The Structure of Robot-Directed Interaction compared to Adult- and Infant-Directed Interaction using a Model for Acoustic Packaging

Lars Schillingmann[1 2], Britta Wrede[1 2], Katharina Rohlfing[2 3], Kerstin Fischer[4] and Gerhard Sagerer[1 2]

[1]Applied Informatics Group, Faculty of Technology, Bielefeld University, Germany
[2]Research Institute for Cognition and Robotics, Bielefeld University, Germany
[3]Emergentist Semantics Group, Center of Excellence, Cognitive Interaction Technology, Bielefeld University, Germany
[4]University of Southern Denmark, Denmark
{lschilli, bwrede, rohlfing, sagerer}@techfak.uni-bielefeld.de
kerstin@sitkom.sdu.dk

***Abstract*— As it is often assumed that an interaction with a robot is similar to an interaction with an infant, one would expect similar characteristics in tutoring behavior in human-robot interaction from which the robot's learning processes can benefit. Especially learning actions can profit from the use of language. As it has been shown by Hirsh-Pasek and Golinkoff introducing the idea of Acoustic Packaging, language not only serves as a social cue but also provides functional structure.**

In order to better understand how robot-directed interaction is structured and how the modifications may help to extract cues relevant for learning actions from a tutoring situation, in the present paper we analyzed the structure of Acoustic Packages in the adult-robot situation. Our results indicate that in human-robot interaction, the phenomenon of Acoustic Packaging is similar in adult-child interaction: In both situations, more Acoustic Packages with less content can be found, compared to adult-adult interaction. This can be interpreted as an encouraging result towards the goal of building robots that can learn from a tutoring situation.

I. Introduction

Envisioning a robot that learns actions from a human, we need to realize its ability to understand where an action starts an ends in a continuous stream of multi-modal information. Such an ability, as we can see it in children, provides sense to actions demonstrated to them, so that children eventually can imitate or emulate them. Indeed, already at a very early age they begin to interpret actions [1], [2]. This skill of interpretation is supported on the one hand by the infant's ability to use synchronous information from multiple modalities. According to the *Intermodal Redundancy Hypothesis*, information picked up by different senses can provide an important basis for perception of unitary events [3]. This means that the child picks up overlapping, redundant information for objects and events from the environment. Gogate and Bahrick showed that when 7 months old infants were presented a syllable with a synchronous movement of the labeled objects, they could learn this syllable more easily and map it onto the presented object than their peers receiving an asynchronous presentation [4].

In accordance with this, the infant's social environment presents specifically designed input to the infant, which appears to be well timed with the infant's cognitive development. For example, Gogate and colleagues could show that parents decreased their use of synchrony in an object-labeling task when the infant increased its lexical abilities [5].

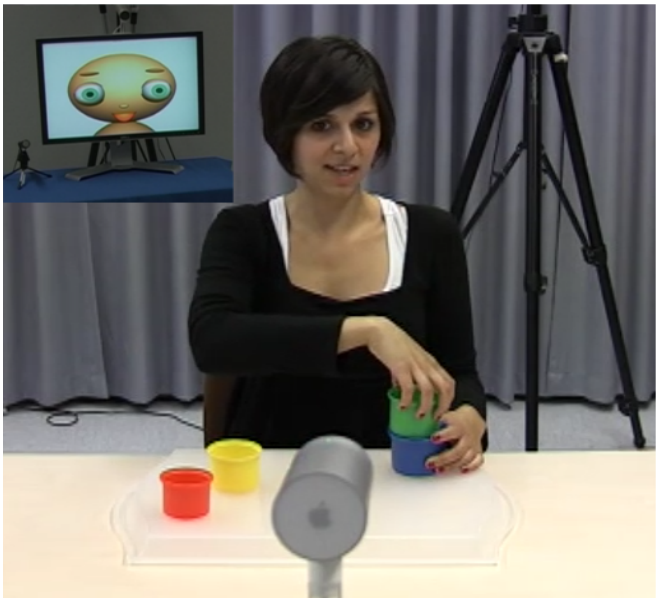

Fig. 1. A test subject demonstrating how to stack cups to a robot simulation. The robot simulation is shown in the top left for illustration purposes.

