Enabling robots to make use of the structure of human actions - a user study employing Acoustic Packaging

Manja Lohse[1], Britta Wrede[2], and Lars Schillingmann[3]

***Abstract*— Human learning strongly depends on the ability to structure the actions of teachers in order to identify relevant parts. We propose that this is also true for learning in robots. Therefore, we apply a method for multimodal action segmentation called Acoustic Packaging to a corpus of pairs of users teaching object names to a robot. Going beyond previous use cases, we analyze how the structure of human actions changes if the robot is learning quickly or slowly. Our results reveal differences between action structuring in the conditions such as longer utterances and more motion when the robot learns slowly. We also evaluate how the partners in the pair influence each other's action structuring. The results show a strong correlation between the participants in the pairs, even more so in the trials where the robot is learning slowly. We conclude that the action structuring based on Acoustic Packaging allows robots to differentiate how well the interaction with multiple users is going and is, thus, a vehicle for feedback generation.**

I. Introduction and Motivation

Throughout their lives humans need to learn thousands of actions, many of these being taught to them by others. For example, parents teach their children to draw shapes, to use objects, or to articulate sentences. To learn these actions, people must develop models to segment them in a reliable manner, i.e., they need to determine when an action starts and ends and which modalities it includes to decide which essential subparts the action consists of. Zacks and Swallow [14] have shown that event segmentation supports memory and learning. Also an increasing number of robots depend on learning from humans because the systems are used in a variety of environments that they cannot be preprogrammed for. Our basic assumption in this context is that robots, just like humans, need to structure the input while processing it. In other words, robots need action segmentation methods. We have previously proposed one such method [9] that has been shown to provide a meaningful bottom-up approach to action segmentation in tutoring situations among humans and between humans and robots. The method is called *Acoustic Packaging*. It segments the actions of the teacher based on auditive and visual cues. In previous work [10], the Acoustic Packaging approach has been used for analyzing one-to-one tutoring situations and comparing adult-adult, adult-child, and adult-robot tutoring. It was found that the tutoring of robots resembled the tutoring of children. In both types of interaction, less action was packaged within an utterance compared to adult-adult interaction. Yet, an important difference between adult-robot interaction and adult-child interaction was the higher verbosity of people interacting with the robot. Taking these findings as a starting point, in the current paper we apply the approach to interactions in small groups (one robot and two humans) as depicted in Fig. 1. The paper contributes a description of how people structure their actions when teaching names of objects to a robot (and not manual actions like in previous work) in a small group. Moreover, it investigates the degree of interpersonal correlation between people's behaviors in pairs, or in other words how strongly participants in pairs align their actions to each other. The study also takes situative factors into account, i.e., by comparing tutoring situations in which the learning capabilities of the robot vary (learning quickly versus learning slowly). Based on this work, the robot can be equipped with better action understanding capabilities for one on one and group interaction and, thus, Acoustic Packaging can be an efficient vehicle for developing appropriate feedback behavior. The approach is particularly suitable here because it operates on bottom-up cues such as speech and movement. In other words it does not depend on the input of any knowledge of the interaction (which we do not have at this stage) such as how people move. It rather generates knowledge about the action structure based on statistics of visual and acoustic cues and develops its own action representations.

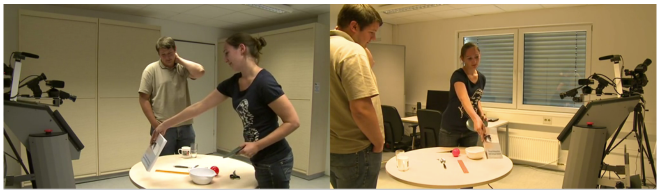

Fig. 1. Tutoring situation with the robot BIRON

II. Acoustic Packaging

In this section, we present the steps and underlying algorithms of the Acoustic Packaging approach. For a more detailed description the reader is referred to [8], [9].

