results from the automatic features were not always similar to the pragmatic analyses regarding the participation of the child or the interventions of the therapist. In the pragmatic grid, the participation of the child and need for support items aggregated a selection of dialog acts. In the automatic analysis, these features globally quantified the duration of speech interventions, regardless of the dialog act type. Finally, from Table 5, it appears that the features based on the therapists' verbal and non-verbal behavior were as informative (discriminative or predictive), if not more so, than the features characterizing the children's behavior. Thus, the therapists' adaptations to the children were the most helpful in discriminating the two groups.

4.2. Implications for ASD children

In this paper, we performed an experiment with children of different ages and different levels of communication and social skills. As expected, we found progressions in the way children performed the tasks and in the number of features that discriminated the groups or developmental ages according to the difficulty of the task. Indeed, we found that the features that predicted the developmental age of the child the best were from the ''Child Follows Instructions'' task compared to the ''Imitation'' task and on the ''Child Gives Instructions'' task compared to the ''Child Follows Instructions'' task. Similarly, when we assessed, within the ASD group, the children with AD or PDD-NOS, separately, we also found the same progression. Children with PDD-NOS that were clinically less severe than children with AD (Table 1) performed better on the tasks. Further, the number of features that discriminated them from typically developing matched controls were less numerous than those found to classify AD children from matched controls. The current results also have meaningful clinical implications. As shown by the pragmatic grid that we used, most of the discriminating cues between ASD children and controls were based on linguistic characteristics. Language and communication impairment in ASD, especially in AD, has led to numerous studies over the last decades that have tried to specify profiles. Language in autism, when present, may display several varying subtypes within the spectrum. Some individuals may have any of the following structural language disturbances: (i) delayed phonology (Kjelgaard & Tager-Flusberg, 2001); (ii) poor comprehension skills (which may occasionally include greater impairments in expressive skills) (Rapin & Dunn, 1997; Tager-Flusberg, 1981), and (iii) immature syntax and increased prevalence of syntactic errors (Kjelgaard & Tager-Flusberg, 2001). Functional deficits are characterized by the following: (i) a core pragmatics disorder (defined as the ability to use and understand the rules governing language as a communicative tool including tone of voice, facial expressions, communicative gesture and affect, accepted as a universal in the whole spectrum and long-lasting, even in adult life (Rapin & Dunn, 1997)); (ii) impairment regarding semantics, i.e., the linguistic meaning of utterances and bounds established between words/utterances and what they do/may represent (Rapin & Dunn, 1997). Furthermore, studies that have compared AD, PDD-NOS and specific language impairment (SLI) have provided evidence that language skills in AD and SLI rely on different mechanisms, while PDD-NOS shows an intermediate profile that shares some characteristics of AD and SLI (Demouy et al., 2011; Ringeval et al., 2011).

Here, using automatic extraction and classification, we found some speech turn-taking features of interest for classifying children according to developmental age and/or clinical status; however, we also found that the features that mattered were those features describing gestural turn-taking, and to a lesser extent, features extracted from synchronized motions. Interestingly, some features that were found to classify ASD vs. TD were not found when classifying younger vs. older children. This finding shows that the communication impairments in children with ASD rely not only on verbal deficits but also on body language impairments, and these difficulties are not a simple delay but rather a pervasive development. Regarding the assessment of the spontaneous interactions between a child with ASD and a partner (e.g., the therapist or caregiver), the computation of a general interactive score based on the raw features described above may be of great clinical value because the behavioral modification of children with ASD is often low and subtle and therefore difficult to measure with available clinical tools.

In terms of both communication and interaction, a striking finding of our study relies on the fact that features characterizing the rhythm of the therapist and the duration of his gestural pauses were particularly adequate in predicting whether the child had ASD. To some extent, the idea that information regarding the behavior of ASD children may be found not in the children's' behavior per se but in the way interactive partners adapt to the ASD child constitutes a paradigm shift. Recently, several studies studying infants with ASD or at-risk of ASD shared the same view. In two related studies (Cohen et al., in press; Saint-Georges et al., 2011) based on home movies (HM) of infants, our group showed that when studying interactive patterns with computational methods to take into account synchrony between partners, (i) deviant autistic behaviors appeared before 12 months; (ii) parents seemed to feel weaker interactive responsiveness and mainly weaker initiative from their infants; and (iii) parents increasingly tried to supply soliciting vocalizations rich in motherese and touching. A typical developing twin case study also has shown that a twin's mother used more parentese with the twin who was less reactive (Niwano & Sugai, 2003). Given that HM are not standardized and that analyses are retrospective from the time of positive diagnosis, other research prefers prospective follow-up of high risk samples (e.g., siblings of AD children) with experimental procedures to assess early infant–parent interaction despite the fact that parents are aware of the risk. The British Autism Study of Infants' Siblings reported that early dyadic interaction between at-risk infants and their parents was associated with later diagnosis of autism (Wan et al., 2012). By suggesting that parents feel the pathological process ongoing, we want to guard against the idea that the parenting behaviors are impaired and cause autism. In fact, when parents respond to their infant they behave as parents of TD infant (Saint-Georges et al., 2011). Rather, we suggest that they are some sort of reaction to early sign that are implicitly perceived by the parents and that make them adapt their behavior during interaction. Together with the current results, it seems that interactive patterns should be considered as Social Signal Signs per se and may offer a new area of research in the field of ASD (Chaby et al., 2012).

