Characterization of Coordination in an Imitation Task : Human Evaluation and Automatically Computable Cues

Emilie Delaherche
Institut des Systèmes Intelligents et de Robotique
CNRS UMR 7222
Université Pierre et Marie Curie - Paris 6
4 Place Jussieu
Paris, France
emilie.delaherche@isir.upmc.fr

Mohamed Chetouani
Institut des Systèmes Intelligents et de Robotique
CNRS UMR 7222
Université Pierre et Marie Curie - Paris 6
4 Place Jussieu
Paris, France
mohamed.chetouani@upmc.fr

ABSTRACT

Understanding the ability to coordinate with a partner constitutes a great challenge in social signal processing and social robotics. In this paper, we designed a child-adult imitation task to investigate how automatically computable cues on turn-taking and movements can give insight into high-level perception of coordination. First we collected a human questionnaire to evaluate the perceived coordination of the dyads. Then, we extracted automatically computable cues and information on dialog acts from the video clips. The automatic cues characterized speech and gestural turn-takings and coordinated movements of the dyad. We finally confronted human scores with automatic cues to search which cues could be informative on the perception of coordination during the task. We found that the adult adjusted his behavior according to the child need and that a disruption of the gestural turn-taking rhythm was badly perceived by the judges. We also found, that judges rated negatively the dyads that talked more as speech intervenes when the child had difficulties to imitate. Finally, coherence measures between the partners' movement features seemed more adequate than correlation to characterize their coordination.

Categories and Subject Descriptors

J.4 [**Social And Behavioral Sciences**]: Psychology; I.5.4 [**Pattern Recognition**]: Applications—*Signal Processing*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

General Terms

Measurement, Algorithms, Experimentation

Keywords

Social signal processing, coordination, imitation

ICMI'11, November 14–18, 2011, Alicante, Spain.

1. INTRODUCTION

Building robots capable of fluent interactions is a persistent goal of social robotics. The understanding of social and cognitive abilities at stake during interactions represents a puzzling but necessary question for social robotics. Multiple attempts have been made to provide human abilities to robots, like showing expressivity [5], decoding emotion [6], giving appropriate feedback [19] or sharing attention [20, 4] with a human peer. The development of such robots able to devise with oneself in a friendly manner surely represents a long-term objective at the cross-over of multiple disciplines ranging from psychology, to engineering, through signal processing and machine learning. A short-term question could be what human abilities are strictly require to fulfill naturalistic interaction with a peer. Or are there unbreakable rules, that once broken, ruins every attempt to carry on the interaction.

Psychologists commonly refer to coordination as one of those rules. Bernieri et al. define interactional synchrony as "the degree to which the behaviors in an interaction are non-random, patterned, or synchronized in both timing and form" [1]. According to [1], coordination is related to the "interrelatedness" of individuals behavior. It is not necessarily linked with positive affect; two persons fighting are in fact very well-coordinated. Moreover, several psychologists have highlighted that human are efficient and reliable judges of human-human coordination [12, **?**]. Even in degraded conditions, without vocal cues or without the details of facial expressions, this trend remains. Thus the great challenge is to define what perceptual cues help us, human, to consider an interaction as smooth and coordinated...

Several successful attempts have been made in the field of social signal processing to characterize high level nuances of interaction with simple and automatically extracted cues. For instance, cohesion in small group meetings was characterized with audio and visual cues like pauses between individual turns, pauses between floor exchanges, motion turn lengths or synchrony [8]... Dominance was characterized with para-linguistic features [18], speaking length [9] or motion [16]. To identify group conversational behavior, [10] proposed an unsupervised approach based on low-level non verbal cues (speaking length, interruptions, speaking distribution across participants, overlaps...).

In this paper, we investigate the notion of coordination between dyadic partners. We take the example of an imita-

tion task fulfilled by a child and a therapist. We propose to measure the correlation between : (1) automatic and semi-automatic audio-visual cues extracted on the different dyads and (2) the perceived degree of coordination of the dyads, rated by external judges. Our goal, in this paper, is to identify which automatic cues are related with the perceived coordination of the dyads (Fig 1). The following step will be to identify the most appropriate cues to predict low or highly coordinated dyads during the task.

We present the experimental setup and the task in section 2. In section 3, we describe the human evaluation procedure and the results we obtained. Then, we present the automatically and semi-automatically extracted cues to characterize the dyads based on turn-taking, gestures and dialog acts in section 4. At last, in section 5, we present the correlations we found between the coordination scores and the cues we extracted from the video clips.

