Dissertation

A Computational Model of Acoustic Packaging

Der Technischen Fakultät der Universität Bielefeld
zur Erlangung des Grades *Doktor-Ingenieur*

vorgelegt von

Lars Schillingmann

Juni 2012

Dipl.-Inform. Lars Schillingmann
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email: lschilli@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.). Der Technischen Fakultät an der Universität Bielefeld am 26. Juni 2012 vorgelegt. Verteidigt und genehmigt am 31. Oktober 2012.

Gutachter:
apl. Prof. Dr.-Ing. Britta Wrede, Universität Bielefeld
Prof. Dr. Giorgio Metta, Italian Institute of Technology, Genoa

Prüfungsausschuss:
apl. Prof. Dr. Jochen Steil, Universität Bielefeld
apl. Prof. Dr.-Ing. Britta Wrede, Universität Bielefeld
Prof. Dr. Giorgio Metta, Italian Institute of Technology, Genoa
Dr. Robert Haschke, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier nach ISO 9706

Acknowledgements

The following thesis would not have been possible without the support of many people. First, I want to thank my advisor Britta Wrede for the valuable discussions and helpful comments which made this work possible. Equally, many thanks go to Katharina Rohlfing for helpful discussions and for pointing out many viewpoints from developmental linguistics and psychology. Furthermore, I want to thank the reviewers Giorgio Metta and Britta Wrede as well as the additional examination board members Jochen Steil and Robert Haschke for taking the time to review this thesis.

During my work I have met many fascinating friends and colleagues in the Applied Informatics Group, the Emergentist Semantics Group, the Central Lab Facilities, and the Cognition and Robotics Lab. In particular I want to thank Ingo Lütkebohle and Sebastian Wrede for technical discussions, exchange, and insights on system integration and software engineering in general. I also want to thank Manja Lohse, Julia Peltason, Ingo Lütkebohle, Agnes Swadzba, and Katrin Lohan for project collaboration, various social activities, and interactions in many different contexts. Furthermore, I give thanks my office mate Iris Nomikou for supporting the office atmosphere with relaxed vibes and linguistic knowledge. Moreover, working in this environment would not be possible without Franz Kummert and Gerhard Sagerer, who gave me this opportunity.

Moreover, I want to thank my former student workers Christian Munier, Oliver Metz, and Fabian Klinke for their contributions. Furthermore, I want to thank Manja Lohse, Iris Nomikou, Angela Grimminger, Maha Salem, and Friederike von Lehmden for their feedback on readability and spelling. Likewise, thanks go to Silke Fischer for pointers to color perception in children.

Last but not least I want to thank all my friends, especially Michael Stachowski, and my parents for supporting me in many ways.

Contents

1. Motivation

In robotics, the problem of action learning is often viewed from a machine learning perspective. Machine learning primarily focuses on the generalization of action data by identifying invariant parts and adapting the result to new goals. These systems typically perceive action information visually in form of trajectories or by recording their own joint data. The existing methods in this domain provide partial solutions to action learning. Typically, it is predefined how the action is structured, which information is relevant, and how the information is transferred from the human to the robot in the interaction.

For example, when a human starts showing the robot an action, the system views this demonstration as one unit which ends once the human gives a command or when a specified goal has been reached. The structure of such action is therefore predefined. Furthermore, relevant knowledge is often preprogrammed as e.g., a set of objects that can be manipulated. In a scenario, usually one part of the interaction contains information which the robot should learn, for example, the shape of a movement. Transferring information to the robot by communicating multiple actions requires the human to follow specific patterns like, for example, providing a name for each action before showing it to the robot. Additionally, linguistic information is mostly integrated on a symbolic level in terms of labels, which makes it difficult for these systems to handle unknown linguistic events as they likely occur in a realistic scenario.

If robots should be able to assist humans in everyday situations in the future, we need to overcome these limitations. Action and language learning requires more flexible methods, since it is not possible to predetermine all possible tasks a robot would be involved in. Future systems need to be able to acquire this knowledge through communication with humans. No special training and means of communication should be necessary for humans if the interface is sufficiently flexible. But how can this problem be approached?

