A Computational Model of Acoustic Packaging

Lars Schillingmann, Britta Wrede, and Katharina J. Rohlfing

Abstract—In order to learn and interact with humans, robots need to understand actions and make use of language in social interactions. The use of language for the learning of actions has been emphasized by Hirsh-Pasek and Golinkoff (MIT Press, 1996), introducing the idea of acoustic packaging . Accordingly, it has been suggested that acoustic information, typically in the form of narration, overlaps with action sequences and provides infants with a bottom-up guide to attend to relevant parts and to find structure within them. In this article, we present a computational model of the multimodal interplay of action and language in tutoring situations. For our purpose, we understand events as temporal intervals, which have to be segmented in both, the visual and the acoustic modality. Our acoustic packaging algorithm merges the segments from both modalities based on temporal overlap. First evaluation results show that acoustic packaging can provide a meaningful segmentation of action demonstration within tutoring behavior. We discuss our findings with regard to a meaningful action segmentation. Based on our future vision of acoustic packaging we point out a roadmap describing the further development of acoustic packaging and interactive scenarios it is employed in.

Index Terms—Acoustic packaging, action representation, action segmentation, robot action learning, tutoring situation.

I. INTRODUCTION

ACOUSTIC packaging has been proposed in developmental research by Hirsh-Pasek and Golinkoff [1] as a possibility of bottom-up action segmentation. This form of bootstrapping is suggested to guide children towards the hierarchically organized action structure known in adults [2]. Moreover, the concept of acoustic packaging complements the research on action segmentation in two important aspects. First, it points out a developmental perspective on action understanding suggesting a way of how meaningful units that are crucial for action perception and action memory in adults can be learned. Second, it adds acoustic signals to the features that have been proposed as mostly visual bottom-up processing for meaningful action parsing in adults [3].

L. Schillingmann and B. Wrede are with the Research Institute for Cognition and Robotics, Bielfeld University, and the Applied Informatics Group, Faculty of Technology, Bielefeld University, Bielefeld, 33615, Germany (e-mail: lschilli@techfak.uni-bielefeld.de; bwrede@techfak.uni-bielefeld.de).

K. J. Rohlfing is with Emergentist Semantics Group, Center of Excellence Cognitive Interaction Technology, Bielefeld University, and the Research Institute for Cognition and Robotics, Bielefeld University, Bielefeld, 33615, Germany (e-mail: rohlfing@techfak.uni-bielefeld.de).

A. Developmental Studies

The main idea of acoustic packaging is that while adults are considered to parse actions in meaningful parts and subparts [3], children have to discover these units. For the developmental period between birth and 9 months of age, Hirsh-Pasek and Golinkoff propose that speech provided by adults has the power of yielding meaningful chunks out of the action stream [1]. This power can be characterized by two achievements.

On the one hand, speech can be associated with some elements of actions [5], [6]. This association can be understood in terms of the Intermodal Redundancy Hypothesis [7], stating that information picked up by different senses is redundant and can be better perceived. Currently, developmental research emphasizes the role of sensory overlap for cognitive, social, and emotional development and Zukow-Goldring states that this way, infants attention is educated [6]. For example, Gogate and Bahrick [8] showed in experimental research that when 7-month-old infants were presented a syllable with a synchronous movement of the labeled objects, infants could remember this syllable more easily and link it to the presented objects than their peers receiving and asynchronous presentation.

On the other hand, the power that speech has also manifests itself in helping children to extract some parts and subparts of actions. The timely coincidence of speech and action, thus, yields meaningful chunks. This extraction has been shown by Brand and Tapscott [9] in a study, in which 7.5 to 11.5-month-old infants were familiarized with video sequences showing short action clips. The acoustic input coincided with portions of the action stream and thus, "packaged" paired clips together. During the test, infants viewed packaged and nonpackaged pairs of actions framed in silence. The results of the study revealed that 9.5-month-olds looked longer at the nonpackaged action sequences, suggesting that acoustic input (i.e., narrations heard during familiarization) influenced the way of how infants perceived the action units.

Since action understanding remains a challenge for the robotics [10], in this paper we suggest that bottom-up mechanisms such as acoustic packaging is relevant for younger children and can pave the way for action learning in social interactions.

B. Motion Segmentation

From a technical point of view, action segmentation can be narrowed down to the problem of segmentation of video sequences. Previous work associated with this area considers different ranges of motion segmentation like detecting scene cuts in movies or segmenting group actions in meeting recordings. In the following, we will group the relevant approaches according to their segmentation goal and look at properties such as online

Reference	Segmentation Goal	Multi-modal	Online	Predefined Classes
[12] Janvier et. al.	Scene cuts	no	no	no
[13] Wolf et al.	Key Frames	no	yes	no
[14] Davis et al.	Aerobic Actions	no	yes	yes
[15] Schuldt et al.	Human Actions	no	?	yes
[16] Rui et al.	Actions: Key poses	no	?	no
[17] Zhang et al.	Group Action	yes	no	no

processing or the capability of handling multimodal input (see Table I).

1) Scene Cut Detection: The problem of finding key frames in video sequences is often regarded with the goal to summarize or index the video. The idea is to extract a sequence of stationary images from the video in which each image represents the salient content of a certain video segment. These images are called key frames. Some of the work is focusing on detecting structure in the video, which results from the video editing such as scene cuts [11], [12]. Other work is focusing on selecting key frames within shots marked by scene boundaries [13]. The key frames are selected at the local minima of a motion feature based on optical flow. To put it in other words, in this approach, discontinuities are detected in the feature stream while some approaches are capable of online processing [13], others are designed for offline processing [12]. The commonality is that all approaches use the visual modality only.

2) Action Segmentation: In many approaches, developments on action segmentation are motivated by recognizing predefined classes as for example, in [14] and [15]. However if the goal is to create a system inspired by developmental learning, the categories and the structure of the action cannot be *a priori* assumed. Following the idea of analyzing video sequences without using pretrained classes, a more complex approach than scene cut detection, but with a similar basis is presented in [16]. This approach specifically aims at segmenting human actions into key poses. A key pose is understood as the boundary of a video segment, which captures important human action changes. The key poses are detected by searching temporal discontinuities in features based on optical flow that are supposed to carry information about the movements of the human in the image. The authors discuss potential applications such as summarizing video sequences, action recognition and segmentation, and selecting key frames in video compression tasks.

3) Segmentation of Meeting Recordings: Multimodality has been considered frequently in the analysis of meeting recordings. This string of research typically focuses on the segmentation of coarse grained categories, which occur in meetings and performs offline processing. In [17], these categories consist of group actions such as one participant speaking continuously or most participants being engaged in conversations. The authors use several high level visual features such as head and hand positions, as well as audio features such as speech activity and pitch. They report evaluation of different hidden Markov model (HMM) based approaches for automatic clustering of group actions is reported. However, although multimodal cues are taken into account in this approach, no explicit use of the synchrony

between the modalities is made. Rather, the relationship between the modalities is modeled statistically through the temporal structure provided by the HMMs.