A. Multi-modal Segmentation of Acoustic Packages

In a first step, we segment the acoustic and the visual modality. In the following section we describe how this works and how the results are merged to form acoustic packages.

1) Acoustic Segmentation: Based on the observation that speech is structured by pauses, it seems appropriate to segment the acoustic signal simply into speech and non-speech (pause)

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

[1] M. Lohse is with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands m.lohse at utwente.nl

[2] B. Wrede is with the Faculty of Technology, Bielefeld University, 33615, Bielefeld, Germany bwrede at techfak.uni-bielefeld.de

[3] L. Schillingmann is with the Graduate School of Engineering, Osaka University, 565-0871, Osaka, Japan lars at ams.eng.osaka-u.ac.jp

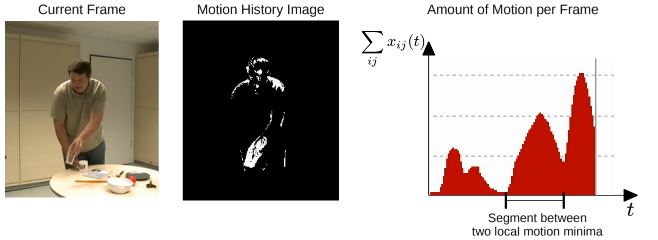

Fig. 2. A person showing an object (left), the corresponding motion history image (middle), and the approach to visually segment actions via the amount of motion per frame (right).

segments. Yet in relatively noisy environments, the separation of speech from non-speech is a difficult task. The Acoustic Packaging approach offers a solution for that [9], however, we do not want to go into detail here because the approach has not been applied in our study. In contrast, our corpus of data came with annotations of speech that provided a clear-cut segmentation of speech and pauses.

2) Visual Action Segmentation: For visual action segmentation we employ a bottom-up approach that makes use of pixel-wise changes in image sequences of the video stream. We compute the amount of pixels that have changed their value during a certain time interval. This is based on the assumption that movement results in pixel changes. The faster a person is moving, the more change occurs in the number of pixels. If there is no movement, no pixel changes occur. Thus, by finding minima in the function describing the amount of pixel change we can segment the video stream into *motion peaks* (MPs). The underlying assumption is that MPs correlate with basic movements such as arm gestures.

The segmentation into motion peaks is technically realized by an approach based on motion history images [1]. The idea is that a motion history image $x_{ij}(t)$ contains non-zero values at the coordinates (i, j) if the corresponding pixels change within a history window because of motion. Thus, the amount of motion can be calculated per frame at time-step t by summing up the motion history image (see Figure 2). In the amount of motion, local minima are detected with the help of a sliding window that is updated at each time step. Here the motion history consisted of 10 frames. If the value at the center of the window is smaller than the local neighborhood, a minimum is detected. Very small changes are considered as no motion and filtered out by applying a threshold. Small local peaks are suppressed by using a sufficient window size that is yet small enough to detect actual human movements because we chose to not use any prior knowledge with respect to relevant areas in and content of visual information. Here the window size was 10 frames. When a local minimum is detected, a motion peak is created describing the motion between the previous and the current motion minimum. The description contains time stamps of the minima. Through these timestamps the visual stream can be temporally associated with the acoustic stream.

3) Temporal Association: The motion peaks and the speech segments need to be temporally associated in order to form acoustic packages. When a new segment arrives, the corresponding time interval is aligned to its modality-specific timeline. In the next step, the temporal relations to the segments on the other timeline are calculated. When overlapping speech and motion segments are found on the timelines, acoustic packages are created. In the case that motion segments overlap with two different speech segments, the one with the larger overlap is chosen (see Figure 3). Thus, a motion segment cannot bind multiple speech segments together. However, multiple motion segments can be associated to one speech segment to form an acoustic package. The boundaries of the acoustic packages are determined by all modalities depending on which one starts earlier and ends later. So the beginning can, e.g., be determined by motion while the end of the same acoustic package is determined by the speech as is the case in both examples in Figure 3.