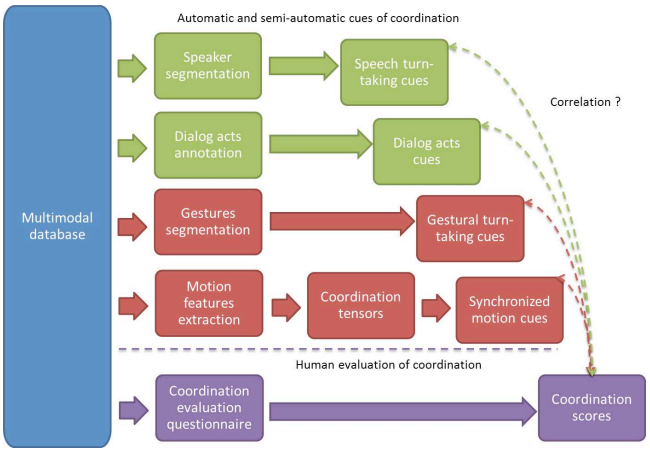

Figure 1: Synopsis

2. EXPERIMENTAL SETUP

We designed an imitation task to evaluate the ability of young children to imitate a partner.

2.1 Participants

21 children participated to the study (developmental ages = 4-9 years) : 14 typically developing children and 7 children followed in the day-care hospital la Pitié-Salpétrière for Pervasive Developmental Disorders. Those children suffered from various social impairments : language disabilities, poor communicative skills, gesture impairments... The current work focuses on the perception of coordination regardless of the children population group. The comparison between groups will be treated in future work.

2.2 Procedure

The task consisted in building a clown with 7 polystyrene elements (2 hands, 2 legs, body, head and hat). The child sat across from the therapist. The same polystyrene elements were arranged on a table in front of them. Therapist led the task and showed to the child each step of the construction task. Speech interventions from the therapist were limited to the times the child encounters difficulties. The children were asked to "do as the therapist" in order to encourage the imitation.

Figure 2: Experimental setup

2.3 Material

The interactions were recorded using a single camera, put up above the participants. The audio recordings were collected at 48 kHz and the video recordings at 25 fps. The participants were equipped with color bracelets to easily track their hands. Audio data were annotated with Anvil annotation tool [11] by two language therapist interns in order to segment the speakers' speech turns and label the utterances according to their dialog acts category (see section 4.2). The total duration was approximately 40min with a mean duration of 1min53s and a standard deviation of 1min09s.

3. EVALUATION OF PERCEIVED COORDINATION

3.1 Evaluation procedure

A questionnaire was built to evaluate the perceived coordination and smoothness of the interactions. The grid was previously used to assess coordination [1]. 17 judges rated the video recordings from the puzzle task. The questionnaire consisted in a 4-item grid :

- Item1 The partners engaged in simultaneous movement
- Item2 The partners had similar tempos of activity
- Item3 The partners' interaction was coordinated and smooth
- Item4 The child matched the adult behavior

The video clips were rated on a six-point Likert scale ranging from 1 (Strongly Disagree) to 6 (Strongly Agree).

3.2 Results

We chose the weighted kappa with quadratic weights to measure the agreement among judges. The weighted kappa was calculated between each pair of annotators for each video and each item of the questionnaire. The mean weighted kappa corresponds to the averaged kappa across all possible pairs of judges. Mean weighted kappa was in average lower for the first item (k = 0.23) and higher for the 4th item (k = 0.34). The lowest kappa value was obtained on item1 (k = 0.13) : this value is usually considered as a slight agreement among raters. We may explain the low agreement by the liberty given to the raters in the interpretation of the 4 items ; no further information was given on the behaviors to rate or the modalities to observe. Considering the small size of our database, we nevertheless decided to keep the 21 video clips in the study.

Fig. 3 represents the evaluation scores versus the mean weighted kappa for the 4 items and the 21 videos. We can see that the plot has a C-shape, similar with the one obtained on a study on cohesion [8] : the agreement tends to be higher at the extreme points of the scale, for the video clips rated as "poorly coordinated" or "highly coordinated".

Finally, the average scores on the 4 items ranged from 1.5 to 5.15, with an average value of 3.54 and a standard deviation of 1.19. The dispersion of the scores showed that our database contains data representative of various degrees of coordination. Moreover, it promotes the idea that coordination is not a all-or-none phenomenon but a continuous notion and that a dyadic interaction can approach or move away from synchrony.

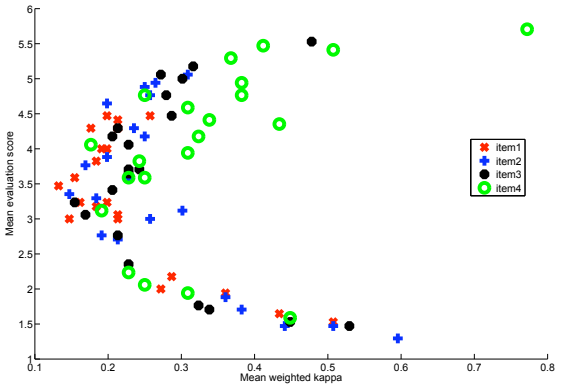

Figure 3: This figure shows the degree of coordination perceived by the raters (mean evaluation score) versus the strength of their agreement (mean weighted kappa), for each dyad and each item of the questionnaire. Agreement tends to be higher for the dyads whose coordination is very low or very high.