Synchrony has not only been shown to be helpful in learning word-referent relations but also for learning actions. Hirsh-Pasek and Golinkoff [6] propose that special properties of the sound signal may help infants to attend to particular units within an action stream. They call this phenomenon Acoustic Packaging. They argue that children can use the information provided by the Acoustic Packages to perceive a link between specific sounds and visual events, thus helping to segment the action into meaningful and non-meaningful parts. Indeed, it has been shown that infants are able to bundle different sub-actions into a meaningful unit depending on the position of the accompanying speech stream [7]. In the study, 7.5 to 11.5 months-old infants were first familiarized with video sequences showing short action clips. The acoustic input coincided with portions of the action stream and thus "packaged" pairs of clips together. During the test, infants viewed packaged and non-packaged

pairs of actions side-by-side in silence. The results of the study showed that 9.5 month-olds looked longer at non-packaged action sequences. This suggests that acoustic input (i.e. narrations heard during familiarization) influenced the way of how infants perceive the action units [7]. These are promising results as they might help robots to interpret action sequences based on the synchrony of the acoustic and the visual channel.

In line with this, it has been suggested that by using modifications in the tutor's behavior, a robot could learn to detect the meaningful structure of the demonstrated action [8], [9]. However, we do not yet know whether robots receive similar input as infants or what behavior of the learner actually elicits such a tutoring behavior. Nagai et al. [10] assume that a robot can trigger such behavior possibly because of its immature cognitive capabilities. However, this assumption still needs to be verified experimentally and the question remains open how the robot has to behave in order to receive the input it needs. Recently, a study by Herberg and colleagues [11] investigated whether people will modify their actions for computers. They presented a picture of an interaction partner – either a child, an adult, or a computer – to the subjects and asked them to demonstrate objects. Results showed that subjects did indeed modify their actions when pretending to speak to a computer, and these modifications differed from their pretended interactions with an adult or a child. The authors interpret these differences with respect to different conceptualizations of the interaction partner: while persons are assigned the capability of reasoning about goals, the subjects would not assign this capability to a computer.

However, it is questionable if this experimental design did indeed invoke different conceptualizations of the interaction partner at all: It has been shown that subjects, when asked to pretend to speak to an (imaginary) infant, were able to modify their speech but failed to produce characteristic features of motherese as observed in natural adult-infant interactions [12]. In addition, this study might not be applicable to (humanoid) robots as it was carried out with a computer. In an fMRI study, Krach et al. [13] show that interactions with a computer produce different activation patterns in human subjects than interactions with a humanoid robot. In detail, when subjects believed that they were interacting with a humanoid robot, they had a significantly higher activation in brain areas generally associated with theory-of-mind (that is, reasoning about the interaction partner's intentions) than when they believed they were interacting with a computer.

Thus, the question remains, if and how humans modify their actions when interacting with a robot. More specifically, if they will structure their demonstrated actions – in terms of Acoustic Packages when interacting with a robot. More specifically, the question is whether they will structure their demonstrated actions – in terms of Acoustic Packages – in the same way as when interacting with infants.

In this paper, we use our Acoustic Packaging model [14] to analyze adult-robot interaction as opposed to adult-child and adult-adult interactions. If robots are offered actions in a similarly structured way as infants, we can draw from these findings on Acoustic Packaging in order to develop algorithms to detect meaningful action in the interaction stream. Previous work shows that tutoring behavior affects many modalities. We were already able to show that our Acoustic Packaging model is able to reflect the structural differences between tutoring in adult-adult and adult-child interactions [14]. In the following, we will analyze structural properties of the Acoustic Packaging results also on adult-robot interaction.

II. Our Approach to Acoustic Packaging

The development of a computational model for Acoustic Packaging benefits from understanding the characteristic features of infant directed interactions. Brand et al. have been the first to report experimental evidence for modifications in the demonstration of action towards infants [15]. These manually coded modifications have been verified by automatic measurements of the velocity of hand movements, their roundness and pace, i.e. the relationship between the duration of a movement and the preceding pause [9]. The results indicate that infant-directed actions have more pauses and are less round than adult-directed actions. Similarly, the speech exhibits more pauses. Moreover, it has been found that speech and action segments tend to coincide more frequently in infant-directed interaction [16]. More recently, it has been shown that these characteristics of motionese are even stronger in interactions with a robot as compared to infant-directed interactions [17].

In order to derive meaningful event segments (acoustic packages) it seems thus promising to use very simple features that are able to detect pauses in the speech and the vision (or rather motion) channel.