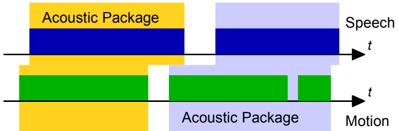

Fig. 3. Motion and speech intervals are assigned to an acoustic package if they overlap. The middle motion interval has been assigned to the second acoustic package due to greater overlap.

B. Structure Detection within Acoustic Packages

In a further step towards providing a learning system with more systematic and structured information, we enhanced our model with mechanisms for structure detection within acoustic packages. This mechanism relies on the assumption that speakers stress important parts of their utterances which results in an increase of the prominence of the signal.

1) Acoustic Prominence Detection: We understand perceptual prominence of linguistic units as the units' degree of standing out relative to its environment [11], thus, leading to a ranking of syllables within an utterance. We perform the prominence detection on a pre-segmentation of the speech stream into syllables which we yield by a modified version of the Mermelstein algorithm [5]. Subsequently, each syllable is rated according to the acoustic parameters which correlate to the perceived prominence. We implemented a simplified version of a prominence algorithm described in [11] using only one feature, namely spectral emphasis, which showed an almost similar performance on our data.

2) Prominence Assignment to Words: For further analysis we were interested in which words are actually stressed by the speaker. We therefore performed a manual transcription of the speaker's utterances and time-aligned the annotations to the speech signal in order to synchronize them with the prominence detection module. This is done during a so called forced alignment by an automatic speech recognition system. We use the ESMERALDA framework [2] for this. We feed the recognizer with the sequence of words together with the speech stream. The recognizer outputs a time alignment at word (and phone) level. This word information is then merged with the time aligned prominence detection output which results in the most prominent word for each acoustic package. Note that this assignment of prominence to words is only possible in an offline version of the module whereas all other detection mechanisms run online.

3) Color Saliency Based Tracking: The AP approach also contains a method for color based saliency tracking that works

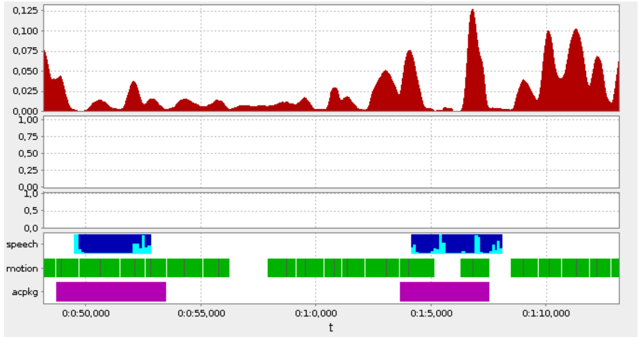

Fig. 4. The tool displays motion activity, speech segments with prominence information, motion segments, and the acoustic packages formed using these cues. The empty row usually displays the results of the color saliency based tracking that has not been used here.

on objects with bright colors that we used in previous studies. However, this aspect of Acoustic Packaging could not be employed here because the objects used in the study were everyday objects that could not reliably be tracked based on their color (e.g., a key, scissors, a colorful book). Identifying a more suitable method for finding emphasis in movement remains for future research. We have shown that tutors do use visual cues to attract attention ([6], [12]). However, we still do not know how this is synchronized with speech in detail and how this can be detected automatically.