4) Summary: As outlined above, both approaches, scene cut detection and action segmentation, have the detection of discontinuities in features derived from the video sequence in common. But as can be seen in Table I, most of the work focuses on one modality exclusively and is rarely online capable. This is especially the case with increasing complexity of the method. Furthermore, most approaches use points in time as the only representation of their segmentation results. Thus, there is no explicit representation of the segments found, which can further be interpreted.

From this summary, the main implications for the acoustic packaging model are: a) to handle multimodal input; b) be capable of online processing; and c) to not require pretrained classes for detecting segments. Additionally, acoustic packages need to be represented in a way, which allows to bootstrap a semantic representation.

C. Programming by Demonstration

Standard approaches within the programming by demonstration paradigm (or imitation learning) tend to be based on one single modality derived from hand and object tracking [18]. Such movements are generally tracked either visually or by a sensor glove. Pardowitz *et al.* use visual cues related to the hand and object movements in order to derive a gestalt-based action segmentation [19]. In other approaches, different kinds of inherent movement structure and implicitly coded world knowledge are used allowing for a meaningful action segmentation [20], [21]. Making use of information derived from parent-infant interactions is a relatively new approach and only few computational systems exist that explicitly make use of characteristics in tutoring behavior [4]. In [22], it has been argued that visual saliency cues may help to detect structural information in parent-infant interaction. However, although tutoring behavior has been reported to affect many modalities, especially gesture and speech [5], [23], to our knowledge no model has been implemented that makes use of several modalities and their synchrony for detecting action structure in demonstrated actions (but see below for first steps in this direction [24]).

In contrast to the above presented approaches, in the domain of object learning, there exist robotic systems that associate modalities. In [25], a system is presented, where both the visual and the acoustic cue are used for learning object names and sizes. In this system, the tutor positively or negatively rewards the learning agent depending on whether the extracted visual features, the extracted acoustic features, and the learned association between the visual and acoustic features result in a right response from the learning agent. The reward is used by the learning agent to tune both the association between modalities and the feature extraction within each modality.

D. Social Robotics

Within the field of social robotics Breazeal *et al.* [26] postulates that, in addition to specific visual cues, verbal descriptions and lexical cues are used that are helpful in deriving a

task hierarchy from a demonstrated action. However, in this approach, again implicit knowledge about the action is used for the process of action segmentation. In general, we can summarize that the inherent synchronous characteristics of multimodal tutoring behavior tend to remain unexploited. Yet, in computational models, it has been shown that the detection of synchrony in different modalities can be a potentially powerful low level approach to perform spatial and temporal segmentation [27]. More recently, it has been shown that differences between infant- and adult-directed interactions can even be found at a very low signal level [24] indicating that synchrony might be a useful cue for analyzing demonstrated actions in a tutoring scenario. However, a temporal model of how synchronous events that extend over time can be related is still lacking. This might be due to the fact that the temporal alignment of segments entails several very different problems. On the one hand, a segmentation in each modality has to be performed. On the other hand, the segments need to be temporally aligned which remains a challenge.

II. Modeling Acoustic Packaging

The development of robots, that are able to interact with humans and learn action from them, requires methods to segment actions into meaningful parts. We transfer the concept of acoustic packaging from developmental research and pursue the goal to create a software module that fulfills 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 further be processed by other modules that infer models about the actions currently presented. Hirsh-Pasek and Golinkoff describe a minimal and a maximal role that acoustic packaging can take [1]. In the minimal role, acoustic packages are formed on repetition of an acoustic chunk in conjunction with a particular event. In its maximal role, Acoustic Packaging can fuse separate events into meaningful macroevents. Our approach aims at 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 for both cues is required. We address the visual and acoustic segmentation problem in Sections II-C and II-D.

A second problem is the *temporal synchronization* of these sensory cues. The difficulty here is that hypotheses from audio and vision processing are typically not generated at the same time, nor in the same rate. Thus, temporal synchrony has to be exploited, which itself can be considered as an amodal cue, that 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

Fig. 1.   A test subject demonstrating how to stack cups to an infant [4].
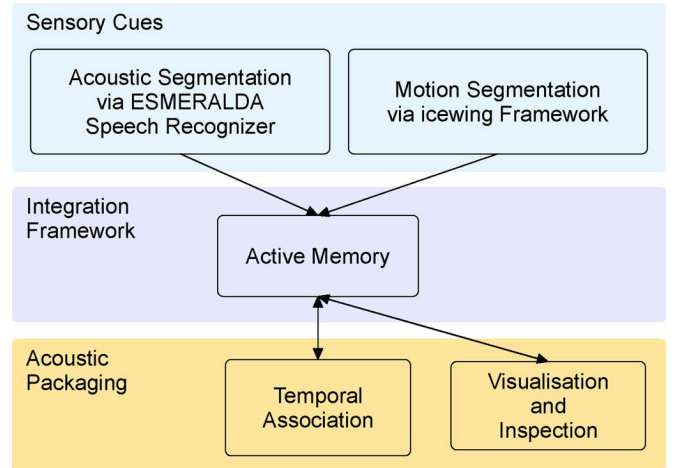

Fig. 2.   System overview with highlighted layers and their relation to the acoustic packaging system.

packages should be facilitated by the architecture. Since a socially interactive robot should give feedback during tutoring, the system has to be *online* usable and able to cope with updating hypotheses.

Finally, tools to debug and evaluate the acoustic packaging system are important. This sets up the requirements of *visualization*, which will provide support for the *inspection* in the development of the system.

B. System Overview

Our system for acoustic packaging proposed here consists of four modules (see Fig. 2). These modules communicate events through a central memory, the so-called Active Memory [28]. The Active Memory notifies components about event types they have subscribed to and is able to store these events persistently. It establishes thus, an integration framework that supports a decoupled design of the participating modules facilitating integration of further processing modules. This directly addresses the architectural requirement of extensibility.

Motion History Image

Amount of Motion per Frame

$$\sum_{ij} x_{ij}(t)$$

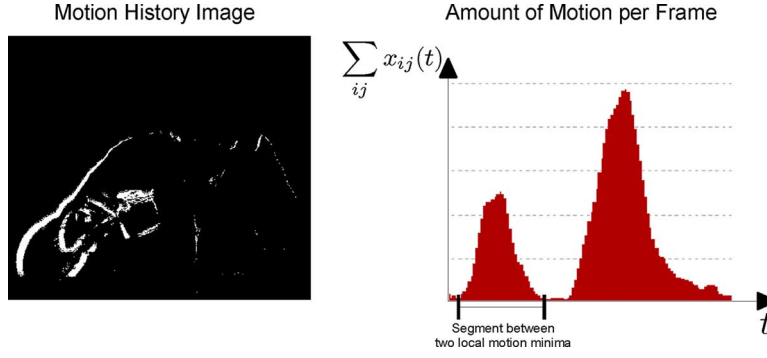Segment between
two local motion minima

t

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.