The development of robots that are able to interact with humans in tutoring situations, requires methods to segment actions into meaningful parts. We transfer the concept of Acoustic Packaging from developmental research in order to create a software module that is able to fulfill two important tasks in human-robot tutoring situations. The first task is to deliver bottom-up segmentation hypotheses about the action presented. The second task is to form early learning units containing multimodal information. These units can be processed further by other modules that infer models about the actions currently presented. Hirsh-Pasek and Golinkoff describe a minimal and a maximal role Acoustic Packaging can take [6]. In the minimal role, acoustic packages are formed on repetition of an acoustic chunk in conjunction with a particular event. In their maximal role, Acoustic Packaging can fuse separate events into meaningful macroevents. Our approach aims towards the maximal role of Acoustic Packaging.

A. Requirements

As a first step towards the development of a computational model of Acoustic Packaging the *segmentation problem* has to be solved. Since the model has to make use of at least one visual and one acoustic cue, a temporal segmentation

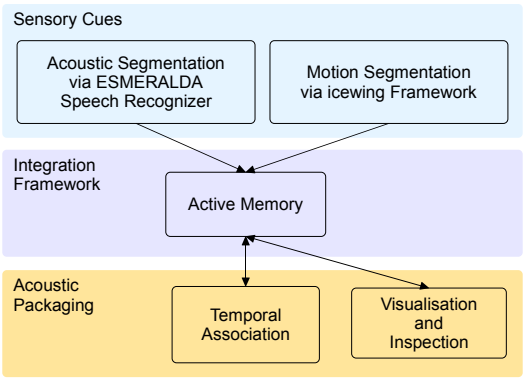

Fig. 2. System overview with highlighted layers and their relation to the Acoustic Packaging system.

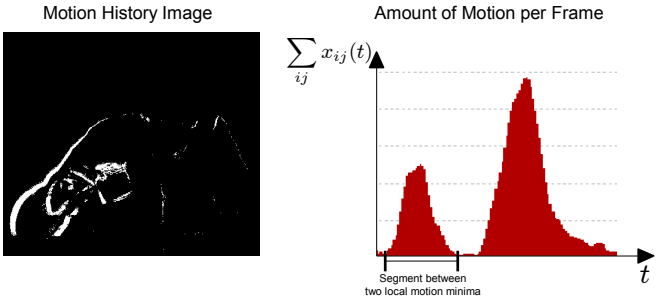

Fig. 3. The left image shows a motion history image from a person showing a cup. The right image illustrates our approach to visually segment actions via the amount of motion per frame.

for both cues is required. We address the visual and acoustic segmentation problem in section II-D and II-C.

A second problem is the *temporal synchronization* of these sensory cues. Hypotheses from audio and vision processing are typically generated neither at the same time nor in the same rate. Temporal synchrony itself can be considered as an amodal cue, which provides information about what segments should be packaged. A *timestamp concept* addresses the amodal property and is used in the Acoustic Packaging process in order to associate the different cues.

Another requirement concerns the architecture which should be *extensible*. The integration of additional cues or modules that perform further processing towards learning on the acoustic packages should be facilitated by the architecture. Since a socially interactive robot should give feedback during tutoring, the system should be *online* usable and able to cope with updating hypotheses. Another important aspect is the support for *visualization* and *inspection* of the involved cues and the resulting acoustic packages. It provides means for debugging and evaluating the Acoustic Packaging system.

B. System Overview

Our system for Acoustic Packaging proposed here consists of four modules (see Figure 2). These modules communicate events through a central memory, the so-called Active Memory [18]. The Active Memory notifies components about event types they have subscribed to and is able to store these events persistently. It is thus an integration framework that supports a decoupled design of the participating modules facilitating integration of further processing modules. This addresses the requirement of extensibility.

All signal processing modules are connected to the Active Memory. The audio signal is processed using the ESMERALDA speech recognizer [19], which is configured to use an acoustic model for monophoneme recognition. Phonotactics are modeled statistically via an n-gram model. The visual signal is processed with the help of a graphical plugin environment [20].

C. Acoustic Segmentation

Based on the observation that infant-directed actions exhibit more, and more structured pauses, it seems appropriate to segment the acoustic signal simply into speech and non-speech (pause) segments. Yet in a relatively noisy environment such as the experimental setting, the separation of speech from non-speech is a difficult task. Therefore, instead of simple voice activity detection, we used a more sophisticated approach: We defined an acoustic segment as speech framed by non-speech. As a consequence, a continuous chain of phoneme hypotheses generated by the speech recognizer is considered as a speech segment. Our speech recognizer inserts phoneme hypotheses as well as the corresponding audio signal into the Active Memory. As the recognition process is incremental the hypotheses are continuously updated during processing of an utterance.