C. *Visualization and Inspection*

For research on synchrony we found it crucial to have a visual representation of the data in order to see relations between modalities on first sight. Figure 4 shows our visualization and inspection tool. The top row (red) shows the amount of pixel change in the visual data over time. As can be seen, action demonstrations result in motion peaks which can be segmented into motion segments. The row below usually shows the color saliency based tracking results. However, this row is empty because, as has been mentioned above, color based tracking was not used here. Also a second row is empty. It usually displays the x- and y coordinates (over time) of the trajectory corresponding to the most salient region. Again, this row is not relevant because color saliency was not used. The next row shows the segmentation into speech (dark blue) and non-speech (white). The light blue bars within each speech segment indicate the level of stress assigned to each syllable. The highest light blue bar indicates the most prominent syllable in the utterance. The main role of this row and the last two rows is to visualize the hypotheses as time intervals coming from the acoustic segmentation, the visual action segmentation, and the temporal association module. Below the speech segmentation, the temporal extensions of the motion peaks are displayed in green. The white lines between the green boxes indicate motion boundaries. The bottom row (purple) visualizes the acoustic packages formed according to the temporal association algorithm. White areas indicate that no significant motion has been performed which is temporally overlapping with speech.

III. Experimental User Study

We conducted an experimental Wizard of Oz type user study with the robot BIRON (Bielefeld Robot Companion) [13]. This research is explorative and not guided by hypotheses because, as has been mentioned in Section I, Acoustic Packaging is a bottom-up approach and we do not presuppose any knowledge of people's actions. We invited pairs of users to complete the task together in order to go beyond previous work on teaching interactions and to learn about how teaching a robot works within small groups. Using the approach of remote controlling the robot's behavior based on pre-defined scripts, we could generate standardized conditions that are explained in the following.

A. *Experimental Conditions and Scripts*

The goal of the study was to find out whether the participants' behaviors differed if the robot learned quickly or slowly and how the participants in the pair aligned to each other. Therefore, all participants completed a positive and a negative trial (within-subjects design) in order to determine whether the structure of the acoustic packages varied between the trials. Positive and negative trials differed in the amount of objects that the robot learned correctly in a given time. Overall, each script consisted of 20 robot utterances. In the case of "success" the object was recognized correctly and the robot replied "This is the [object] then". In the case of a "failure" the robot inserted the wrong object name in the same utterance. The scripts of the positive trial contained 10 instances of success (50% of all utterances) and 3 of failure (15% of all utterances). The scripts of the negative trials contained 5 instances of success (25% of all utterances) and 8 failures (40% of all utterances). In the negative trial the wizard went through the script twice in order to give the participants the chance to teach all ten objects to the robot. The order of the trials was counterbalanced. Based on findings from previous research [3], all other utterances were distributed equally in all scripts ("Pardon?" twice, "I don't know the word" once, etc.). All utterances were translated for this paper as the study was conducted in German. Additionally to the utterances in the scripts, the participants were greeted and asked for their names and the robot said good-bye.

Moreover, the robot displayed a pair of eyes on its screen. The eyes were also remote controlled and were directed at the person who was speaking. If nobody was speaking, the wizard made the eyes look straight (between the participants) or at the table with the objects. The gazing behavior was the same in all conditions.

B. *Setup*

The setup was identical in all experimental conditions. The participants and the robot BIRON stood around a round table (see Figure 1). The participants' approximate position was marked on the floor in order to get them to stand equally far away from the other person as from the robot and to avoid effects of different distances. We put the objects that the participants had to teach to the robot (e.g., book, plate, scissors) on the table. The wizard controlling the robot was sitting behind a wall in the back of the participants who could not see him. He controlled the robot by clicking on options in a remote control interface.

We positioned two video cameras in the corners opposite of the participants. Each of the cameras recorded a frontal view of one person, the other person from the side, the robot also from the side, and the table. For more information about the setup and the scripts see [4].

C. Procedure

On arriving, we welcomed the participants and provided them with a written introduction. Thereafter, they could ask questions about the study, signed a consent form, and received a first questionnaire. This questionnaire included items on demographic information, relationship to the other person and the robot, experience with computer and robot usage, and expectations toward the robot in the given scenario. Thereafter, the participants interacted with the robot. Person A was asked to start the interaction. He or she greeted the robot and began to show objects. No information on how to teach the objects was provided to the participants.