4. EXTRACTION OF SEMI-AUTOMATIC AND AUTOMATIC CUES OF COORDINATION

In the following sections, we adopted the abbreviation *Chi* for the Child and *The* for the Therapist. We present in this section the extracted audio and visual features that we assumed were linked with the perceived coordination between the child and therapist.

4.1 Speech turn-taking cues

Raw characteristics on turn-taking have shown to predict efficiently high level characteristics of interaction (cohesion [8], dominance [9]...).

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:

- Pause durations (Pause_dur_statname). A time-segment was considered as a pause if none of the participants was talking. We extracted all the pause segments from the task and calculated standard statistics on the duration of the segments (mean, median, standard deviation, range, minimum and maximum).
- Pause ratio (Pause_ratio). We measured the percentage of interaction time that was silent.
- Speech turn durations (Speech_dur_statname). A speech turn is a continuous time segment when one participant is speaking. We extracted all speech turns for each participant and calculated standard statistics on the duration of the segments.
- Speech ratio (Speech_The_ratio and Speech_Chi_ratio). We measured the percentage of interactional time when each of the participant was talking.
- Overlap ratio (Ovlp_ratio). We measured the percentage of interactional time when both participants were talking at the same time.

4.2 Dialog acts cues

Therapists and children utterances were labeled in Anvil according to the following categorization : initiative assertions, questions, retorts, answers, orders/requests. Moreover, the children answers were labeled according to their adequacy. Conventional head gestures (head nods or head shakes) were also annotated and accounted for adequate answers. We selected the categories that were predominant in the database and related to coordination. We describe in the following paragraphs the cues we chose for the child and the therapist.

4.2.1 Therapist dialog act cues

For the therapist, we focused on the number of questions and orders/requests which were the predominant categories in the database. Moreover, we believed the presence of questions and requests could be linked with a need to engage the child more and thus to a lack of coordination in the dyad. Therapist' questions were more precisely labeled as :

- closed (yes/no) : the answer is yes or no (e.g. "Are you okay?"),
- closed (alternative) : the question holds the answer (e.g. "Is it the blue one or the red one?"),
- categorial : who, what, when, where ... questions (e.g. "Where is the hat?"),
- open : the form of the answer is free (e.g. "What shall I do next?"),
- reopening : repetition of a previously enounced question.

4.2.2 Children dialog act cues

For the children, we focused on the presence and the adequacy of their answers. We hypothesized that an absence of answer or an inadequate answer would be perceived as disruptive for the interaction by the judges. We counted the number of answers according to their adequacy :

- adequate,
- unexpected. The form of the answer if correct but the content is unexpected (e.g. "Q:Where may I put the hands?...A:*In* the arms" instead of "*At the tip of* the arms").
- inadequate. Inadequacy is in both form and context of the answer (e.g. using a spatial deixis (here, there) while not sharing the same visual information with your partner).

Moreover, we counted the number of unanswered questions. An answer was linked to a question if it came after the beginning of the question, in a 10 seconds delay and before a new utterance from the therapist. This way, we took into account a possible overlap between the therapist question and

the child answer. The 10 seconds delay may seem outsize but took into account observed time responses from disabled children.

4.3 Gestural turn-taking cues

Turn-taking is a pre-requisite of imitation. In the task we proposed, the child needs to look at the demonstrator and then reproduce the same actions. While the child is imitating, the demonstrator waits that he has finished to show him the next stage of the assembly. So we would expect to find an alternation of sections when the demonstrator is gesturing and the child stays still and sections when the demonstrator observes and the child assembles the pieces : a gestural turn-taking. We assumed the respect of this balance would be a positive predictor of perceived coordination.

We tracked the participants hands trajectory with the coupled Camshift algorithm [3]. The participants were equipped with salient color bracelets to easily follow their gestures. We then compressed x and y Cartesian coordinates to the polar coordinate $r = \sqrt{x^2 + y^2}$ and derived r to obtain the hands velocity.

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

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

- The_off_Chi_on_ratio. Percentage of time when the child is gesturing and the therapist stays still,
- The_on_Chi_off_ratio. Percentage of time when the therapist is gesturing and the child stays still,
- The_on_Chi_on_ratio. Percentage of time when therapist and child are gesturing at the same time,
- The_off_Chi_off_ratio. Percentage of time when therapist and child are staying still at the same time.