D. Visual Action Segmentation

Similar to the acoustic segmentation, we segment the visual signal into motion and motion-pauses. In our model, the visual signal is segmented into motion peaks where each peak ranges between two local minima in the amount of change in the visual signal. For example, if someone shows a cup, there is typically a motion minimum at the point where the cup is hold still or slowed down for a short moment. When the cup is accelerated again, on its way to be put on the table, a local maximum in the amount of motion can be observed. Another local minimum occurs when the cup is eventually put on the table. This observation is the motivation for our heuristic approach to segment actions into motion peaks.

This segmentation into motion peaks is technically realized by an approach based on motion history images [21]. The amount of motion is calculated per frame by summing up the motion history image (see Figure 3). In the amount of motion local minima are detected with the help of a sliding window that is updated at each time step. 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 can be suppressed by choosing the window size accordingly. Our current model considers the complete image when detecting local motion minima. It is therefore also sensitive to motion in the video that is not related to the demonstrated action – which can be coped with by ignoring certain parts of the image.

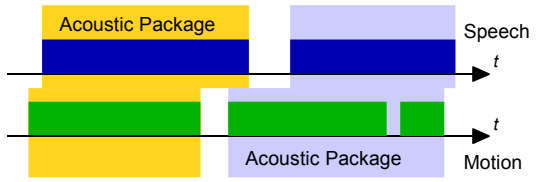

Fig. 4. Motion and speech intervals are assigned to an acoustic package if they overlap. The middle motion interval has been assigned to the second acoustic package due to greater overlap.

When a local minimum is detected then an event describing the motion peak between the previous and the current motion minimum is inserted into the Active Memory. The description contains the peaks' time interval and the frames at the minima from the beginning and end of the motion peak.

E. Temporal Association

As already pointed out as a requirement, both the motion peaks and the speech segments need to be temporally associated in order to form acoustic packages. Our temporal association module subscribes to events on the Active Memory and maintains a timeline for different types of time intervals. In our current version of the system motion peaks and speech segments are processed. When a new event arrives, the segment is aligned to the timeline. In the next step, the temporal relations to the segments on the other timeline are calculated for which a subset of the relations defined in [22] is used. 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 4 for the association process). If hypotheses from the signal processing modules are updated (e.g. a speech segment is extended), the corresponding acoustic package is updated as well. The temporal association module has to process a large number of events. These events can either be new hypotheses or updates of existing hypotheses. Since our aim is to process these events online, this approach requires inserting and updating of incoming time intervals to be handled computationally efficient: Each incoming time interval has to be aligned to the timelines of the other modality. Furthermore, the module should allow asynchronicity between the incoming events of the different modalities. This, on the one hand, requires handling potential processing delays, on the other hand it eases debugging and offline processing. Since the hypotheses for each modality are generated in independent processes, the association module should not rely on the order of events. The strategy, which addresses these requirements is explained in the following.

Maintaining a structure, which preserves the order of time intervals is a central concept of the temporal association module. For example, the timeline for speech contains intervals with the hypotheses of the speech recognizer. Since intervals of a single timeline have the property of being sorted and do not overlap, the insertion point can easily be found by performing a binary search on the timeline. The same method is used when modalities are associated in the process of forming Acoustic Packages. In the case of an incoming speech interval, the insertion point of the speech interval in the motion timeline is determined. After that, the temporal relations of the speech interval to each interval in the local neighborhood in the motion timeline are calculated. Motion peaks overlapping with the speech intervals are associated to the same Acoustic Package as the speech interval or a new Acoustic Package is created. If a motion peak is already associated with an acoustic package, the motion peak is reassigned depending if it has a larger overlap with the current speech interval. In case of an incoming motion peak, the same procedure is applied. The insertion point of the motion peak in the speech timeline is determined, and the motion peak is associated to the Acoustic Package with the most overlapping speech interval. The construction and update of packages is mirrored into the Active Memory. This accords with the idea of an online usable system.

F. Visualization and Inspection

Since the temporal synchrony is one important cue for this system, tools are needed that analyze the Acoustic Packaging process and the temporal relations of the involved sensory cues. Figure 5 shows our visualization tool, which monitors events. Other processing modules communicate them to the Active Memory. The first plot displays the amount of motion over time. The second row shows the signal energy that gives an estimate about speech activity. The third row visualizes the hypotheses as time intervals coming from the acoustic segmentation, the visual action segmentation, and the temporal association module. More specifically, the first line displays the speech recognition results: The lighter areas mark non-speech hypotheses like for example noise. The second line displays the temporal extensions of the motion peaks. The third line visualizes the results of the Acoustic Packaging module. Since the case is possible, that under certain conditions the temporal extensions of two neighboring acoustic packages overlap, only the range of motion peaks (which have been associated to one acoustic package) is visualized currently.