All signal processing modules are connected to the Active Memory. The audio signal is processed using the ESMER-ALDA speech recognizer [29], 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 [30].

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 nonspeech (pause) segments. Yet in a relatively noisy environment, such as the described experimental setting (see Section III), the separation of speech from nonspeech 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 nonspeech. As a consequence, a continuous chain of phoneme hypotheses generated by the speech recognizer is considered as a speech segment. Our speech recognizer inserts those phoneme hypotheses, as well as the corresponding audio signal into the Active Memory. As the recognition process is incremental during processing of an utterance the hypotheses are continuously updated.

D. Visual Action Segmentation

Unimodal action segmentation approaches attempt to find discontinuities in the visual signal. We follow this idea in its basic way and segment the visual signal at minimums in local motion. As a result, the visual signal is segmented into motion peaks, where each peak ranges between two local minimums in the amount of change in the visual signal. To understand this approach, the occurrence of motion peaks is related to action in the following example. If someone shows a cup, there is typically a motion minimum at the point where the cup is held 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.

Acoustic Package

Speech

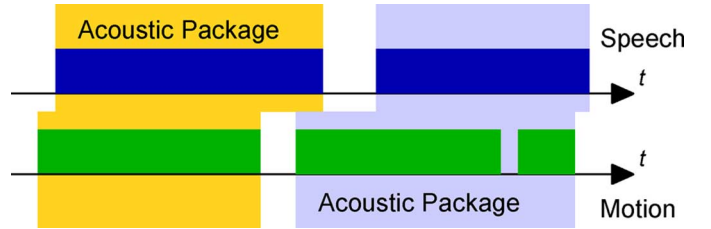t

Acoustic Package

Motion

t

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.

The segmentation into motion peaks is technically realized by an approach based on motion history images [14]. The amount of motion is calculated per frame by summing up the motion history image (see Fig. 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 are suppressed by using a sufficient window size that is small enough to not affect human movements. 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 (in a more focused approach) could be coped with by ignoring certain parts of the image. However, in this approach we chose to not use any prior knowledge with respect to space and content of visual information.

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 communicated through 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
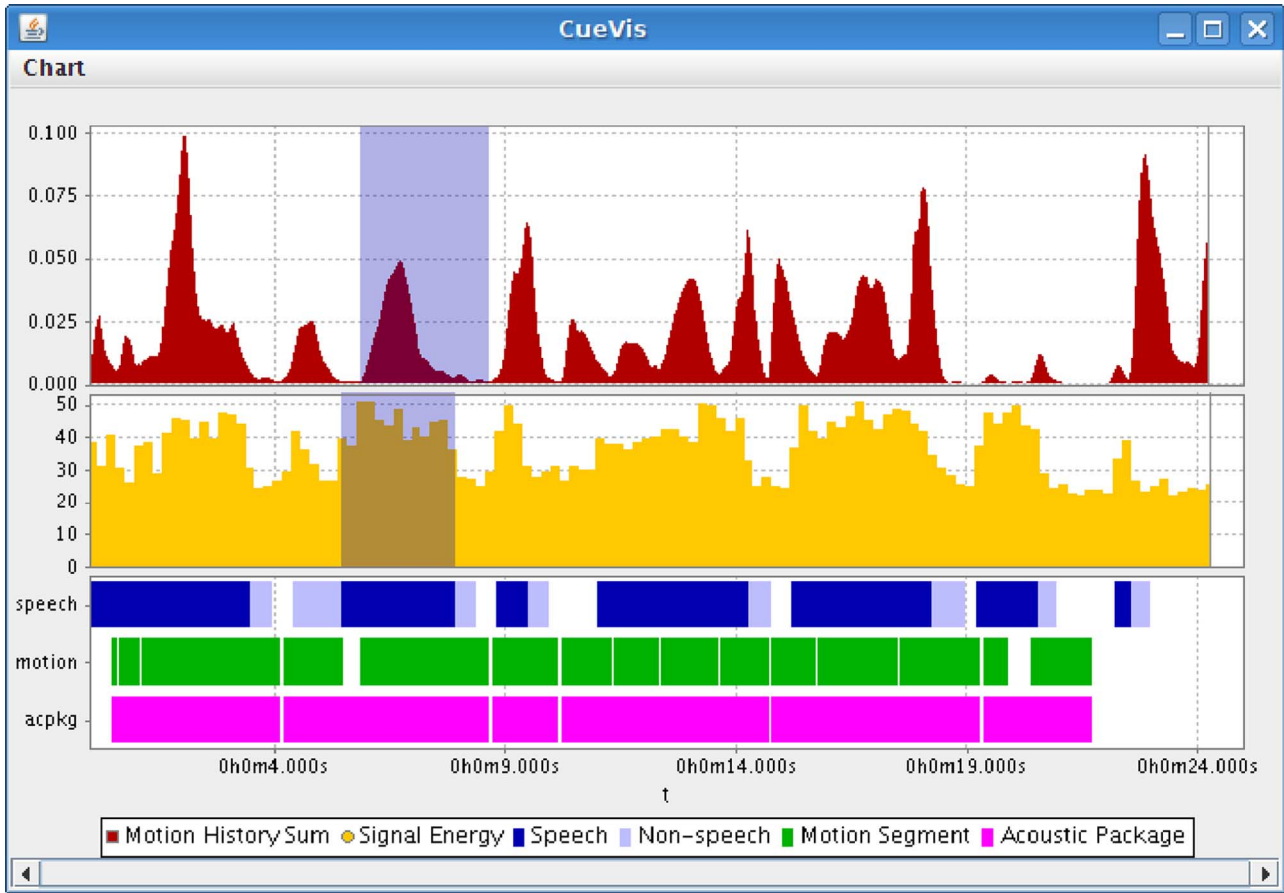
Fig. 5. Cue visualization tool.

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 [31] 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 Fig. 4 for the association process). When 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 requires handling potential processing delays on the one hand. 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 that 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. In the case, in which a motion peak is already associated with an acoustic package, the motion peak is reassigned. This depends on whether it has a larger overlap with the current speech interval. In the 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 step accords with the idea to realize an online usable system.