Interaction between adults and children provides a source of insights: children are able to acquire knowledge about new actions although they have limited experience with the events they observe. More specifically, they seem to be able to identify which parts of an action are relevant and adapt this newly-won knowledge to new situations. Typically this does not happen in an isolated way but in an interaction with an adult. In these interactions, multiple modalities are used concurrently and redundantly. Research on child development has shown that the temporal relations of events in the acoustic and visual modality have a significant impact on how this information is processed. Particularly, temporally overlapping events seem to have a stronger effect on action and language

learning than non-overlapping events (Gogate and Bahrick, 1998, 2001; Bahrick et al., 2004). Thus, developing a model of action and language synchrony would be beneficial for action and language learning in robotics. It would bring forward building robots that are able to learn action through interaction: in substance, it allows for identifying meaningful chunks in interaction. Furthermore, it helps to discover structural properties through analysis of interaction. Therefore, the main question this thesis will cover is: "how can we take advantage of speech and action synchrony?". The answer, which will be elaborated in this thesis, is a computational model of *acoustic packaging*.

The idea of acoustic packaging has been proposed by Hirsh-Pasek and Golinkoff (1996). They suggest 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. Modeling acoustic packaging requires transforming this idea into an architecture which automatically processes acoustic and visual sensory signals and chunks this stream into acoustic packages. Both the foundations from developmental research and the methods which have been proposed in the computer science community need to be considered when developing a computational model for acoustic packaging. Thus, aspects of infant development, the perception of events, automatic action and speech segmentation, as well as modeling temporal information need to be taken into account.

An important part of the related work for the computational model of acoustic packaging developed in this thesis stems from psychological and linguistic research which is therefore reviewed first. Two relevant main areas have been identified: The first area concerning event and action segmentation is reviewed in Chapter 2. The second area, which is reviewed in Chapter 3, comprises multimodal processing and the literature on acoustic packaging. Subsequently, the computational model of acoustic packaging will be introduced in Chapter 4. In this context, a more detailed description of the tutoring scenario is given. Furthermore, insights from the domain of speech processing, vision processing, and robotics are reviewed to substantiate findings from the previous chapters for a practical model and implementation of acoustic packaging for the given scenario. Chapter 5 focuses on evaluating the resulting model and performing analyses of action demonstrations. Acoustic packaging is used to analyze the interaction between pairs of adults, between an adult and a child as well as between an adult and a robot. In Chapter 6, further development steps of the acoustic packaging system are described which allow the system to extract more specific information from the actions perceived. As a result, the system is able to provide feedback to human users based on the content of the action, which was tested using the iCub robot (Metta et al., 2010). Subsequently, additional analysis of the content of acoustic packages is provided. Based on a review of the current representational capabilities of acoustic packages, Chapter 7 provides a roadmap which describes the future development of acoustic packaging in the context of action and language learning. Finally, Chapter 8 concludes this work.

2. Event and Action Segmentation

How humans perceive ongoing behavior and how it is segmented into meaningful units has been part of psychological research for many years. This area provides relevant insights for the question how action is perceived. It is relevant for this work since identifying meaningful chunks in interaction requires perceptual processes that handle action segmentation. This requirement leads to the follow-up question which sensory features need to be considered by these processes and how action can be represented. Therefore, in this chapter findings regarding action representations, sensory features, and perceptual mechanisms are reviewed to find sensible approaches. Typically most publications in this field cover results and at least theoretical considerations on all three topics. Thus, in this chapter several publications will be reviewed multiple times considering each topic separately. This separation supports the conceptual design of the acoustic packaging model that will be discussed in Section 4.3. Initially insights on how humans perceive ongoing behavior were inferred from experimental results with adult participants but in more recent work, action segmentation is studied also in infants. To facilitate classification of these findings, a short overview over the most common methods will be given first.