As mentioned above, the users completed two trials: a positive one in which the robot performed well and a negative one in which it performed badly. The necessity of the two trials was explained by telling the participants that the robot needed to learn various sets of objects because objects in real life differ in appearance. Thus, the objects were exchanged between the trials. The two sets contained the same items with different appearance. After the trials, the participants completed a second questionnaire to evaluate the robot and the interaction. Finally, they were paid for participation.

D. Sample

40 pairs of participants (80 participants) took part in the study (50 female, 30 male). Of the 40 pairs, 17 were female/female, 16 female/male, and 7 male/male. One of the pairs had to be excluded from the analysis. Most participants were students. Their age ranged between 20 and 57 years (mean age=25.46, standard deviation (sd)=5.90). They came from all kinds of disciplines (linguistics, electrical engineering, law, etc.). Their mean experience with computer usage was 2.98 (all means on a scale of 0 to 4) (sd=0.80), with computer programming 0.86 (sd=1.18), and with robot usage 0.51 (sd=0.87). So people had some experience with computers but hardly any with programming and robots.

The mean duration that participants had known each other was 44 months (3 years and 8 months, minimum 0 months, maximum 296 months, sd=61).

IV. Data Analysis

Overall, the corpus that we acquired in the study contained 4588 annotations of utterances, 2924 in the negative and 1664 in the positive trials. This equals roughly 60 utterances per person in both trials. The mean number of utterances per person in the negative trials (37.49; sd=15.30) was higher than in the positive trials (21.33; sd=7.41). This was due to the fact that the interactions in the negative trials took longer because the participants needed twice as many utterances to teach all objects to the robot. In the following we describe how we analyzed the data with the Acoustic Packaging approach.

A. Procedure of the Analysis

The analysis was based on the annotation of speech and on the videos that were acquired with the two cameras. For each person, the frontal view video was included in the analysis. As the experimenter told the participants where to stand, there was little variation between the videos and the relevant part could be cropped. This was necessary because the videos contain parts in which the other participant is visible and, as has been mentioned in Section II-A.2, the Acoustic Packaging approach takes the whole video into account. Thus, a crop area was defined manually. It remained the same for all videos. On this area the visual action segmentation was employed.

The motion segments detected in the videos were associated with speech segments by the temporal association module system according the previous description (see Section II-A.3). Regarding audio segmentation, as has been mentioned in Section II-A.1, annotation was used to segment the audio into utterances. Utterances were further segmented into syllables and the prominence rating was calculated for each syllable by the prominence detection module (see Section II-B.1). Utterances including their syllable segmentation were associated with motion segments by the temporal association module. After running the Acoustic Packaging module on all data, the statistics described in the following were calculated.

B. Measures and Statistic Tests

To compare the structure of the acoustic packages (APs) between participants in a pair and between the two trials of one participant, we analyzed the following measures: mean number of utterances in APs per participant, mean number of APs per participant, mean number of MPs in APs per participant, mean number of MPs per AP, mean length of APs, mean length of MPs in AP, mean height of MP in AP, mean length of utterances in AP, and maximum deviation from mean prominence. All lengths are calculated in seconds. The height of motion peaks refers to the pixels that have changed in a picture in relation to the overall number of pixels in the picture. Thus, the value is normalized. The prominence rating refers to the energy of a certain frequency band. The maximum of the rating is 1.

For the comparison between positive and negative trials for each participant we conducted Asymptotic Wilcoxon Mann-Whitney Rank Sum Tests because not all data could be assumed to be normally distributed and the test does not make this assumption. For the comparison between the two participants in a pair we calculated Intraclass Correlations (ICC). These take into account that we randomly decided which participant was in position A or B (A being to the left of the robot and B to its right).