F. Visualization and Inspection

Since the temporal synchrony is an important cue for this system, tools are needed that analyze the acoustic packaging
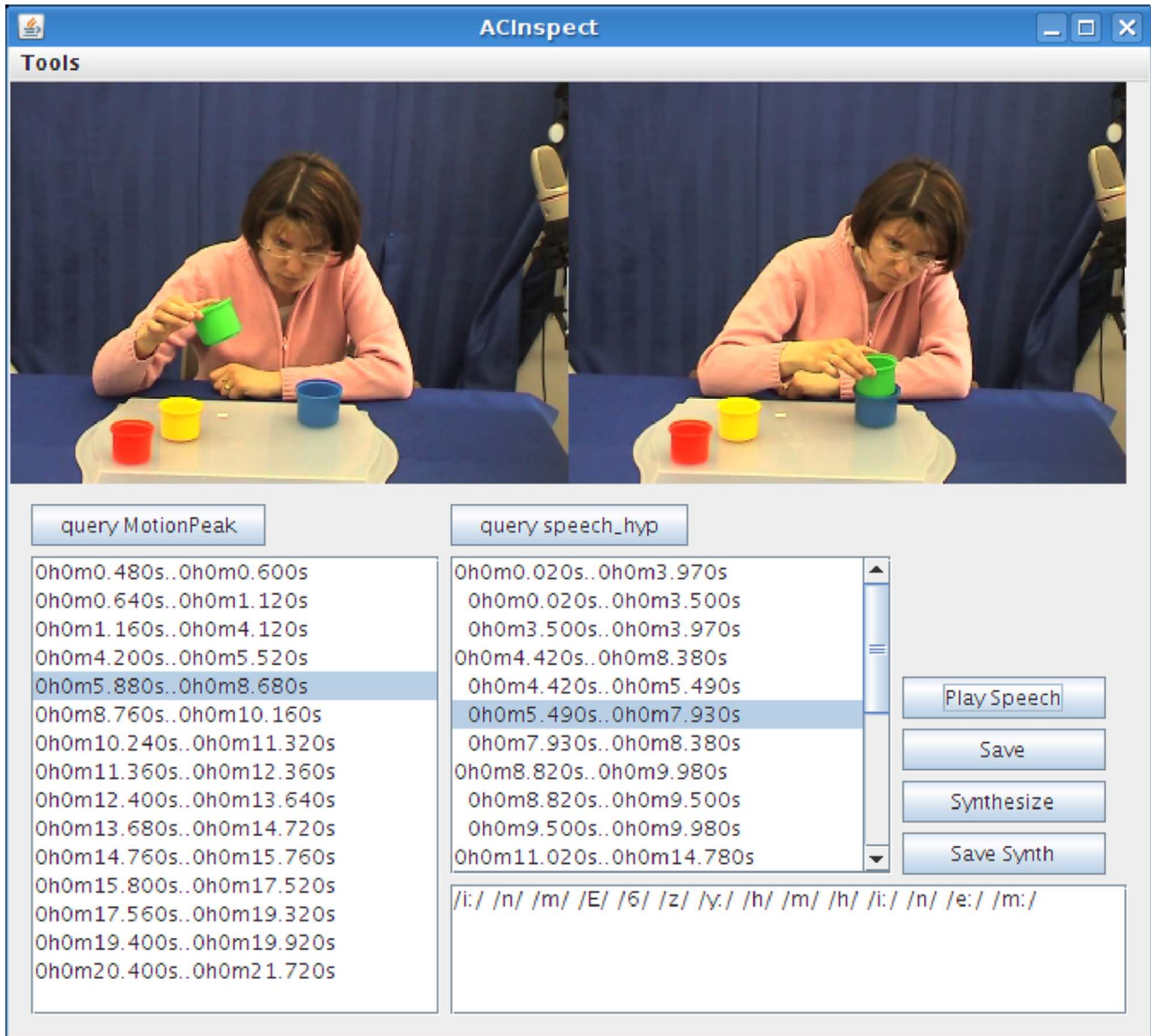
Fig. 6. Inspection tool.

process and the temporal relations of the involved sensory cues. Fig. 5 shows our visualization tool, monitoring events, which are communicated to the Active Memory by other processing modules. 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 nonspeech 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, Fig. 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 (Fig. 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 enabling 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, this functionality is a topic of further development. Taken together, these features of the inspection tool help to rate, optimize, and debug the acoustic packaging system and its parameters.

III. EVALUATION

It is important to emphasize that our approach delivers bottom-up hypotheses for acoustic packages and provides no high level classification on the semantic level of the processed sequences. In a sophisticated cognitive system, these obtained bottom-up hypotheses need to be further processed by learning modules. A robot that interacts frequently can verify and refine these hypotheses. Only then an evaluation of acoustic packaging within an interaction can be undertaken.

However, at this stage of development, an appropriate way to evaluate this system is to compare tutoring behavior in situations with children and adults. We therefore exposed our acoustic packaging system to a corpus containing video and audio data on adult- and infant-directed interactions [4]. From this corpus, we selected 11 subjects interacting with their 8 to 11-month-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, Fig. 7 illustrates the experimental setting). In the evaluation reported below, we focus on one task, namely the stacking cups (see Fig. 1).

Using this corpus, we analyzed the following hypotheses. Firstly, with reference to the research [4], [9] it can be hypothesized that parents structure their actions more when interacting with their children. Therefore, we expect the acoustic packaging system to generate more packages in an adult–child condition than in an adult–adult condition. For our purpose, we processed and compared 11 videos with adults demonstrating the stacking of cups towards children with 11 videos of the same adults demonstrating the same task to an adult (see Table II). A paired t-test revealed a significant difference in the amount of acoustic packages between these groups: $t = 3.618$, $df = 10$, and $p = 0.005$. This result strongly suggests that more acoustic packages can be found in an interaction towards a child.

Another expectation was that adult–adult interaction is less structured when compared to adult–child interaction. Since adults perform their actions and narrations more fluently when interacting with each other, we expect a larger amount of motion segments per package compared to the adult–child condition. We tested this hypothesis applying a paired t-test on the ratio of motion peaks to acoustic packages in both conditions. We found a significant difference: $t = 4.654$, $df = 10$, and $p = 0.001$. This result strongly suggests that more motion segments are packaged together in an interaction towards an adult. Table II shows the motion peak counts per subject.

What is also noticeable is that in adult–adult interaction is that the variance of motion peaks per acoustic package is higher than in adult–child interaction. This stems from the fact that our subjects displayed highly individual communication styles. For example, some subjects tended to be quite verbose in adult–adult interaction while demonstrating the action, which resulted in a large number of motion peaks per acoustic package. Other subjects behave in the opposite way. Thus, although on average, more motion peaks per utterance are packaged as compared to adult–child interaction, the difference is smaller. It is important
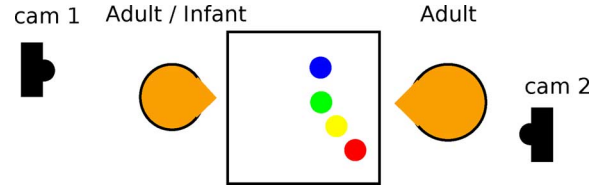

Fig. 7. Adult–Child/Adult–Adult interaction setting. The interaction partners are seated at a table facing each other. In this evaluation, recordings from camera 1 are used.