2.1. Experimental Methods to Investigate Action Segmentation

The main difficulty in understanding how humans segment action is that corresponding perceptual processes cannot be observed directly. Thus, researchers typically use methods that allow for drawing conclusions on possible mechanisms of action perception. The following methods can be identified: The first method requires the participant to actively segment actions observed into chunks, for example, by pressing a button (e.g. Newtson, 1973; Zacks et al., 2001; Zacks and Swallow, 2007; Meyer et al., 2010) or by comparing them with a set of predefined units (Schack and Mechsner, 2006). The resulting segments can directly be analyzed, for example, by comparing their lengths and agreement to other participants' results.

In the second method, participants do not directly segment actions or corresponding video data but watch a slide show displaying frames of an action sequence. Since here participants decide on their own when to continue with the next slide, their dwell time

on each slide can be measured (e.g. Hard, 2006; Meyer et al., 2011a). Similarly to the first method, timing is compared between participants' or the participants' segmentation judgments on a subsequent video segmentation task to draw conclusions on human action perception.

The third method is used when infants participate in experiments. Since they are not able to actively segment videos the preferential looking paradigm is used in these experiments. A common layout of such experiments is that actions are presented on two screens or stages next to each other. By measuring the infants looking time for both screens it can be determined which action is more familiar or more novel to the infant (e.g. Baldwin et al., 2001; Saylor et al., 2007; Hespos et al., 2009). This allows to infer if the infant remembers, for example, a previously seen action sequence.

The forth method is recording participants' brain activity using neuroimaging methods such as fMRI while they are watching the stimuli (see Zacks and Swallow, 2007, for examples). One possibility is to compare the participants' brain activity with the results of a segmentation task they performed subsequently. Thus, conclusions can be drawn on features relevant for action segmentation based on the neuronal activity in certain brain areas.

Regarding action segmentation the first method provides more direct measurements compared to the other methods. Nevertheless, the first method might be undesired, since participants need to actively segment action, which might affect the experimental results. Furthermore, this method cannot be used with infant participants. To overcome these problems, methods two and three are typically used. Method four is not applicable to infants and additionally is resource intensive on the one hand, but allows for direct observation of brain activity during action segmentation on the other hand. However, the activity patterns still need to be interpreted. In general, the explanatory power of studies on action segmentation reviewed in the following sections is limited concerning the underlying features and mechanisms, due to the indirect experimental methods used.

2.2. Representation and Memory of Meaningful Event Units

In this section, evidence for two hypotheses related to the question of how humans internally represent actions will be reviewed. The first hypothesis is that humans segment actions into event units and that these units play an important role in memorizing actions. The second hypothesis is that humans segment events into hierarchically organized parts where smaller units share boundaries with larger units (Kurby and Zacks, 2008). Typically research focuses primarily on the second hypothesis, which includes the first hypothesis to a certain extent. However, it is likely that hierarchical segmentation requires more conceptual knowledge than "flat" event segmentation where an organizational dimension in addition to the temporal dimension does not exist or is developed. Therefore, evidence for event segmentation and event organization will be reviewed separately.

2.2.1. Humans Segment Action into Variably Sized Units

According to Newtson (1973) it was assumed in the past that humans observe ongoing behavior and infer fixed sized behavior units. Newtson questioned this view and conducted studies focusing on this problem by letting participants perform a video segmentation task. In one study two groups of human participants segmented a video into meaningful units of action by pressing a button if according to their judgment one unit ended and the other began. Both groups watched a video but one group watched a modified version of this video including an unexpected action. The results showed that participants generate more units after an unexpected action than participants who were not exposed to this situation. The studies' results provide strong evidence for rejecting the theory that behavioral units have a fixed size. Humans are capable of adapting their unit of perception according to demands of the current task but also non-explicitly according to situational constraints.