The acoustic packaging method also allowed us to determine which of the words were most prominent in the users' utterances and to compare these between the trials. We added up the numbers of times any object name was most prominent, words that were synonyms (e.g., different ways of praising the robot such as ”correct”, ”good”, ”right”), and articles in female and male form. We removed alignments that went wrong, i.e., alignments where the system found the most

TABLE I

OVERVIEW OF THE RESULTS

(numbers in parentheses refer to standard deviations, the Z-test refers to differences between the positive and negative trials, the Intraclass Correlations (ICC) are presented for all trials and for the positive / negative trials as a comparison between participants in a pair)

measure	all trials	pos. trials	neg. trials	Z	p	ICC all trials	ICC pos. trials	ICC neg. trials
mean number of utterances in APs / participant	29.41(14.47)	21.33(7.41)	37.49(15.30)	8.40	.00	.76**	.10	.70**
mean number of APs / participant	22.76(12.15)	16.95(6.74)	28.56(13.53)	6.78	.00	.62**	.06	.52*
mean number of MPs in APs / participant	38.23(27.07)	26.81(14.52)	49.65(31.63)	5.80	.00	.68**	.41*	.63**
mean number of MPs per AP	1.59(0.41)	1.52(0.36)	1.66(0.44)	1.98	.05	.55**	.49**	.57**
mean length of APs	2.50(0.67)	2.45(0.69)	2.55(0.65)	1.34	.18	.78**	.86**	.69**
mean length of MPs in AP	1.36(0.17)	1.38(0.20)	1.33(0.14)	−1.63	.10	−.09	−.08	−.24
mean height of MPs in AP	0.04(0.01)	0.04(0.02)	0.03(0.01)	0.03	.98	.57**	.54**	.64**
mean length of utterances in AP	1.49(0.70)	1.41(0.72)	1.56(0.68)	1.79	.07	.78**	.86**	.68**
maximum deviation from mean prominence	0.47(0.08)	0.45(0.07)	0.48(0.08)	2.18	.03	.60**	.59**	.59**

prominent part of an utterance between two words. This error rate was 4% which is quite low given that the corpus was recorded in a noisy environment with, e.g., the robot making functional sounds. All statistical tests were calculated using the R software package [7].

V. RESULTS

Table I gives an overview of the results. It presents the measures for the overall data set and for the single trials.

Each participant used a mean number of 29.41 (sd=14.47) utterances per trial. The mean number of APs detected in these utterance was 22.76 (sd=12.15). Thus, in 77% of the utterances an AP was detected or in other words some movement of the user coincided with the utterance. Hence, participants did not move in the other 23% of the utterances. The overall number of motion peaks (MPs) within all APs per trial was 38.23 (sd=27.07) which results in a mean number of MPs per AP of 1.59 (sd=.41). All these numbers show a high standard deviation which points to large variations between the participants. However, the mean number of APs, MPs in APs, and utterances per participant were strongly correlated within pairs when looking at all trials (ICC =.62**, .68**, .76**, respectively). This points to an alignment in the pairs that will be discussed in more depth later in the paper.

In the next section we first compare the APs of the positive and the negative trials to determine whether such manipulation of the situation has an influence on the way people tutor the robot. Thereafter, we compare the relationship of behaviors of the two people in a pair.

A. Comparison of APs in the Positive and the Negative Trials

One result that has already been mentioned is that the mean number of utterances was higher in the negative trials (see Section IV). The table shows that this difference was actually statistically significant (Z=8.40, p<.01). The same was true for the mean number of APs per participant (Z=6.78, p<.01) and for the mean number of MPs in APs per participant (Z=5.80, p<.01). The table shows that the mean length of the APs did not differ significantly between the trials. Thus, the ratio between the length of the motion packages and the utterances stayed the same. As has been explained above, the APs are determined by the users' motions and their utterances. For the overall APs, no difference between the motion peaks was found with respect to their length and height between the trials. However, there was a (non-significant) trend that the utterances in the APs in the negative trials were longer than in the positive trials (1.56 seconds vs. 1.41 seconds; Z=1.79, p=.07). One explanation for this is that the users explained more in the negative conditions because they guessed that the robot needed more information to learn the objects. Moreover, the mean number of MPs per AP was higher in the negative condition (1.66 vs. 1.52; Z=1.98, p=.05) which is in line with the longer utterances. Generally, shorter utterances are higher structured. Thus, we can conclude that the interaction in the negative condition was less structured.