Moreover, the task is particularly repetitive (succession of assembling and observing phases); so we can assume that a rhythm is established during the task. All pause segments should last approximately the same duration, so as 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 duration of the gestural and pause segments :

- Gestural Pause Durations (Pause_dur_statname). From the pause segments, we calculated standard statistics on the duration of the segments (mean, median, standard deviation, range, minimum and maximum).
- Gestural Segments Durations (Gest_dur_statname). We extracted the same statistics on the duration of the segments when the participant was moving.
- Pause ratio (Pause_The_ratio or Pause_Chi_ratio). We measured the percentage of interaction time that the participant stayed still.

Lastly, we simply counted the number of continuous sequences of movements for each participant (Seq_The_nb and Seq_Chi_nb) and the ratio between the number of sequences of each participant (Seq_The_Chi_ratio).

4.4 Synchronized motion cues

We also propose to measure the coordination between the child and the therapist as the degree of similarity between their motion features. Thus, we extracted various motion features from the database, computed similarity measures on windows of interaction. We then applied a bootstrap method to classify the windows as "coordinated" or "not coordinated" and deduced a measure of coordination from the percentage of windows classified as "coordinated". The method was previously described to study multimodal coordination in a similar task [7].

4.4.1 Video features

We extracted the following features at 25Hz for each participant. First, Motion energy (ME) is defined as the number of pixels in movement between the current video frame and a reference image in a region of interest. Preciously studied to assess interactional synchrony [13], motion energy informs about a shared dynamics between the two partners movement. For each speaker, we defined three regions of interest in the video : one centered on the head and torso of the participants, to capture postural movement, one centered on the hands to capture manipulative gestures and a last region that covered all the participant. We also extracted Hands velocity as described in section 4.3.

4.4.2 Coordination tensors

Various measures have been proposed to evaluate the similarity between two time series on the time or the frequency domain : cross-correlation [13], phase synchronization [17], mutual information [15], cross-spectral coherence [14] ...

We opted for cross-correlation and cross-spectral coherence for the simplicity of their estimation, opposed to mutual information or phase estimation and the fastness of their computation. We selected cross-correlation to capture the similarity of the time series in the time domain.

Cross-spectral coherence measures the similarity of the participants in the frequency domain, comparing the similarity of the power spectral densities of the time series in each band of frequency. Thus it may be useful to identify variations that share the same spectral properties, the same rhythm, regardless of an eventual time-lag between the processes occurring in the time series.

We performed the two measures across all possible pairs of features on 1s, 2s and 5s frames delayed of 0s, 1s or 2s lags. Thus large 3D coordination tensors (feature x x feature y x duration of the video) were obtained where each coefficient represents the similarity between a pair of features for a given time window (see Fig 1).

4.4.3 Ruling out random coordination

A critical question when we try to detect dependence relationships between features is beyond which score should we consider the measure as significant. Consequently, we needed a reference to compare our score and conclude about the significance of our measure.

[2] originally proposed a rating method ("the pseudo synchrony experimental paradigm") to evaluate the interactional synchrony that occurred in a dyadic interaction. The method consisted in synthesizing pseudo interactions : video images of dyadic partners were isolated and re-combined in a random order. After judges rated original videos and pseudo interactions videos. Pseudo interactions scores constituted

a baseline to judge the scores obtained on the original interaction.

We used an extension of this method to automatic computation to classify each window of interaction as "coordinated" or "not coordinated" [13]. First, features were extracted for each dyadic partner on the original video. The features extracted for first partner were time-shuffled and reassociated with the second partner original features. With this bootstrap method, 100 pseudo interactions were generated. Genuine coordination measures were assessed on original features and 100 pseudo interactions coordination measures were computed on the time-shuffled features. Then, to replace human judges from the Bernieri method, a one-sided z-test (p = 0.05) was performed to compare genuine interaction measures and pseudo interactions measures. We concluded that a given window was "coordinated" if genuine coordination measure was two standard deviations above the pseudo coordination measures.

We finally characterized the degree of synchronized movements between partners with the percentage of frames for which the measure was significant.

5. RESULTS

In section 3, we obtained a human evaluation of coordination for each video clip in our database. In section 4, we extracted semi-automatic and automatic cues that we assumed were related with the coordination of the dyads. In this section, we finally measure the Pearson correlation coefficient between the scores from the evaluation and the cues extracted on the video clips. Only the cues that presented significant relations with the coordination scores are provided in the tables below.

5.1 Speech turn-taking cues

First, judges rated better interaction with longer speech pauses: pauses duration (mean, median and minimum) was positively correlated with coordination scores from all items (see Table 1). Also, when silence covered a larger part of the interaction, the coordination scores were higher. Therapist shorter speech turns corresponded to the interaction rated as less coordinated and the amount of therapists speech turns was negatively correlated with coordination scores. The duration of children speech turns (mean, median, minimum and maximum) were significantly higher in interaction poorly coordinated. Then, children turns durations varied more in interactions poorly coordinated (range and standard deviation). And the greater was the proportion of the child speech turns the worst the raters perceived the dyad. At last, the amount of speech overlaps were not significantly linked with perceived coordination.