Different lengths of units have further been analyzed by Hard (2006). She conducted three experiments on action segmentation with adult participants. The participants were exposed to a slide show and decided themselves when to forward to the next slide. In the first study, slides were presented that displayed frames from a filmed activity in one second intervals. Subsequently, the participants segmented the corresponding video into fine, intermediate, and coarse units. Looking time was measured for slides that were marked as a unit boundary in the video segmentation task using a tolerance of one second for matching slide frames with unit boundary markings. The results confirmed the hypothesis that participants looked longer at slides with action breakpoints than at slides between breakpoints. A significant linear trend in looking time was found from coarse to fine units. Additionally, participants with longer looking time at coarse boundaries were able to recall more actions. These results suggest that smaller units are integrated at coarse unit boundaries resulting in higher processing time. Thus, it seems that the unit size correlates with humans' internal processing and linking of information into representation of action segments.

Meyer et al. (2011a) showed similar results for 3–4 year old children. Here, children were exposed to a slide show that displayed images of an adult interacting with toys. The images were extracted from a movie with a temporal distance of one second between frames. After an instruction and a training phase children viewed the main slide show and decided on their own when the next slide should be displayed by clicking a mouse. The dwell times of the children were measured and analyzed for three groups of slides. Namely, slides before or close to coarse unit boundaries as well as fine unit boundaries and slides within units. The linear trend in looking time from fine to coarse units was also shown here. Additionally, Meyer et al. were able to show that for a subgroup of children their memory for object and actions seems to be related to the dwell time.

Zacks and Swallow (2007) summarize further findings strengthening the link between action segmentation and memory. Their insight is that event segmentation supports memory and learning. Studies have shown that participants which segment a movie

similar to generally agreed boundaries better memorize visual contents of this movie than participants deviating from these boundaries. Further studies suggest that if participants tend to segment tasks into hierarchical units, better performance in these tasks is achieved. Indeed Schack and Mechsner (2006) have shown that expert tennis players exhibit a significant deeper hierarchical organization of functional units compared to non- or low-level players. Zacks and Swallow further conclude that event boundaries form anchors for longterm memory. A related conclusion suggests if people are supported in correctly segmenting events they will remember theses events better. Additionally, learning the corresponding task from these events is improved.

Also infants seem to be able to segment ongoing action and to memorize events. Baldwin et al. (2001) present results regarding infants' ability to parse ongoing behavior. Infants of 10 to 11 months age were exposed to videos displaying intentional actions. Infants' looking times were measured after a familiarization phase for two different conditions. In the first condition pauses were inserted into the video at the middle of an intentional action, whereas in the second condition the pauses were inserted at the completion of an intentional action. The infants showed significantly different looking times between these conditions: They looked longer at the videos in the first condition. Based on these findings the authors inferred that infants parsed the actors' behaviors at boundaries of intentions. Hence, the pauses while the actor realized his intentions raised the infants level of attention at these points in contrast to the pauses at the boundaries of intention.

In the experiment described above possible units were interrupted by pauses. However pauses are not necessarily a prerequisite. Hespos et al. (2009) were able to show that infants are also able to spot actions they have previously seen even if they are embedded in a larger chunk of continuous action. They conducted studies with 6 and 8 month-old infants. The infants were first habituated to short action presentations that showed motion of a ball between different target positions such as on and over an object as well as on and under an object. In the test run, either a short novel action sequence or a short familiar action was shown. Since infants looked significantly longer at the novel test sequences, it is concluded that infants were able to detect the target action which was embedded in the habituation sequence shown before.

Another experiment by Hespos et al. (2009) revealed that infants have limited capabilities remembering actions with different transitions along a trajectory with the same endpoints. They investigated the possibility that categories of actions are more salient to infants compared to actions from the same category but with different transitions. Here the habituation phase consisted of longer actions containing different transitions between events. In the test condition, novel and familiar transitions were tested. The results did not reveal a significant difference in looking time between novel and familiar transitions. Furthermore, the looking times were compared to the habituation phase and revealed significant differences for both the familiar and the novel transitions. However, other experiments they conducted in the course of the same study showed similar looking times

between habituation and familiar events. Thus, the authors inferred that infants encode transitions differently and both the novel and familiar transitions are perceived as novel by the infant.

Kurby and Zacks (2008) summarize findings regarding brain activity and event segmentation. It is very likely that changes in brain activity correlate with humans subjective experience of event boundary locations. The participants did not know about the event segmentation task before the brain activity was measured, which suggests that these effects are task independent.