TABLE II

PROMINENT WORDS IN THE POSITIVE AND THE NEGATIVE TRIALS

rank	pos.trials	count	neg. trials	count
1	object names	400(25.84%)	object names	823(31.51%)
2	this	218(14.08%)	this	328(12.56%)
3	praise	209(13.50%)	praise	205(7.85%)
4	yes	108(6.98%)	no	160(6.13%)
5	a	90(5.81%)	yes	158(6.05%)
	all prom. words	1548 (100%)	all prom. words	2612 (100%)

Furthermore, we compared the maximum deviation from the mean prominence rating. This value was significantly higher in the negative trials than in the positive trials (0.48 vs. 0.45; Z=2.18, p=.03), pointing to the fact that single parts of the utterance were more prominent.

We were also interested in which words the participants stressed in both trials. Table II gives an overview of the five most prominent words / categories of words. These accounted for about 65% of all prominent words in both trials. The words actually display the most common sentence in the task: "This is a [object name]" and the praise if the robot did well. Alone the object names and the "this" added up to about 40% of the most prominent words in both trials. There was a clear tendency towards praising the robot more prominently in the positive trials (13.50% vs. 7.85% of all prominent words). Another obvious difference between the five most prominent categories of words, was the usage of "no" in the negative trials. Surprisingly, also the word "yes" was prominent very often in the negative trials. Anyway, the statistical distribution of prominent words in the interaction could actually help the robot to determine how well the interaction is going.

B. Comparison between the Participants in the Pairs

As has been mentioned before, the ICC for all trials was high for the number of APs, MPs in APs and utterances. The ICC for the number of MPs in APs was still significant when looking at the positive and the negative trials individually (ICC=.41* for the positive trials and .63* for the negative trials). However, this was not true for the number of APs and utterances per participant. Both of these were not correlated in the positive trials, but they were correlated in the negative trials (ICC=.52* and .70**, respectively). Thus, the participants aligned more in the negative trials with respect to these measures which was probably mainly due to the fact that the number of utterances became more similar.

While we found many strong correlations within the pairs, all the measures were not correlated at all when calculating ICCs for all participants. Thus, the effects really point to alignment within the pairs and it can be excluded that all participants in the experiment acted in the same way.

Table I furthermore shows that the mean length of APs was strongly correlated for both trials (ICC=.78** overall, .86** in the positive trials, and .69** in the negative trials). Also the length of the participants' utterances within the APs was strongly correlated (ICC=.78** overall, .86** in the positive trials, and .68** in the negative trials). The mean number of MPs per AP was strongly correlated for the overall data set and the individual trials (ICC=.55** overall, .49** in the positive trials, and .57** in the negative trials). Thus, the participants aligned with respect to this feature in both trials. Hence, we can assume that, overall, the participants in a pair align with respect to structuring their actions. Surprisingly, the mean lengths of MPs in the APs were not correlated at all, even though the mean heights were (ICC=.57** overall, .54** in the positive trials, and .64** in the negative trials). We cannot offer a conclusive explanation for that but will have to conduct additional qualitative analyses. Possible explanations might be person specific properties, artefacts of the perspective and the resulting size of the participants in the video.

Finally, we analyzed the maximum deviation from the mean prominence. This correlation was strong overall and equally strong in both trials (ICC=.60** overall, .59** in the positive trials, and .59** in the negative trials). Thus, both participants had similar "patterns" of stressing prominent words.