This paradigm shift, is not surprising when referring in a more general context. Indeed, modeling human communication dynamics becomes easier when considering the complementarities and synchrony between people's verbal and non-verbal behavior. Speaker behavior triggers listeners' back-channels (Gravano & Hirschberg, 2009). Likewise, Morency (2010) predicted the occurrence of listener's backchannel (head nods and gaze aversion) based on the speaker's verbal and non-verbal actions. Lee and Narayanan (2010) predicted interruptions in dialog with three different set of features: based on the interrupter, the interruptee or a combination of both. The set combining features from both partners outperformed the performance of individual sets.

4.3. Current limitations and future work

These results should be taken with care for several reasons: the size of the training set is very limited. In fact, gathering interaction data with disabled children was particularly delicate. Recruiting the children, obtaining the parental consents and collecting the data (with children who often have a tight agenda) were time-consuming and made it more complex to gather large databases. The population may seem undersized compared to previous works on social signal processing. But, gathering interaction data with disabled children is particularly time-consuming and makes it more complex to gather large databases. Moreover, despite the increase of autism prevalence (in the USA, 11.3 per 1000 in 2008, 6.7 per 1000 in 2000[1]), the recruitment entailed to select children in a limited age range and with sufficient interactive abilities to complete the tasks. These restrictive criteria accentuated the difficulty to recruit numerous children. Nevertheless such work could help to identify possible automatic cues to evaluate the children coordination. Moreover, the task we designed is special and consequently our findings may not hold for communicative tasks for instance. Yet, studying such tasks is relevant for therapists to evaluate general abilities of children required in everyday life: turn-taking, joint attention or planning. Another limitation concerned the features that were used in the present task. We initially wanted to track each pieces of the jigsaw to compute the synchronized movement features. However, on many occasions, hands or other pieces were hiding the smaller pieces of the figure. Consequently we decided to use global features such as motion energy or hand tracking to capture the motion of the participants. Consequently we were not able to discriminate if the participants were moving synchronously the same pieces of the jigsaw or different pieces.

In future work, we will extend our efforts regarding the characterization of interactive behavior; for example, we will better characterize gestures that would give a more accurate knowledge of imitation (Delaherche et al., 2012a). The segmentation of the speakers' turns should be automated in order to compute the automatic features online. Then, we will also work to gather a larger database that will generalize our results. These are milestones to be passed before envisionning the use of such system in real practice.

Acknowledgments

The authors would like to thank A.L. Cornuault and C. Debray for their assistance in gathering and annotating the data. This work was supported by the UPMC "Emergence 2009" program, the European Commission (FP7: MICHELANGELO under grant agreement no. 288241) and the Agence Nationale de la Recherche (SAMENTA program: SYNED-PSY).

[1] http://www.cdc.gov/ncbddd/autism/data.htmlhttp://www.cdc.gov/ncbddd/autism/data.html.

Appendix A

See Tables A1 and A2 and Fig. A1.

Table A1
Annotation grid.