5.2 Therapist and children dialog acts cues

Although we did not find a significant correlation, we observed a negative trend that linked the number of yes/no questions with the perceived coordination (see Table 2). The number of categorial and open questions, as well as the overall number of questions from the therapist was negatively linked with the evaluation scores. The number of requests was also negatively correlated with the first item of the questionnaire.

There was no significant relation between the adequacy of the answers and the coordination scores (see Table 3). On the other hand the number of not answered questions was directly linked with a negative perception of the coordination.

Table 1: This table presents the correlation between speech turn-taking cues and coordination scores. The less the therapist and the child needed to speak the better coordinated the dyad was perceived.

Features	Item1	Item2	Item3	Item4
Pauses				
Pause_dur_mean	0.59^{**}	0.61^{**}	0.65^{**}	0.64^{**}
Pause_dur_min	0.42^{1}	0.48^{*}	0.48^{*}	0.46^{*}
Pause_dur_med	0.57^{**}	0.61^{**}	0.63^{**}	0.61^{**}
Therapist				
Speech_dur_min	0.53^{*}	0.5^{*}	0.5^{*}	0.44^{*}
Child				
Speech_dur_mean	-0.63^{**}	-0.59^{**}	-0.64^{**}	-0.62^{**}
Speech_dur_std	-0.5^{*}	-0.43^{1}	-0.47^{*}	-0.38^{1}
Speech_dur_min	-0.35	-0.31	-0.37^{1}	-0.39^{1}
Speech_dur_max	-0.71^{***}	-0.67^{***}	-0.7^{***}	-0.67^{***}
Speech_dur_range	-0.74^{***}	-0.7^{***}	-0.72^{***}	-0.68^{***}
Speech_dur_med	-0.59^{**}	-0.56^{**}	-0.61^{**}	-0.6^{**}
Ratios				
Pause_ratio	0.62^{**}	0.64^{**}	0.66^{**}	0.65^{**}
Speech_The_ratio	-0.35	-0.37	-0.38^{1}	-0.38^{1}
Speech_Chi_ratio	-0.64^{**}	-0.65^{**}	-0.65^{**}	-0.66^{**}

^{1} p<0.1 * p<0.05 ** p<0.01 *** p<0.001

Table 2: This table presents the correlation between therapist dialog acts cues and coordination scores. Questions and requests occurred when the child needed more support to accomplish the task.

Features	Item1	Item2	Item3	Item4
Closed (yes/no)	-0.37^{1}	-0.38^{1}	-0.38^{1}	-0.41^{1}
Categorial	-0.45^{*}	-0.49^{*}	-0.48^{*}	-0.46^{*}
Open	-0.52^{*}	-0.47^{*}	-0.44^{*}	-0.45^{*}
Total questions	-0.5^{*}	-0.52^{*}	-0.51^{*}	-0.54^{*}
Order/Request	-0.45^{*}	-0.35	-0.41^{1}	-0.36

^{1} p<0.1 * p<0.05 ** p<0.01 *** p<0.001

5.3 Gestural turn-taking cues

Large variations of the gestural pause durations of the therapist (range and standard deviation) were negatively linked with coordination, as larger maximal duration of gestural pause during the interaction (see Table 4). Therapist's maximum and range of gestural segments duration were lower for highly coordinated video clips (significant correlation for item 4 and trend for item 1 and 3).

The child's gestural pauses (minimum, mean and median duration) tended to last longer for well-coordinated dyads. Several items showed this significant relationship.

The proportion of time when the therapist was gesturing and the children was not, was higher in interaction perceived as highly coordinated. Finally, there was significantly more movement sequences from both child and adult on poorly coordinated dyads whereas the number of puzzle elements stayed constant. And there were proportionally more movement sequences from the child than from the therapist on those dyads. In fact, the interactions were more fragmented, the child had to start again several times before adjusting

Table 3: This table presents the correlation between children answers and coordination scores. More adequate answers and less unanswered questions characterized the well coordinated children.

Features	Item1	Item2	Item3	Item4
Adequate	-0.3	-0.38[1]	-0.34	-0.32
Not answered	-0.41[1]	-0.39[1]	-0.39[1]	-0.45[*]

[1] p<0.1 [*] p<0.05 [**] p<0.01 [***] p<0.001

Table 4: This table presents the correlation between gestural turn-taking cues and coordination scores. In the well coordinated dyads, the duration of the therapists gestural pauses were more homogeneous and the child spent less time gesturing while the therapist was demonstrating