TABLE II
COUNTS OF ACOUSTIC PACKAGES (AP) AND MOTION PEAKS (M) ON SUBJECTS IN ADULT–ADULT INTERACTION COMPARED TO THE SAME ADULTS INTERACTING WITH CHILDREN

Subject	Adult-Adult-Interaction			Adult-Child-Interaction		
	AP	M	M/AP	AP	M	M/AP
1	3	7	2.33	17	33	1.94
2	3	8	2.67	7	14	2.00
3	3	13	4.33	17	30	1.76
4	3	9	3.00	3	5	1.67
5	10	24	2.40	34	60	1.76
6	1	4	4.00	3	7	2.33
7	2	7	3.50	8	10	1.25
8	2	7	3.50	13	29	2.23
9	2	6	3.00	6	13	2.17
10	3	16	5.33	7	14	2.00
11	5	10	2.00	8	14	1.75
M	3.36	10.09	3.28	11.18	20.82	1.90
SD	2.42	5.70	0.99	8.99	16.10	0.30

to note that in adult–child interaction, the variance is lower. This suggests that adult–child interaction is not affected by the subject's specific communication style to the same extent as it is in an adult–adult interaction.

A further question we investigated is whether it is possible to use acoustic packaging as an analysis tool for other types of interaction such as human–robot interaction. In [32], we compared adult–adult and adult–child interaction to adult–robot interaction. The simulated robot reacted to the environment using a saliency based attention model [33]. The results suggest a similar level of structuring in this adult–robot interaction compared to adult–child interaction although the verbosity towards the robot was higher than in adult-child interaction.

Currently the evaluation is limited to the statistical properties of acoustic packages calculated on adult–adult and adult–child interaction data. A detailed evaluation of the content of acoustic packages could reveal further relations between these groups. As a further step in this direction, commonalities and peculiarities between acoustic packages in the different conditions of the interaction could be identified.

IV. DISCUSSION

Our results show that when comparing the same subjects in two different conditions, significantly more acoustic packages were found in parent-infant interactions than in adult–adult interactions. In addition, the number of motion segments in the acoustic packages was significantly higher in adult–adult interactions than in parent-infant interactions. These results indicate that infant-directed interaction is more structured than

adult–adult interactions, which is in line with previous findings [4], [6], [9], [34].

Based on these results, we can assume that acoustic packaging provides a meaningful bottom-up action segmentation in tutoring situations. The segmentation consists of acoustic packages, which bind acoustic and visual events into a common unit. A sequence of acoustic packages can therefore be seen as a low level action representation of tutoring situations. This action representation contains information about the visual changes in the scene and the corresponding acoustic description. Furthermore, their temporal relationships are explicitly modeled. In the following, two examples are presented, which illustrate the contribution of acoustic packages to the segmentation of action. Based on these examples, we will discuss issues concerning the evaluation of segmentation correctness.

The first example stems from a tutoring situation. It consists of a mother taking a red cup, raising it, and finally turning it towards the child. While showing it to the child she says, "the red one". After a short pause, the mother continues to move the red cup over the yellow cup while saying, "in the yellow one," and drops it afterwards. In the second example, another mother takes the red cup and puts it directly into the yellow cup while saying, "the red one in the yellow one". When the first example is processed, two acoustic packages are formed: The first package consists of the acoustic segment "the red one" associated with taking and raising the cup. The second contains the utterance "in the yellow one" associated with moving and dropping the red cup. In contrast, the second example results in a single acoustic package containing the utterance "the red one into the yellow one". It is associated with a visual event which ranges from taking the cups to the cup in its final position. In both examples, the task is the same, but the way of communicating the task to the learner differs in the way the action is structured, which in is reflected in the segmentation provided by acoustic packaging. Although the packages differ, both segmentations are meaningful in the sense that the key frames and the acoustic segments associated with the acoustic packages contain the necessary information to describe the action.

As shown by the two examples, the fact that–given the same task–acoustic packaging can deliver different results in segmentation, can be an advantage for the learner on the one hand: It simply enables the learner to collect different segmentations for the same action. This way and over time, the learner will be able to form a representation on a more conceptual level. On the other hand, the variability in segmentation makes it more difficult to determine an objective ground truth for action segmentation on the level on which acoustic packaging operates.

Another reason why it is not desirable and applicable to perform a detailed evaluation of segmentation correctness is that acoustic packaging is a bottom-up process, which delivers segmentation hypotheses based on relatively simple cues. Thus, it is possible that motion observed by the robot is packaged, although it is not related to manipulation of the scene. A typical example is head movement such as nodding, which parents exhibit during communication with the infant. Here, the movement leads to quite large motion peaks, which are related to the communication with the child rather than to the action demonstration. It will be a future task to filter acoustic packages respec-

tively. We see a possible solution in a qualitatively different processing of, e.g., communication cues as packages that are unrelated to scene changes.

The method proposed here has been applied to interactions containing tutoring situation, in which the tutor performed manipulative actions. This specific situation limits the extent to which the benefit of acoustic packaging can be generalized. The motion that constitutes a manipulative action can be expected to provide a meaningful cue for segmenting the visual signal, and in its current realization, our implementation of acoustic packaging relies on this assumption: We segment motion by finding discontinuities in the visual signal as visual processing step. The discontinuities are detected by using motion history images to measure the amount of motion over time. The use of motion history images makes the approach "blind" to scenarios with no motion or to scenarios, in which motion plays a secondary role. Thus, certain actions such as holding an item still could lead to problems in this motion based segmentation approach: The visual segment containing the important conceptual aspect would not be captured, since the item is not moving. Scenarios, in which the motion cue is less important and other concepts play the primary role could, for example, consist of a situation with static objects where joint attention (thus, rather a social information) between the tutor and the learner provides a better cue to segment the interaction. In this case, acoustic packages would describe more than merely manipulative actions. This course of development is supported by the Emergentist Coalition Model [35], which makes a statement about the cues that children take into account when learning words: Initially, predominantly perceptual cues are processed. During the further development, social cues play an increasingly important role.

V. OUTLOOK

In our approach to acoustic packaging, events separate from vision and speech are fused into macroevents. This is accomplished by looking at the temporal relationship of events in both signals and combining synchronous events into acoustic packages. At the current stage, acoustic packaging can be used for two purposes. On the one hand, it can be used as a vehicle for feedback behavior in human–robot tutoring situations. The expected effect is here that the tutor gets insights into robot's processing, and the tutor gets an impression about the robot's stage of development. On the other hand, it can be used as an analysis tool for tutoring interactions. It provides an automatic measurement for the level of structuring in these interactions assuming that highly structured interactions are beneficial for action learning.

We envision further developments of such bottom-up segmentation methods for action learning. These developments can be pushed forward along two dimensions (see Fig. 8). One dimension spans across the improvements of acoustic packaging as a method and acoustic packages as a representation format (see Sections V-A–V-E).