VI. Discussion and Conclusion

We set out to show how the Acoustic Packaging approach can help to research alignment in groups of humans and robots, to structure human actions for robots and, thus, to support the generation of appropriate robot feedback. Indeed, we found that the way people tutor the robot differed between a positive situation, in which the robot learned object names quickly, and a negative situation, in which it needed more turns to learn. Particularly, the users' utterances were longer in the negative trials, the utterances contained more motion peaks, and the maximum deviation from mean prominence was higher. Also the patterns of which words were prominent in the utterances differed between the trials.

With respect to the two participants in a pair, we found a strong alignment with the exception of the mean lengths of the motion peaks in the acoustic packages and the number of motion peaks per acoustic package in the positive trials. Overall the alignment was stronger in the negative trials.

In the short term the Acoustic Packaging approach can be used to give immediate feedback about the associations that the system has made, e.g., by replaying prominent syllables or adapting the timing of the robot utterances to the prominence detection. This goes beyond the current analysis by employing the system online in interaction (which has been done already in other contexts). In the long run, the approach can be used to differentiate between properties such as (non-)successful interaction. If statistics indicate that the interaction is not running well for some time, the behavior of the system can be adapted, e.g., by simplifying dialog responses or providing more explicit feedback. This adaptation could be enhanced by contextual knowledge. Thus, Acoustic Packaging allows to adapt the interaction to the capabilities and knowledge of the users by identifying the current course of the interaction. Future work will also research whether the approach can be deployed to determine not only the current success of the interaction but also personality specific behavior that allows for an adaptation of robot behavior that is even more tailored to individual users. This might come with the need for a more sensitive system that includes further cues such as facial expressions and eye gaze.

References

[1] J. W. Davis and A. F. Bobick. The Representation and Recognition of Human Movement Using Temporal Templates. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928–934, San Juan, Puerto Rico, 1997.

[2] G. A. Fink. Developing HMM-Based Recognizers with ESMERALDA. In V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, pages 229–234. Springer, 1999.

[3] C. Lang, M. Hanheide, M. Lohse, H. Wersing, and G. Sagerer. Feedback interpretation based on facial expressions in human-robot interaction. In *RO-MAN*, pages 189–194. IEEE, 2009.

[4] M. Lohse. Treating robots as social beings - a matter of personal preconceptions or interpersonal alignment? In *21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012.

[5] P. Mermelstein. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 58(4):880–883, 1975.

[6] K. Pitsch, A. L. Vollmer, J. Fritsch, B. Wrede, K. Rohlfing, and G. Sagerer. On the loop of action modification and the recipient's gaze in adult-child interaction. In *Gesture and Speech in Interaction*, Poznan, Poland, 2009.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2011.

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

[9] L. Schillingmann, B. Wrede, and K. J. Rohlfing. A Computational Model of Acoustic Packaging. *IEEE Transactions on Autonomous Mental Development*, 1(4):226–237, Dec. 2009.

[10] L. Schillingmann, B. Wrede, K. J. Rohlfing, and K. Fischer. The structure of robot-directed interaction compared to adult- and infant-directed interaction using a model for acoustic packaging. *Spoken Dialogue and Human-Robot Interaction*, 2009.

[11] F. Tamburini and P. Wagner. On automatic prominence detection for German. In *Interspeech 2007*, pages 1809–1812, 2007.

[12] A. L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wrede. People Modify Their Tutoring Behavior in Robot-Directed Interaction for Action Learning. In *International Conference on Development and Learning*, Shanghai, China, 2009.

[13] S. Wachsmuth, F. Siepmann, D. Schulze, and A. Swadzba. Tobi - team of bielefeld: The human-robot interaction system for robocup@home 2010. 06/2010 2010.

[14] J. M. Zacks and K. M. Swallow. Event Segmentation. *Current Directions in Psychological Science*, 16(2):80–84, Apr. 2007.