Features	Item1	Item2	Item3	Item4
Therapist' gestures pauses				
Pause_dur_mean	-0.35	-0.43[1]	-0.33	-0.3
Pause_dur_std	-0.43[1]	-0.52[*]	-0.45[*]	-0.45[*]
Pause_dur_max	-0.58[**]	-0.64[**]	-0.61[**]	-0.62[**]
Pause_dur_range	-0.58[**]	-0.64[**]	-0.61[**]	-0.62[**]
Therapist' gestures				
Gest_dur_max	-0.41[1]	-0.36	-0.37[1]	-0.49[*]
Gest_dur_range	-0.41[1]	-0.36	-0.37[1]	-0.49[*]
Child' gestures pauses				
Pause_dur_mean	0.3	0.32	0.35	0.37[1]
Pause_dur_min	0.44[*]	0.46[*]	0.4[1]	0.4[1]
Pause_dur_med	0.41[1]	0.44[*]	0.48[*]	0.48[*]
Ratios				
Pause_The_ratio	-0.3	-0.37[1]	-0.3	-0.21
The_on_Chi_off_ratio	0.42[1]	0.48[*]	0.44[*]	0.38[1]
The_off_Chi_off_ratio	-0.35	-0.42[1]	-0.35	-0.31
Number of sequences				
Seq_The_nb	-0,83[***]	-0,8[***]	-0,84[***]	-0,86[***]
Seq_Chi_nb	-0,88[***]	-0,89[***]	-0,89[***]	-0,91[***]
Seq_The_Chi_ratio	-0,54[*]	-0,62[**]	-0,55[**]	-0,54[*]

[1] p<0.1 [*] p<0.05 [**] p<0.01 [***] p<0.001

two elements together. The therapist needed to demonstrate several times the same stage of the assembly. We did not find significant relations between the movement features extracted on the children movements duration statistics and the coordination scores.

5.4 Synchronized motion cues

5.4.1 Measures based on correlation

We measured a significant negative relation between synchronized motion cues and coordination scores for Global Motion (ME) on windows of 2s and 5s with a delay of 1s between the therapist and the child (see Table 5). We also observed this negative trend for Posture Motion (ME) on windows of 1s with 1s delay and on windows of 2s with 0s or 1s delay.

For Hands Motion (ME), we measured a negative relation with raters evaluation for windows of 5s with no delay but the relation was positive for windows of 1s and 5s with 2s delay. For Left Hand (Tracking), there seems to be a negative trend for windows delayed of 1s but the tendency seems to be inverse for windows delayed of 2s. Nevertheless, those relations did not reach significance.

Table 5: This figure presents the correlation between synchronized motion cues (measures based on correlation) and coordination scores.

Features	Item1	Item2	Item3	Item4
Global (ME)				
Win=2s_delay=1s	-0,39[1]	-0,38[1]	-0,36	-0,32
Win=5s_delay=1s	-0,41[1]	-0,37[1]	-0,43[1]	-0,28
Posture (ME)				
Win=1s_delay=1s	-0,42[1]	-0,39[1]	-0,4[1]	-0,32
Win=2s_delay=0s	-0,37	-0,44[*]	-0,31	-0,28
Win=2s_delay=1s	-0,55[**]	-0,54[*]	-0,57[**]	-0,49[*]
Hands (ME)				
Win=1s_delay=2s	0,39[1]	0,37	0,31	0,37[1]
Win=5s_delay=0s	-0,42[1]	-0,41[1]	-0,44[*]	-0,46[*]
Win=5s_delay=2s	0,47[*]	0,5[*]	0,42[1]	0,41[1]
Left hand (Tracking)				
Win=5s_delay=1s	-0,39[1]	-0,36	-0,32	-0,24
Win=5s_delay=2s	0,29	0,27	0,31	0,38[1]

[1] p<0.1 [*] p<0.05 [**] p<0.01 [***] p<0.001

5.4.2 Measures based on coherence

For Global Motion (ME), we measured a positive relation between coordination parameters and scores (trend on windows of 2s with 2s delay on item3 and significant relation on windows of 5s with 2s delay)(see Table 6). We also observed a significant positive link between Posture Motion (ME) of the participants measured on 1s windows with 2s delay. For Hands Motion (ME), we assessed a positive significant link on windows of 2s and 5s with 2s delay. But the relation was negative for Right Hand (Tracking) on 2s windows with 1s delay. For Left Hand (Tracking), there were some positive and negative trends for windows of 1s without delay, windows of 2s with 2s delay and windows of 5s with 1s delay. The only relation that reached significance was between Left Hand (Tracking) measured on 5s windows with 2s delay and scores from item2.

6. DISCUSSION

6.1 Speech as a last resort

First, the raters tended to judge negatively the dyads where the therapist was more talkative. In fact, there is no clear necessity to speak to fulfill the task ; the clinician tended to use speech when the visual demonstration was not sufficient. Consequently, the longer the speech pauses lasted the more coordinated the dyad was perceived. No words were needed.