The other dimension extends when acoustic packaging is put into human–robot interaction scenarios. Here, the general goal is that the robot will learn actions by interaction for which the necessary step is the development of feedback strategies [36], [37] in the robot and the initiation of interaction loops [38].
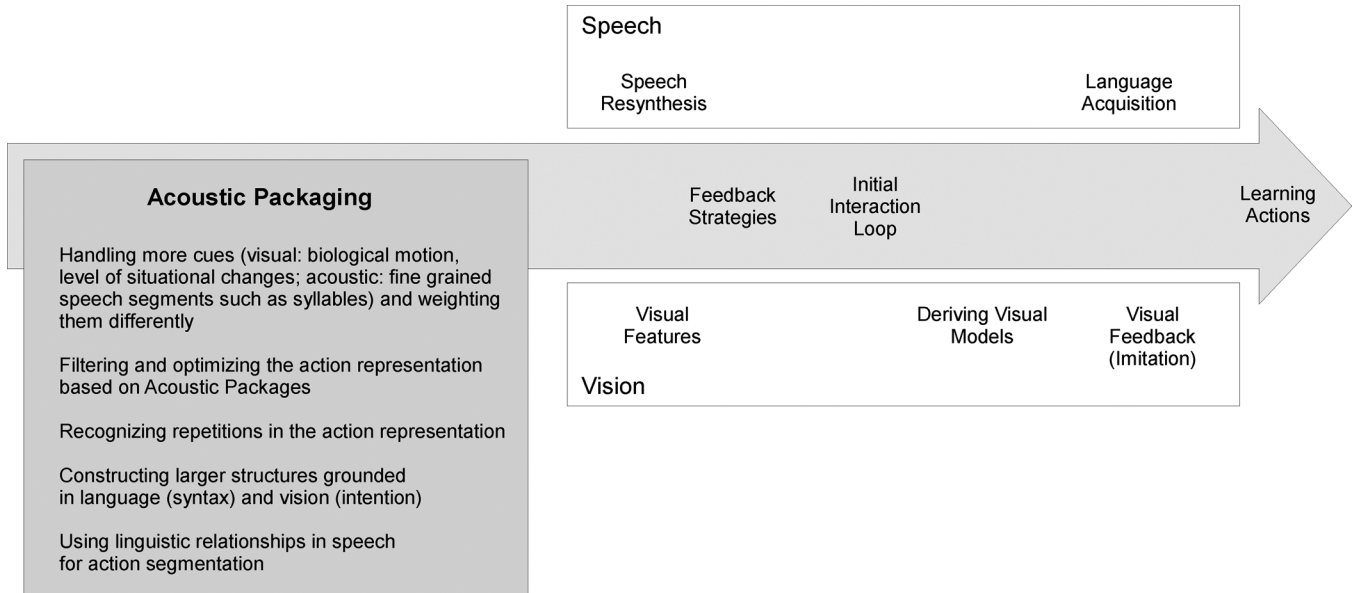
Fig. 8.   Roadmap showing future improvements of acoustic packaging in one dimension and next steps towards action learning in the other dimension.

A. Handling More Cues

Concerning the further development of acoustic packaging, we envision that more cues are helpful to bootstrap action representations which are grounded both visually and acoustically. Currently, it is not part of acoustic packaging to analyze how the environment is manipulated and what is manipulated. However, in addition to research about what and when to imitate [39]–[41], it might be important for the robot to distinguish between human motion and objects in the environment manipulated by humans. Especially the recognition of biological motion could help to further structure visual events. In infants, the sensitivity towards biological motion has been recognized as a fundamental experience. For example, predictive tracking as a basic cognitive capability emerges around 3 months of age, but when tested with faces, this capability can be observed significantly earlier [42].

The combination with another cue, which detects the level of situational change, could also help to structure human action according to the impact on the environment. For example, consider somebody lifting a cup and highlighting it in contrast to the action of lifting and stacking the cup into another one. In the former situation, the situational change is a minimal one, since probably only the position of the cup has changed. In the latter situation, the situational change is more significant, since the scene's appearance has changed: One cup disappeared in the other one.

As mentioned in the discussion, an inclusion of more social cues is possible as well. This will result in taking a different nature of cues into consideration as it is suggested in the Emergentist Coalition Model [35]. In this model, cues of different sources (perceptual, social, and linguistic) interplay with each other, but depending on the child's development, they are weighted differently. More specifically, in the first stage of development, predominantly perceptual cues are taken into consideration. Starting from the tenth month, children are increasingly paying attention to social cues as well. Once modules responsible for extracting these different cues are developed, they could be integrated in the acoustic packaging system. A weighting mechanism could further be adapted to model different developmental stages.

Considering acoustic cues (or in terms of the Emergentist Coalition Model: linguistic cues), the robot needs to detect which parts of an utterance are highlighted by the tutor when an action is presented. This would help the system to link segments of speech with actions. For this, a more fine grained speech segmentation than at the current state is required (see Section V-E). For the realization, syllables might be an appropriate level. Features, such as prominence [43]–noticeable by stress–could help to relate parts of speech particularly relevant to action structure.

B. Filtering and Optimizing the Action Representation Based on Acoustic Packages

Acoustic packages contain segments from different cues which are associated based on their temporal relationship. Concerning the features described in the previous subsection, the action representation based on acoustic packages needs to be further developed. This development should be motivated by memory processes, such as transforming a short-term action representation into a format that is appropriate for long-term storage. In this format, a higher conceptualization and stronger linkage to other concepts, as well as consolidation needs to be implemented.

Possibly consolidation and conceptualization can be achieved in a similar way as outlined for perceptual symbol systems [44]. When new acoustic packages are acquired, they are compared against other packages and the system relates them. Over time, a memory process can determine the invariant parts of actions and relate other parts as specializations to them. This way, filtering could be realized: If newly perceived packages are close to an abstract concept and cannot further contribute to it, they are not preserved anymore during memory consolidation.

C. Recognizing Repetitions in the Action Representation

The ability to bootstrap action concepts requires that similar parts in the action stream the robot perceives are clustered. As a method, the recognition of repeated chunks can be used, allowing to cluster these. This method should take both the visual and acoustic cues of the action representation into account. The resulting clusters will form recognition and synthesis units, on which speech recognition and synthesis can operate.

Methods for imitation learning could help in training units for visual action recognition and synthesis. The modules implementing the training methods do not necessarily need to run online during the interaction with the human. Instead they could run offline as part of a reorganization or consolidation process restructuring the data acquired during the human–robot interaction in the background.

D. Constructing Larger Structures Grounded in Language and Vision

Based on the clusters formed, as described in the previous subsection, larger sequences can be targeted. Clusters of grasping and lifting cups, as well as stacking and releasing them, are not sufficient to model a complete task. What is lacking is a larger construction encompassing the complete task and putting the several actions in a specific order.