Verbal	Sentence	Transcription	
Therapist			
	Dialog act	Initiative assertion	
		R[a] – Backchannel	
		R – Repetition/Rephrasing	
		R – Follow-up	
		R – Opposition/Correction	
		R – Digression	
		A[b] – Induced	Follow closed questions
		A – Implied	Follow open questions
		A – Indirect	Require and inference to link the answer with the question
		A – Metadiscursive	Comment on the question (“Why are you asking me that?”)
		A – Inadequate	
		Q[c] – Closed	Answer is yes or no (“Are you okay?”)
		Q – Alternative	Question holds the answer (“Is it the blue one or the red one?”)
		Q – Categorial	Who, what, when, where...
		Q – Open	The form of the answer is free (“What shall I do next?”)
		Q – Reopening	Repetition of a previous question
		Order – Request	
		Expressive	Exclamative/Greetings
	Deixis	True/False	Here, there...
Gesture	Type	Effective pointing	Help the child
		Ineffective pointing	Does not help the child
		Conventional	Clapping/Nodding
		Therapeutic help	
Child			
	Dialog act	Initiative assertion	
		R – Backchannel	
		R – Repetition/Rephrasing	
		R – Follow-up	
		R – Opposition/Correction	
		R – Digression	
		A – Induced	Follow closed questions
		A – Implied	Follow open questions
		A – Indirect	Require and inference to link the answer with the question
		A – Metadiscursive	Comment on the question (“Why are you asking me that?”)
		A – Inadequate	
		Q – Closed	Answer is yes or no (“Are you okay?”)
		Q – Alternative	Question holds the answer (“Is it the blue one or the red one?”)
		Q – Categorial	Who, what, when, where...
		Q – Open	The form of the answer is free (“What shall I do next?”)
		Q – Reopening	Repetition of a previous question
		Order – Request	
		Echolalia	Automatic repetition of vocalizations made by the therapist
		Expressive	Exclamative/Greetings
	Deixis	True/False	Here, there...
	Answer Adequacy	Adequate	
		Unexpected	The form of the answer if correct but the content is unexpected (“Q: Where may I put the hands?...A:*In* the arms” instead of “*At the tip of* the arms”)
		Inadequate	
Gesture	Type	Pointing	
		Conventional	Clapping/Nodding
		Stereotypy	Repetitive or ritualistic movement
Posture &Gaze	Toward activity	True/False	

[a] R: Retort (follow an assertion).
[b] A: Answer (follow a question).
[c] Q: Question.

Table A2
Mean Cohen's Kappa across all dyads.

Grid item	Category agreement		
	Imitation	Child Follows Instructions	Child Gives Instructions
Therapist dialog acts	0.85	0.84	0.89
Child dialog acts	0.87	0.74	0.89
Child verbal deixis	0.93	0.94	0.95
Child answer adequacy	0.94	0.82	0.92
Child gestures	0.91	1.00	0.93

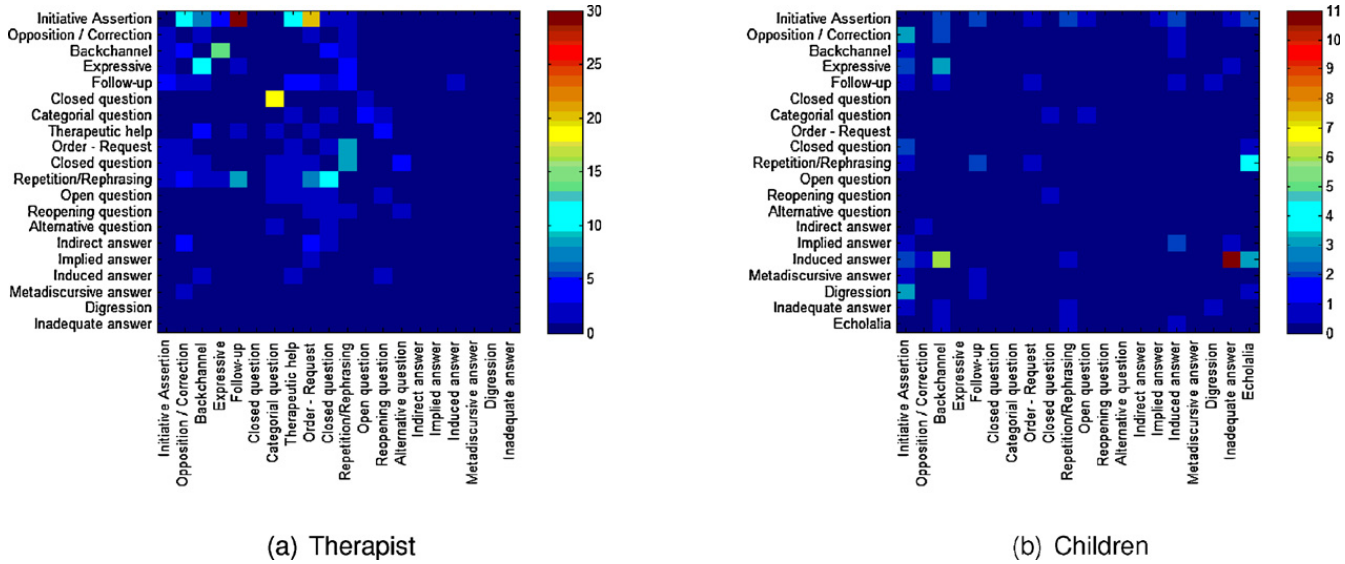


Fig. A1. Dialog acts confusion matrices of the pragmatic grid annotation.