In fulfilling the requirement of support for visualization and inspection Figure 6 shows our inspection tool, which is able to query speech and motion peak hypotheses from the Active Memory. In conjunction with the tool for visualization of the cues (Figure 5), it is possible to inspect hypotheses persistently stored in the Active Memory. The time intervals selected currently in both, the visual and the acoustic cues, are highlighted which enables inspection of their temporal relations. The inspection tool displays the frames at the beginning and the end of the selected motion peak. The speech segment can be replayed or resynthesized. This resynthesis uses the phoneme chain displayed in the bottom right corner. However it is a current topic of development. Taken together, these features of the inspection tool help to rate, optimize and debug the Acoustic Packaging system and its parameters.

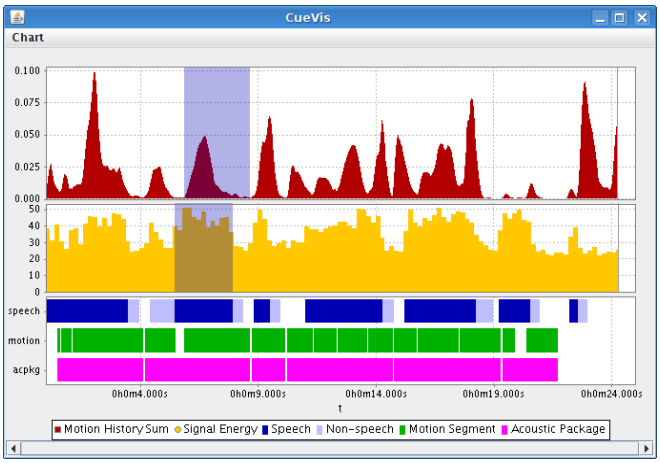

Fig. 5. Cue visualization tool.

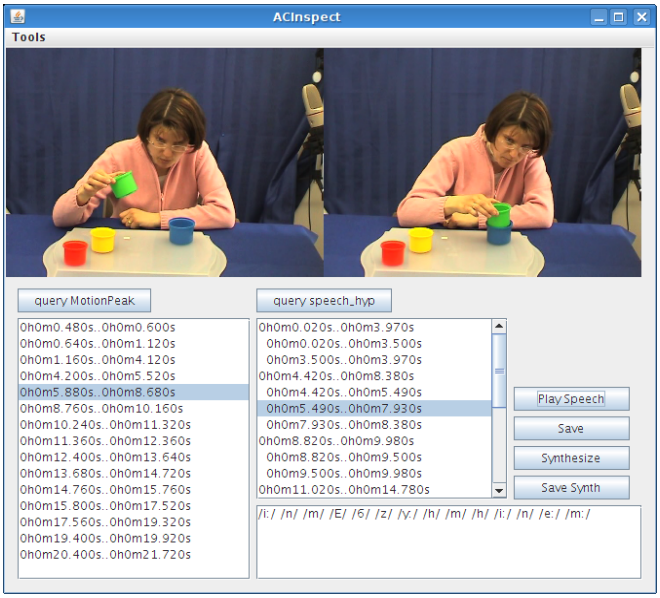

Fig. 6. Inspection tool.

III. Evaluation

Our goal here is to analyze the structural properties of the Acoustic Packages in adult-robot (ARI), adult-child (ACI) and adult-adult (AAI) interaction. We used the Acoustic Packaging system described above, to temporally segment the data and associate the modalities.

A. Data Description

The data consists of two corpora. One is a corpus containing video and audio data with adult- and infant-directed interactions [9]. From this corpus, we selected 26 subjects with 8 to 11 months old children. The subjects were asked to demonstrate functions of 10 different objects to their children as well as to another adult (partner or experimenter, Figure 8 illustrates the experimental setting). In this evaluation, we focus on one task, namely the stacking cups (see Figure 1).