We also observed that the length of the therapist utterances tended to be shorter for "low coordination" interactions. We can hypothesize that the therapist adapted the length of his speech according to the child need. She used shorter utterances for children with poor abilities. Therapist intervenes more on "low coordination" videos trying

Table 6: This figure presents the correlation between synchronized motion cues (measures based on coherence) and coordination scores. Global Motion (ME) and Hands Motion (ME) with 2s delay were more synchronized for well coordinated dyads. For shorter delays, we observed opposite relations for hands motions.

Features	Item1	Item2	Item3	Item4
Global (ME)				
Win=2s_delay=2s	0,27	0,35	0,37[1]	0,33
Win=5s_delay=2s	0,44[*]	0,48[*]	0,42[1]	0,4[1]
Posture (ME)				
Win=1s_delay=2s	0,6[**]	0,66[**]	0,6[**]	0,54[*]
Mains (ME)				
Win=2s_delay=2s	0,46[*]	0,48[*]	0,47[*]	0,44[*]
Win=5s_delay=2s	0,44[*]	0,41[1]	0,39[1]	0,36
Right hand (Track.)				
Win=2s_delay=1s	-0,48[*]	-0,43[*]	-0,42[1]	-0,36
Left hand (Track.)				
Win=1s_delay=0s	0,4[1]	0,36	0,37	0,37
Win=2s_delay=2s	0,41[1]	0,41[1]	0,38[1]	0,38[1]
Win=5s_delay=1s	-0,43[1]	-0,39[1]	-0,4[1]	-0,33
Win=5s_delay=2s	0,38[1]	0,44[*]	0,42[1]	0,42[1]

[1] p<0.1 [*] p<0.05 [**] p<0.01 [***] p<0.001

to engage or help children with semantic informations. We tended to have more categorial questions and open questions for those videos. There were also more orders and requests : "look", "go", "do just like me"...

For the children, we observed that the utterances mean duration and standard deviation were lower for highly coordinated dyads. In fact, children who performed the task without difficulty only used speech for short utterances. They gave back-channels to the clinician for example. Some of them even did not speak at all. On the contrary, when the interactions were less smooth, children asked more questions... We also observed digressions and echolalia for some of them. Standard deviation of child turn duration represents the alternation of short turns with long turns. Children who showed high coordination produced only short utterances (back channels). Absence of answer to the therapist questions was also typical of interactions judged with low coordination.

6.2 Fluency of gestural turn-taking

Therapist gestural pauses represent the time that therapist waits for the child to fulfill next step of the assembly. The interactions in which the therapist had to wait for the children where perceived as poorly coordinated. The variation of the pause durations of the therapist (range and standard deviation) is related to rhythm : when pause durations were stable, the interactions were perceived as well coordinated. There was a comparative trend for the range duration of gestural sequences of the therapist, which was smaller for highly coordinated dyads.

On the contrary when there was a large range between short and long pauses durations, it means that some steps of the assembly took longer than the others. Those interactions were perceived as less fluent by the annotators. We also observed that the movement maximal durations were higher in low coordinated tasks. We can assume that sometimes the therapists repeated the movement several times or did it more slowly to help the child decompose the steps. This was confirmed by a smaller number of movement sequences of the therapist in the well-coordinated dyads.

Lastly we noted that the percentage of time that the child was moving while the therapist was demonstrating was correlated with the perception of coordination. We can assume a link between these features and the attention of the child towards the therapist. Children who gestured a lot while the therapist was demonstrating were perceived less attentive and concentrated on the task. We also observed that the gestural pauses were longer for children who appeared well-coordinated.

6.3 Similar rhythm more than synchronized movements

When we designed our automatic cues, we hypothesized that coordination parameters would be positively linked with evaluation scores. But surprisingly, we found several coordination parameters, for measures based on correlation, that were negatively linked with the evaluation scores. Thus, we can make several assumptions to explain these results.

First, the features we used might be too global to discriminate between the actual gestures required to perform the task and interfering gestures that we implicitly filter when we watch the video clips. Also, our features are not capable of discriminating between two people moving the same piece of the puzzle and two persons moving two different elements at the same time. Despite the bootstrap stage, there are certainly windows of interaction that are classified as synchronous whereas they are not.

Then, we can remark that the tendency tends to reverse when we increase the time-lag between the windows on which the correlation is calculated (negative relation for 5s windows with no delay, but positive relation for 5s windows with 2s delay). Thus, we may not deal properly with the time-lag between the partners. In our model, we measure the percentage of windows that are "coordinated" for arbitrarily fixed time-lags of 0s, 1s and 2s. But this lag certainly vary in a dyad from one puzzle element to another and from one dyad to another. A related question is how do we humans perceive and tolerate the lag between the partners. For instance, the first item of the coordination questionnaire ("the partners engaged in simultaneous movement) was the most puzzling for the annotators as illustrates the low kappa coefficient. In fact, there are no perfectly simultaneous movements during the task as the child imitates the therapist. There is always a certain lag between the demonstration from the therapist and the imitation from the child. And the judges according to their tolerance to the time-lag rated more or less harshly the dyads.