References

Ajmera, J., & Burkhardt, F. (2008). *Age and gender classification using modulation cepstrum*. Proc. Odyssey. p. 25.

Ashraf, A., Lucey, S., Chen, T., Prkachin, K., Solomon, P., Ambadar, Z., et al. (2007). The painful face: Pain expression recognition using active appearance models. In *Proceedings of the ACM international conference on multimodal interfaces (ICMI'07)* (pp. 9–14).

Bemelmans, R., Gelderblom, G. J., Jonker, P., & de Witte, L. (2012). Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *Journal of the American Medical Directors Association, 13*, 114-120.e1 http://dx.doi.org/10.1016/j.jamda.2010.10.002.

Bernieri, F., Reznick, J., & Rosenthal, R. (1988). Synchrony, pseudo synchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology, 54*, 243–253.

Black, M. P., Bone, D., Williams, M. E., Gorrindo, P., Levitt, P., & Narayanan, S. (2011). The usc care corpus: Child-psychologist interactions of children with autism spectrum disorders. In *Proceedings of Interspeech*.

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods, 7*, 338–355.

Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal, 2*, 12–21.

Chaby, L., Chetouani, M., Plaza, M., & Cohen, D. (2012). Exploring multimodal social–emotional behaviors in autism spectrum disorders. *Workshop on wide spectrum social signal processing, 2012 ASE/IEEE international conference on social computing* (pp. 950–954).

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893–910.

Cohen, D., Cassel, R., Saint-Georges, C., Mahdhaoui, A., Laznik, M., Apicella, F., et al. Do motherese prosody and fathers' commitment facilitate social interaction in infants who will later develop autism? *PLoS ONE*, in press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Condon, W., & Ogston, W. (1967). A segmentation of behavior. *Journal of Psychiatric Research, 5*, 221–235.

Delaherche, E., Boucenna, S., Karp, K., Achard, C., & Chetouani, M. (2012a). *Social coordination assessment: Distinguishing between shape and timing*. First IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human Computer Interaction MPRSS 2012 pp. 9–18.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012b). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing, 3*, 349–365.

Demouy, J., Plaza, M., Xavier, J., Ringeval, F., Chetouani, M., Périsse, D., et al. (2011). Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment. *Research in Autism Spectrum Disorders, 5*, 1402–1412 http://dx.doi.org/10.1016/j.rasd.2011.01.026.

Fu, Y., Guo, G., & Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 1955–1976 http://dx.doi.org/10.1109/TPAMI.2010.36.

Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. *INTERSPEECH* (pp. 1019–1022).

Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio–visual nonverbal behavior. *IEEE Transactions on Multimedia, 12*, 563–575.

Hung, H., Huang, Y., Friedland, G., & Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization.. *IEEE Transactions on Audio, Speech & Language Processing, 19*, 847–860.