Similar to larger constructions in action segmentation, according to usage based theories [45], speech production can be seen in constructions as well. Children build up their linguistic inventory by experiencing the language use of other speakers. At the beginning, children's utterances are simple. According to [45], their early utterances are concrete in their meaning as they are instantiations of item-based schemas or constructions. At a later stage, children integrate constructions of different abstraction levels from their linguistic inventory to form new utterances, that are chosen as appropriate for a current usage event.

Along this idea, acoustic packaging needs to combine the larger constructions into tasks that the robot can recall. Initially, these constructions can be used for a more complex imitation behavior of the robot. They have to be augmented in order to link the goals that this task implies with situations to which they can be applied. In the case that the task is communicated by using speech, like in instructing the robot verbally to do something, it is necessary to apply linguistic models. It is not sufficient for such models to make use of already trained acoustic descriptors. Instead, additional *syntactic relationships* between these descriptors must be regarded.

E. Using Linguistic Relationships in Speech for Action Segmentation

On the other hand, once linguistic constructions have been learned and can be recognized in the speech stream, these constructions may help in new demonstrations to segment actions. This means that by the use of a bottom-up strategy for speech and action segmentation, as provided by the acoustic packaging approach, top-down strategies can be built to segment action based on previously learned speech segments. For example, consider the case that the system has learned through repeated observation that the propositional phrase, "in den grünen" ("into the green one") coincides with the end of an action. This

information may then help the system to expect an action end the next time it hears this construction, even if the sensorial data is noisy and there is no clearly visible end of the action. This effect may even be enhanced by prosodic information such as intonation, for example, through correlation of falling intonation patterns with action ends.

F. Feedback Strategies

In a learning scenario, in which the robot interacts with the user and learns action, acoustic packaging might serve as a cue, on which basis a feedback behavior can be provided. The main challenge here is to investigate what form of feedback is effective during action learning in human robot interaction. Effective refers to the impression, that the tutor has and therefore, believes that the robot is actually learning about the ongoing task. At the current stage we think social cues might be considered in realization of such feedback behavior. For example, during tutoring, the robot could react by nodding, eye gaze, or some facial expressions. Even more elaborated verbal feedback such as repetition of words could help the tutor to interpret the systems' level of development. In an interaction with a tutor, this feedback behavior could signal that the robot knows the action or that the demonstrated action consists of new unknown movements. In accordance with this idea, Pitsch *et al.* [38] observed that when a child knows an action, his or her gaze is on the target (for example the target cup in the stacking cups task) instead of on each single demonstration movement.

In the case that the robot is actually capable of performing the demonstrated action, its manipulation can itself be seen as a form of feedback. Any kind of imitation is viewed as visual information about the internal representation of action [41] to the tutor.

These ideas of different feedback signals need to be modeled and tested in concrete human–robot interaction. Especially the integration of verbal behavior as a feedback form requires the integration of speech resynthesis.

G. Initial Interaction Loop

Analyses of human learners have shown that during tutoring, feedback is consequential for the characteristics of the presentation the tutor carries out [38]. For example, when children's attention is distracted, parents produce salient movements with the purpose of attracting children's attentions to the demonstrated objects and actions. In contrast, when children's attention follows the demonstration, less modified movements can be observed [38]. Thus, it seems that the modifications in movements, called motionese (as summarized in [4]), are a product of the interaction loop. The development of feedback forms can therefore only be the first step. We envision that the tutor's teaching behavior is guided by the learner's needs, monitored by feedback. This means that there is a constant loop between the tutor's and the learner's activities, in the sense that the teaching strategies that the tutor chooses are adjustments to the learner's exhibited capabilities.

VI. CONCLUSION

In this article, we presented a computational approach modeling acoustic packaging for human–robot interaction in a tu-

toring scenario. In process of acoustic packaging, speech binds visual events to acoustic packages. This binding is facilitated by the temporal overlap of events. We implemented this approach following a modular concept, being capable of online processing multimodal input. The resulting system fulfills the prerequisites necessary for being integrated in our robotic platforms. In an evaluation performed on natural data, we showed that acoustic packaging is able to reflect the structural differences between adult–adult and adult–child interaction. Based on these first experiences, we envisioned future developments of acoustic packaging as a method and acoustic packages as a representation format. Furthermore, we elaborated on how human–robot interaction scenarios can benefit from acoustic packaging and what are the next steps towards a system, which learns action in interaction with a tutor.

ACKNOWLEDGMENT

REFERENCES

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

[2] J. M. Zacks and B. Tversky, "Event structure in perception and conception," *Psychol. Bulletin*, vol. 127, pp. 3–21, 2001.

[3] J. M. Zacks and K. M. Swallow, "Event segmentation," *Current Direct. Psychol. Sci.*, vol. 16, no. 2, pp. 80–84, Apr. 2007.

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

[5] P. Zukow-Goldring, "Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words," *Early Develop. Parenting*, vol. 5, no. 4, pp. 195–211, 1996.

[6] P. Zukow-Goldring, "Assisted imitation: Affordances, effectivities, and the mirror system in early language development," in *Action to Language Via the Mirror Neuron System*, M. A. Arbib, Ed. Cambridge, MA: Cambridge Univ. Press, 2006, pp. 469–500.

[7] L. E. Bahrick, R. Lickliter, and R. Flom, "Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy," *Current Direct. Psychol. Sci.*, pp. 99–102, Jun. 2004.

[8] 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.

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

[10] J. K. Aggarwal, "Problems, ongoing research and future directions in motion research," *Mach. Vis. Appl.*, vol. 14, no. 4, pp. 199–201, Sep. 2003.

[11] U. Gargi, R. Kasturi, and S. Antani, "Performance characterization and comparison of video indexing algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Santa Barbara, CA, 1998, pp. 559–565.

[12] B. Janvier, E. Bruno, T. Pun, and S. Marchand-Maillet, "Information-theoretic temporal segmentation of video and applications: Multiscale keyframes selection and shot boundaries detection," *Multimedia Tools Appl.*, vol. 30, no. 3, pp. 273–288, Sep. 2006.

[13] W. Wolf, "Key frame selection by motion analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, Aug. 2002, vol. 2, pp. 1228–1231.

[14] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. Conf. Comput. Vis. Pattern Recog.*, Washington, DC, 1997, p. 928.

[15] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. 17th Int. Conf. Pattern Recog.*, Barcelona, Spain, Sep. 2004, vol. 3, pp. 32–36.

[16] Y. Rui and P. Anandan, "Segmenting visual actions based on spatio-temporal motion patterns," in *Proc. Conf. Comput. Vis. Pattern Recog.*, Atlantic City, NJ, Aug. 2000, vol. 1, pp. 1111–1118.