The other corpus [17] contains 31 German interactions between human participants and a simulated robot using

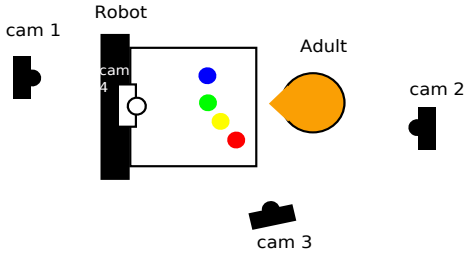

Fig. 7. Adult-Robot Interaction Setting. The test subject is facing the robot simulation, which is displayed on a screen. In this evaluation, recordings from camera 1 are used.

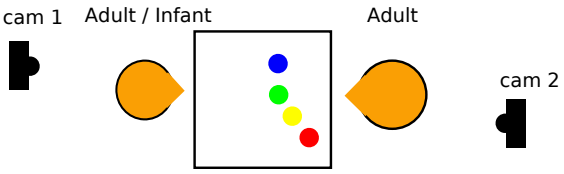

Fig. 8. Adult-Child / Adult-Adult interaction setting. The interaction partners are seated at a table facing each other. In this evaluation, recordings from camera 1 are used.

the same tasks as in the corpus with ACI/AAI (Figure 7 illustrates the experimental setting). The robot is a simulation of a child-like face that is presented on a screen, whose eyes are moving according to a saliency model [10]. Thus, the eyes focus on salient points, like moving or colorful objects. We selected the 16 subjects who performed the stacking cups task comparable to the subjects in the corpus with ACI/AAI.

For better comparison of the actions across subjects, a single task presentation was selected from each video. The Acoustic Packaging system processed the data starting with 2 seconds before the first cup was lifted until 2 seconds after the last cup has been stacked by the subject.

B. Results

We have taken several measurements on the Acoustic Packaging results (see Table I). Each row contains a measurement calculated on the three types of interaction described above. The results are averaged over the number of test subjects in each group. An asymptotic Wilcoxon Mann-Whitney rank sum test was performed to check which measurements show significant differences between the interaction types. We will explain and interpret the results in the following.

Firstly, results concerning the speech and visual modality are presented. After that, we will take a look at Acoustic Packages and their structural properties.

Utterances: There is a significant difference in the number and length of utterances (cp. row 7, 8): ARI (M:10.17) $>$ ACI (M:7.34) $>$ AAI (M:5.48). This shows that the verbosity is much higher in ARI than in ACI and AAI.

Pauses: The average length of pauses is not significantly different between ACI and ARI but between AAI and ARI (cp. row 12). There is a trend for ARI (M:1.23) $\gtrsim$ ACI (M:1.17) $>$ AAI (M:0.89). This is in line with findings on infant- and foreigner-directed speech [23], [24]. In foreigner-directed speech subjects tend to lengthen pauses, while in child-directed speech segments are lengthened. The