At last, the task is very complex to analyze automatically; there are plenty of manners to assemble two elements together. If the two participants don't follow the same strategy, the model won't find them coordinated while extern judges may find they are. To circumvent the features issues, to rule out the interfering gestures and to limit the degree of freedom of the participants, we are setting up a computerized jigsaw task. The principle will be similar to the clown task but the assembly will be completed on a computer. This way, we may only record gestures related to the task

(mouse trajectories) and know which element of the jigsaw is actually moved.

Finally, coordination parameters based on coherence gave more interesting results for global motion, posture and hand motion. As it is independent of the lag between partners, coherence captured a shared rhythm of the dyadic partners. This measure appears to be more adapted to our task and to the perception of the judges to characterize the coordination of the dyad.

7. CONCLUSION

We presented in this work several automatically computable cues that are related with the perception of coordination in an imitation task. The population we chose to work with is special as it involves children, including several ones with social impairments. The task we propose also bears special features. Imitation is a certainly a special type of interaction but it was originally suggested by psychologists as it requires multiple abilities to coordinate and cooperate with the partner : turn-taking, joint attention or planning. So, even if the task is specific, it can give a good insight on general abilities required in everyday life. And the evaluation questionnaire rates general characteristics of coordination : smoothness, behavior matching, similar tempos. Then, identifying low level cues to evaluate and compare children interactive abilities is particularly relevant for psychologists. Imitation is also a simple framework for human-robot interaction. Our work could give insight on how to evaluate or adapt robots to fulfill such tasks. We observed for example that the similarity of gestures was not particularly relevant. On the contrary, demonstrator's variations of rhythm and difference of rhythm between the partners were badly perceived by the judges.

Acknowledgments

The authors would like to thank F. Bigouret, A.L. Cornuault and C. Debray for their assistance in gathering and annotating the data. The authors also wish to thank L. Chaby, M. Plaza and D. Cohen for providing useful comments and advices. This work was supported by the UPMC "Emergence 2009" program.

8. REFERENCES

[1] F. Bernieri, J. Reznick, and R. Rosenthal. Synchrony, pseudo synchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, 54(2):243–253, 1988.

[2] F. Bernieri and R. Rosenthal. *Interpersonal coordination: Behavior matching and interactional synchrony. Fundamentals of nonverbal behavior*. Cambridge University Press, 1991.

[3] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2), 1998.

[4] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146–1153, 1999.

[5] C. L. Breazeal. *Sociable machines: expressive social exchange between humans and robots*. PhD thesis, 2000.

[6] A. Delaborde and L. Devillers. Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 75–80, 2010.

[7] E. Delaherche and M. Chetouani. Multimodal coordination: exploring relevant features and measures. In *Second International Workshop on Social Signal Processing, ACM Multimedia 2010*, 2010.

[8] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[9] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech & Language Processing*, 19(4):847–860, 2011.

[10] D. B. Jayagopi and D. Gatica-Perez. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, 12(8):790–802, 2010.

[11] M. Kipp. Spatiotemporal coding in anvil. In *LREC*, 2008.

[12] D. Lakens. Movement synchrony and perceived entitativity. *Journal of Experimental Social Psychology*, 46(5):701 – 708, 2010.

[13] F. Ramseyer and W. Tschacher. Nonverbal synchrony or random coincidence? how to tell the difference. In A. Esposito et al., editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967, pages 182–196. Springer Berlin / Heidelberg, 2010.

[14] M. J. Richardson, K. L. Marsh, R. W. Isenhower, J. R. Goodman, and R. Schmidt. Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, 26(6):867 – 891, 2007.

[15] M. Rolf, M. Hanheide, and K. Rohlfing. Attention via synchrony : Making use of multimodal cues in social learning. *IEEE Trans. Auton. Mental Develop.*, 1(1):55–67, 2009.

[16] G. Varni, A. Camurri, P. Coletta, and G. Volpe. Toward a real-time automated measure of empathy and dominance. In *CSE (4)*, pages 843–848, 2009.

[17] G. Varni, M. Mancini, G. Volpe, and A. Camurri. Sync'n'move: social interaction based on music and gesture. *Proceedings of the 1st International ICST Conference on User Centric Media*, 2009.

[18] S. F. Worgan and R. K. Moore. Towards the detection of social dominance in dialogue. *Speech Communication*, In Press:–, 2011.

[19] B. Wrede, S. Kopp, K. Rohlfing, M. Lohse, and C. Muhl. Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*, 42(9):2369 – 2384, 2010. How people talk to Robots and Computers.

[20] Z. Yucel, A. A. Salah, C. Mericli, and T. Mericli. Joint visual attention modeling for naturally interacting robotic agents. In *ISCIS*, pages 242–247, 2009.