[17] D. Zhang, D. G. Perez, S. Bengio, I. McCowan, and G. Lathoud, "Multimodal group action clustering in meetings," in *Proc. ACM 2nd Int. Workshop Video Surveillance Sensor Netw.*, New York, NY, USA, 2004, pp. 54–62.

[18] S. Schaal, "Is imitation learning the route to humanoid robots?," *Trends Cogn. Sci.*, vol. 3, no. 6, pp. 233–242, 1999.

[19] M. Pardowitz, R. Haschke, J. J. Steil, and H. Ritter, "Gestalt-based action segmentation for robot task learning," in *Proc. 8th IEEE-RAS Int. Conf. Humanoid Robot.*, Daejeon, Korea, Dec. 2008, pp. 347–352.

[20] S. Ekvall and D. Kragic, "Integrating object and grasp recognition for dynamic scene interpretation," in *Proc. IEEE/RSJ Int. Conf. Adv. Robot.*, 2005, pp. 331–336.

[21] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction from perception-recognizing a grasp from observation," *IEEE Trans. Robot. Autom.*, vol. 9, pp. 432–443, Sep. 1993.

[22] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 1, pp. 44–54, May 2009.

[23] 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 Develop.*, vol. 71, no. 4, pp. 878–894, 2000.

[24] M. Rolf, M. Hanheide, and K. J. Rohlfing, "Attention via synchrony: Making use of multimodal cues in social learning," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 1, pp. 55–67, May 2009.

[25] Y. Zhang and J. Weng, "Conjunctive visual and auditory development via real-time dialogue," in *Proc. 3rd Int. Workshop Epigenetic Robot.*, Boston, MA, Aug. 2003, pp. 974–980.

[26] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration," in *Proc. 3rd Int. Joint Conf. Autonom. Agents Multiagent Syst.*, Washington, DC, USA, 2004, pp. 1030–1037.

[27] J. Hershey and J. Movellan, "Using audio-visual synchrony to locate sounds," *Adv. Neural Info. Process. Syst. 12*, vol. 12, pp. 813–819, 1999.

[28] J. Fritsch and S. Wrede, "An integration framework for developing interactive robots," *Software Eng. Experimental Robot.*, pp. 291–305, 2007.

[29] G. A. Fink, "Developing hmm-based recognizers with esmeralda," in *Proc. 2nd Int. Workshop Text, Speech, Dialogue*, London, U.K., 1999, pp. 229–234.

[30] F. Lömker, S. Wrede, M. Hanheide, and J. Fritsch, "Building modular vision systems with a graphical plugin environment," in *IEEE Int. Conf. Comput. Vis. Syst.*, New York, NY, 2006, p. 2.

[31] J. F. Allen, "Maintaining knowledge about temporal intervals," *Comm. ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983.

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

[33] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *IEEE Int. Conf. Robot. Autom.*, Pasadena, CA, 2008, pp. 3545–3550.

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

[35] G. J. Hollich, K. Hirsh-Pasek, R. M. Golinkoff, R. J. Brand, E. Brown, H. L. Chung, E. Hennon, and C. Rocroi, "Breaking the language barrier: an emergentist coalition model for the origins of word learning," *Monographs Soc. Res. Child Develop.*, vol. 65, no. 3, 2000.

[36] B. Wrede, S. Kopp, K. J. Rohlfing, M. Lohse, and C. Muhl, "Appropriate feedback in asymmetric interactions," *J. Pragmat.*, to be published.

[37] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

[38] 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, pp. 1–17.

[39] C. Breazeal and B. Scassellati, "Challenges in building robots that imitate people," in *Imitation in Animals and Artifacts*. Cambridge, MA: MIT Press, 2001, pp. 363–390.

[40] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Trends Cogn. Sci.*, vol. 6, no. 11, pp. 481–487, Nov. 2002.

[41] M. Carpenter and J. Call, "The question of 'what to imitate': Inferring goals and intentions from demonstrations," in *Imitation and Social Learning in Robots, Humans and Animals*, C. Nehaniv and K. Dautenhahn, Eds. Cambridge, MA: Cambridge Univ. Press, 2007, pp. 135–151.

[42] T. W. Boyer and B. I. Bertenthal , "Predictive tracking of social and non-social stimuli," presented at the Biennial Int. Conf. Inf. Studies, Vancouver, BC, Canada, 2008.

[43] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Proc. INTERSPEECH-2007*, Antwerp, Belgium, 2007, pp. 1809–1812.

[44] L. W. Barsalou, "Perceptual symbol systems," *Behav. Brain Sci.*, vol. 22, no. 4, pp. 577–660, Aug. 1999.

[45] M. Tomasello, "First steps toward a usage-based theory of language acquisition," *Cogn. Ling.*, vol. 11, no. 1–2, pp. 61–82, Feb. 2001.

[46] L. Schillingmann, B. Wrede, and K. Rohlfing, "Towards a computational model of acoustic packaging," in *Proc. Int. Conf. Develop. Learn.*, Jun. 2009.

Lars Schillingmann received the M.Sc. degree in computer science from the Bielefeld University, Bielefeld, Germany, in 2007. He wrote his diploma thesis about integrating visual context into speech recognition.

Subsequently, he joined the research group for Applied Informatics (Angewandte Informatik) at the Bielefeld University. He has been working for the BMBF (German Federal Ministry of Education and Research) Joint-Project DESIRE. Currently he is working in the EU-Project iTalk (Integration and Transfer of Action and Language Knowledge in Robots) on the topic of acoustic packaging. His research interests include learning and feedback processes embedded in human–robot interaction.

Britta Wrede received the M.Sc. degree in computational linguistics and the Ph.D. degree in computer science from Bielefeld University, Bielefeld, Germany, in 1999 and 2002, respectively.

She is head of the Hybrid Society Group at the Research Institute for Cognition and Robotics (CoR-Lab) at Bielefeld University, Germany. After her PhD, she received a DAAD Postdoc fellowship in the speech group of the International Computer Science Institute (ICSI) in Berkeley, CA. Her research interests concern the modeling and bootstrapping of human–robot interaction through analysing parent-infant interactions based on pattern recognition methods.

Katharina J. Rohlfing received the M.Sc. degree in linguistics, philosophy, and media studies from the University of Paderborn, Paderborn, Germany, in 1997. She received the Ph.D. degree in linguistics from the Bielefeld University, Bielefeld, Germany, in 2002.

Since May 2008, she has been head of the Emergentist Semantics Group within the Center of Excellence Cognitive Interaction Technology, at Bielefeld University, where she is interested in learning processes. Her postdoctoral work at the San Diego State University, the University of Chicago, and Northwestern University was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD) and by the German Research Foundation (DFG). In 2006, she became a Dilthey-Fellow (Funding initative "Focus on the Humanities") and her current project on the "Symbiosis of Language and Action" is supported by the Volkswagen Foundation. In her research, she is investigating the interface between cognitive development and early stages of language acquisition.