TABLE I

STATISTICS CALCULATED ON THE ACOUSTIC PACKAGING RESULTS OF ADULT-ROBOT, ADULT-CHILD AND ADULT-ADULT INTERACTION

		ARI	ACI	AAI	ACI-AAI		ACI-ARI		AAI-ARI	
		M (SD)	M (SD)	M (SD)	Z	p	Z	p	Z	p
1	Total number of APs	6.25 (3.15)	4.08 (2.17)	2.19 (0.85)	3.4	0.00	-2.2	0.02	-4.8	0.00
2	Total length of APs [s]	14.89 (5.91)	10.45 (4.86)	6.30 (1.92)	3.4	0.00	-2.4	0.02	-4.8	0.00
3	Average length of APs [s]	2.62 (0.92)	2.90 (1.36)	3.26 (1.46)	-1.2	0.24	0.4	0.68	1.6	0.11
4	Total Number of MPs (in APs)	12.19 (4.92)	8.77 (4.32)	6.27 (1.93)	2.3	0.02	-2.3	0.02	-4.3	0.00
5	Total Length of MPs (in APs) [s]	13.44 (5.39)	9.52 (4.57)	5.50 (1.72)	3.6	0.00	-2.3	0.02	-5.0	0.00
6	Average length of MPs (in APs)	2.37 (0.86)	2.64 (1.24)	2.84 (1.28)	-0.9	0.38	0.7	0.51	1.5	0.14
7	Total number of utterances	7.00 (3.74)	4.69 (2.71)	2.77 (1.21)	2.7	0.01	-2.1	0.03	-4.2	0.00
8	Total length of utterances [s]	10.17 (4.30)	7.34 (3.57)	5.48 (1.77)	1.9	0.05	-2.0	0.04	-3.8	0.00
9	Average utterance length [s]	1.67 (0.85)	1.95 (1.41)	2.48 (1.48)	-1.8	0.08	0.4	0.68	1.9	0.06
10	Total Number of pauses in speech	6.00 (3.74)	3.69 (2.71)	1.77 (1.21)	2.7	0.01	-2.1	0.03	-4.2	0.00
11	Total Length of pauses in speech [s]	6.57 (3.25)	4.44 (3.09)	1.71 (1.30)	3.3	0.00	-2.1	0.04	-4.8	0.00
12	Average length of pauses in speech [s]	1.23 (0.39)	1.17 (0.75)	0.89 (0.68)	1.9	0.06	-0.8	0.45	-2.5	0.01
13	Average number of MPs per AP	2.19 (0.93)	2.47 (1.17)	3.14 (1.14)	-2.6	0.01	1.0	0.31	2.9	0.00
14	Ratio of interaction length to speech length	1.79 (0.49)	2.34 (2.31)	1.38 (0.55)	3.4	0.00	0.4	0.68	-3.2	0.00
15	Ratio of AP length to speech length (in APs)	1.56 (0.32)	1.49 (0.25)	1.28 (0.29)	3.4	0.00	-0.3	0.74	-3.5	0.00
16	Ratio of AP count to speech length (in APs) 1/[s]	0.66 (0.31)	0.59 (0.23)	0.45 (0.28)	2.4	0.02	-0.5	0.59	-2.4	0.01
17	Ratio of all MPs to MPs assigned to APs	1.16 (0.19)	1.60 (1.37)	1.21 (0.41)	1.6	0.11	0.9	0.39	-0.7	0.48
18	Ratio of interaction length to AP length	1.17 (0.17)	1.66 (1.78)	1.19 (0.55)	1.4	0.18	-0.2	0.88	-1.3	0.19

ARI = Adult-Robot Interaction, ACI = Adult-Child Interaction, AAI = Adult-Adult Interaction, AP = Acoustic Package, MP = Motion Peak
Z, p are the results of an asymptotic Wilcoxon Mann-Whitney rank sum test between the interaction type specific results.

high standard deviation also indicates that the subjects are uncertain about their unfamiliar communication partner's capabilities, resulting in a high variance of communication strategies.

Acoustic Packages: There is a significant difference in the total number of Acoustic Packages per test subject (cp. row 1): ARI (M:6.25) $>$ ACI (M:4.08) $>$ AAI (M:2.19). Thus, the interaction in ARI is in general longer than in ACI. This is also shown by the number and length of motion peaks and utterances (cp. row 4, 5). However there is no significant difference in the average length of Acoustic Packages (cp. row 3). Therefore a "unit" in the interactions seems to be temporally the same regardless of the kind interaction type.

Structural properties of Acoustic Packages: Looking at the average number of motion peaks per Acoustic Package (cp. row 13), the results show a significant difference between ACI-AAI and ARI-AAI. The difference between ACI and ARI is not significant: ARI (M:2.19) $\lesssim$ ACI (M:2.47) $<$ AAI (M:3.14). The average number of motion peaks per Acoustic Package can be interpreted as a measurement for the amount of structuring in the interaction: Few motion peaks per packages indicate high structuring, since only a small part of the task is demonstrated within a package. Less structuring is indicated by a higher number of motion peaks per package. The result indicates more structuring for ACI and less structuring for AAI, which is expected. The results reveal that structuring in ARI is on a similar level as in ACI. The ratios in row 15 and 16 can be interpreted similarly. Note that looking at the average utterance length alone does not provide significant results (cf. row 9).

IV. DISCUSSION

We used our Acoustic Packaging system to segment and analyze statistical properties of adult-adult, adult-child and adult-robot interaction in tutoring scenarios. Acoustic Packaging has been observed as a means of communication that is used towards infants [7]. In previous work, we showed that our Acoustic Packaging model is able to reflect the structural differences between tutoring in adult-adult and adult-child interactions. In this paper, we additionally analyzed adult-robot interaction and in comparison with adult-adult and adult-infant interaction. According to the Acoustic Package analysis, the multimodal structure of events is similar between ARI and ACI. In both types of interaction, less action is packaged within an utterance compared to AAI. In ARI and ACI, the subjects seem to package a similar amount of action. This might be an indication for similar units of tutoring in these situations. Yet, an important difference between ARI and ACI is the higher verbosity in ARI.