Jayagopi, D. B., & Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia, 12*, 790–802.
Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica, 32*, 100–125.
Kipp, M. (2008). Spatiotemporal coding in anvil. In *Proceedings of the 6th international conference on Language Resources and Evaluation*.
Kjelgaard, M. M., & Tager-Flusberg, H. (2001). An investigation of language impairment in autism: Implications for genetic subgroups. *Language and Cognitive Processes, 16*, 287–308.
Kozima, H., Michalowski, M., & Nakagawa, C. (2009). Keepon. *International Journal of Social Robotics, 1*, 3–18 http://dx.doi.org/10.1007/s12369-008-0009-8.
Kupper, Z., Ramseyer, F., Hoffmann, H., Kalbermatten, S., & Tschacher, W. (2010). Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. *Schizophrenia Research, 121*, 90–100.
Lakens, D. (2010). Movement synchrony and perceived entitativity. *Journal of Experimental Social Psychology, 46*, 701–708 http://dx.doi.org/10.1016/j.jesp.2010.03.015.
Lee, C.-C., & Narayanan, S. (2010). Predicting interruptions in dyadic spoken interactions. *ICASSP'10* (pp. 5250–5253).
Lucey, P., Cohn, J., Prkachin, K., Solomon, P., & Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. *2011 IEEE international conference on automatic face gesture recognition and workshops (FG 2011)* (pp. 57–64) http://dx.doi.org/10.1109/FG.2011.5771462.
Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., et al. (2007). Comparison of four approaches to age and gender recognition. In *Proceedings of the international conference on acoustics speech and signal processing ICASSP*.
Michalowski, M., Simmons, R., & Kozima, H. (2009). Rhythmic attention in child-robot dance play. In *Proceedings of RO-MAN* (pp. 2009–).
Morency, L. (2010). Modeling human communication dynamics. *IEEE Signal Processing Magazine,* 112–116.
Mower, E., Lee, C.-C., Gibson, J., Chaspari, T., Williams, M., & Narayanan, S. (2011). Analyzing the nature of ECA interactions in children with autism. In *Proceedings of Interspeech*.
Nadel, J., Carchon, I., Kervella, C., Marcelli, D., & Réserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science, 2*, 164–173 http://dx.doi.org/10.1111/1467-7687.00065.
Niwano, K., & Sugai, K. (2003). Maternal accommodation in infant-directed speech during mother's and twin-infants' vocal interactions. *Psychological Reports, 92*, 481–487.
Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* (pp. 61–74).
Prepin, K., & Gaussier, P. (2010). How an agent can detect and use synchrony parameter of its own interaction with a human? In (Series Ed.) & & A. Esposito (Vol. Eds.), *Development of multimodal interfaces: Active listening and synchrony* (Vol. 5967, pp. 50–65). Berlin/Heidelberg: Springer.
Ramseyer, F., & Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference. In (Series Ed.) & & A. Esposito (Vol. Eds.), *Development of multimodal interfaces: Active listening and synchrony* (Vol. 5967, pp. 182–196). Berlin/Heidelberg: Springer.
Rapin, I., & Dunn, M. (1997). Language disorders in children with autism. *Seminars in Pediatric Neurology, 4*, 86–92.
Ringeval, F., Demouy, J., Szaszák, G., Chetouani, M., Robel, L., Xavier, J., et al. (2011). Automatic intonation recognition for the prosodic assessment of language impaired children. *IEEE Transactions on Audio, Speech and Language Processing, 19*, 1328–1342.
Rutkowski, T. M., Mandic, D. P., & Barros, A. K. (2007). A multimodal approach to communicative interactivity classification. *VLSI Signal Processing, 49*, 317–328.
Saint-Georges, C., Mahdhaoui, A., Chetouani, M., Cassel, R. S., Laznik, M.-C., Apicella, F., et al. (2011). Do parents recognize autistic deviant behavior long before diagnosis? Taking into account interaction using computational methods. *PLoS ONE, 6*, e22393 http://dx.doi.org/10.1371/journal.pone.0022393.
Salamin, H., Mohammadi, G., Truong, K., & Vinciarelli, A. (2010). Automatic role recognition based on conversational and prosodic behaviour. In *Proceedings of the ACM international conference on multimedia* (pp. 847–850).
Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences, 10*, 70–76.
Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 326–332.
Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland adaptive behavior scales: Second edition*. Circle Pines, MN: American Guidance Service.
Tager-Flusberg, H. (1981). On the nature of linguistic functioning in early infantile autism. *Journal of Autism and Developmental Disorders, 11*, 45–56 http://dx.doi.org/10.1007/BF01531340.
Tartaro, A., & Cassell, J. (2008). Playing with virtual peers: bootstrapping contingent discourse in children with autism. In *Proceedings of the 8th international conference on international conference for the learning sciences – Volume 2 ICLS'08* (pp. 382–389).
Viaux-Savelon, S., Dommergues, M., Rosenblum, O., Bodeau, N., Aidane, E., Philippon, O., et al. (2012). Prenatal ultrasound screening: False positive soft markers may alter maternal representations and mother-infant interaction. *PLoS ONE, 7*, e30935 http://dx.doi.org/10.1371/journal.pone.0030935.
Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing, 27*, 1743–1759 http://dx.doi.org/10.1016/j.imavis.2008.11.007.
Wan, M. W., Green, J., Elsabbagh, M., Johnson, M., Charman, T., Plummer, F., et al. (2012). Quality of interaction between at-risk infants and caregiver at 12–15 months is associated with 3-year autism outcome. *Journal of Child Psychology and Psychiatry* http://dx.doi.org/10.1111/jcpp.12032.
Wiltermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. *Psychological Science, 20*, 1–5.
Wolters, M., Vipperla, R., & Renals, S. (2009). Age recognition for spoken dialogue systems: Do we need it? In *Proceedings of Interspeech*.
Worgan, S. F., & Moore, R. K. (2011). Towards the detection of social dominance in dialogue.. *Speech Communication, 53*, 1104–1114 http://dx.doi.org/10.1016/j.specom.2010.12.004 Sensing emotion and affect – Facing realism in speech processing.