The data in our experiment is automatically processed by the Acoustic Packaging system. Therefore, it is likely that this approach has a higher error rate than manual annotation. For example, the speech recognizer might not always correctly segment speech, which contains parents' whispering towards their children. Thus, it is important to emphasize our goal to develop strategies that enable robots to react to and learn from tutoring situations. In this context, it is important to be able to segment the users' interaction with the system and to determine the presence of tutoring behavior. The measurements compared here are possible indications how to design features which can be used to discriminate between tutoring and non-tutoring situations.

We conclude that applying Acoustic Packaging Model in our analysis, subjects exhibited a similar tutoring behavior to children as to robots. However, the question how to reliably detect differences between ARI and ACI remains open for further research and includes the problem of robots being less familiar to their communications partners than children are. It remains a challenge to decrease the unfamiliarity by providing proper feedback.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from the FP7 European Project ITALK (ICT-214668).

REFERENCES

[1] C. Rovee-Collier and H. Hayne, "Reactivation of infant memory: implications for cognitive development." *Advances in child development and behavior*, vol. 20, pp. 185–238, 1987.

[2] A. L. Woodward and J. A. Sommerville, "Twelve-month-old infants interpret action in context." *Psychological science : a journal of the American Psychological Society / APS*, vol. 11, no. 1, pp. 73–77, January 2000.

[3] L. E. Bahrick, R. Lickliter, and R. Flom, "Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy," *Current Directions in Psychological Science*, pp. 99–102, June 2004.

[4] L. J. Gogate and L. E. Bahrick, "Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations," *Infancy*, vol. 2, no. 2, pp. 219–231, 2001.

[5] L. J. Gogate, L. E. Bahrick, and J. D. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child Development*, vol. 71, no. 4, pp. 878–894, 2000.

[6] K. Hirsh-Pasek and R. M. Golinkoff, *The Origins of Grammar: Evidence from Early Language Comprehension*, The MIT Press, 1996.

[7] R. J. Brand and S. Tapscott, "Acoustic packaging of action sequences by infants," *Infancy*, vol. 11, no. 3, pp. 321–332, 2007.

[8] Y. Nagai and K. J. Rohlfing, "Can motionese tell infants and robots. what to imitate?" *4th International Symposium on Imitation in Animals and Artifacts*, pp. 299–306, 2007.

[9] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.

[10] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2008, pp. 3545–3550.

[11] J. S. Herberg, M. M. Saylor, P. Ratanaswasd, D. T. Levin, and M. D. Wilkes, "Audience-contingent variation in action demonstrations for humans and computers," *Cognitive Science*, vol. 32, no. 6, pp. 1003–1020, September 2008.

[12] M. Knoll and L. Scharrer, "Acoustic and affective comparisons of natural and imaginary infant-, foreigner- and adult-directed speech," in *INTERSPEECH-2007*, 2007, pp. 1414–1417.

[13] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PLoS ONE*, vol. 3, no. 7, 2008.

[14] L. Schillingmann, B. Wrede, and K. Rohlfing, "Towards a computational model of acoustic packaging," vol. 8, IEEE. Shanghai, China: IEEE, 2009.

[15] R. J. Brand, D. A. Baldwin, and L. A. Ashburn, "Evidence for 'motionese': modifications in mothers' infant-directed action," *Developmental Science*, vol. 5, no. 1, pp. 72–83, March 2002.

[16] B. Wrede, J. Fritsch, and K. Rohlfing, "How can prosody help to learn actions?" Poster presented at the 4th International Conference on Development and Learning (ICDL), 2005.

[17] 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," vol. 8, IEEE. Shanghai, China: IEEE, 2009.

[18] J. Fritsch and S. Wrede, "An integration framework for developing interactive robots," 2007, pp. 291–305.

[19] G. A. Fink, "Developing hmm-based recognizers with esmeralda," in *TSD '99: Proceedings of the Second International Workshop on Text, Speech and Dialogue*. London, UK: Springer-Verlag, 1999, pp. 229–234.

[20] F. Lömker, S. Wrede, M. Hanheide, and J. Fritsch, "Building modular vision systems with a graphical plugin environment," in *Computer Vision Systems, 2006 ICVS '06. IEEE International Conference on*, 2006, p. 2.

[21] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*. Washington, DC, USA: IEEE Computer Society, 1997.

[22] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, November 1983.

[23] S. Biersack, V. Kempe, and L. Knapton, "Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech," in *Interspeech-2005*, 2005, pp. 2401–2404.

[24] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak e-ng-l-i-sh? a comparison of foreigner- and infant-directed speech," *Speech Commun.*, vol. 49, no. 1, pp. 2–7, 2007.