In summary, all these findings support the idea that humans even in infant age are capable to segment actions into temporally variably sized units. Furthermore, these units seem to be tightly related to the internal representation the human brain uses to memorize events. Another advantage of segmentation is that discrete events are an economic way to represent action that additionally allows for recombination of segments to solve new problems (Zacks and Swallow, 2007). However, the experimental findings suggest that these representations are limited in infant age and are in the process of development.

2.2.2. Humans Organize Action Segments Hierarchically

The hypothesis that action segments created by humans are not only variably sized but in addition they typically follow a hierarchical structure has been formulated in different terms, such as hierarchical bias hypothesis (Zacks et al., 2001), hierarchical encoding hypothesis (Hard, 2006), or hierarchical event perception (Kurby and Zacks, 2008). The argument most commonly used to support this hypothesis is that boundaries of fine segments tend to fall together with boundaries of coarse units. Therefore, fine units can be seen as subsets of larger units forming a hierarchical structure of action segments.

A study reported by Newtson (1973), is an early example of this approach. In one experiment two groups of human participants segmented a video into meaningful units of action by pressing a button if according to their judgment one unit ended and the other began. The video showed actions of a human actor. One group was instructed to perform a fine grained segmentation while the other group was instructed to segment the video into gross units. The results are consistent with the hypothesis that fine units are subsets of larger units.

According to Zacks et al. (2001) this hierarchical bias can be found in narrative comprehension, memory, and perception. Experiments designed similarly to Newtson's were conducted where the participant's main task was to segment video data showing activity while watching. The results revealed an alignment of the segmentations into fine and coarse units which is interpreted as support for the hierarchical bias hypothesis. Furthermore, participants were requested to describe video material verbally at coarse and fine units. A hierarchical bias could be observed on the fine units level, namely that fine unit descriptions close to coarse unit boundaries showed significant statistical differences in

syntactic and semantic features compared to fine units not close to coarse units. When participants were asked to describe events from their memory the syntactic and semantic properties of the resulting descriptions were similar to those described on-line.

Hard (2006) interpret their experimental results similarly. Here the hypothesis is that humans hierarchically encode observed behavior. In their experiments, participants were asked to segment tasks displayed as a slide show (see Section 2.2.1 for details). A significant linear trend in looking time was found from fine to coarse units, which was interpreted as support for the hierarchical encoding hypothesis.

According to Kurby and Zacks (2008) first evidence of hierarchical processing can also be found in infants starting at about 12 months of age. They seem to be sensitive to the way actions are grouped to achieve higher level goals. Furthermore, infants seem to be able to distinguish goal appropriate actions from non-goal appropriate actions, even if they are physically similar. At 24 months of age, infants are capable of forming hierarchical goals, which also affects their memory organization. Recent findings by Meyer et al. (2011a) with 3–4 year old children showed a linear trend in dwell time from coarse to fine event units in a slide show task further supporting the hierarchical processing hypothesis.

In general, the hypothesis that humans tend to segment events into a hierarchical structure is well supported by studies with different methods. The capability to hierarchically perceive events seems to develop in infant age. Kurby and Zacks (2008) suggest this capability is important to integrate existing knowledge on activities with information that is currently perceived.

2.3. Features Used for Event and Action Segmentation

The insight that humans parse ongoing events into units that form a hierarchical structure raises the question which features are used to realize this segmentation process. Currently it is assumed that humans take both low level sensory features and top down knowledge into account during action segmentation. Top down knowledge can be categorized into conceptual features such as goals and intentions of the actors performing a certain activity and as schemata that form a conceptual frame of an activity. Schemata specify, for example, the order of smaller steps and the objects involved that are required to realize a certain task (Zacks et al., 2009). However, it is still unclear how humans internally represent the information that schemata refer to, how this information exactly interacts with human event segmentation processes, and how this knowledge is acquired (Kurby and Zacks, 2008; Zacks et al., 2009). Furthermore, it can be hypothesized that schemata of activities play a role later in development compared to simpler sensory or conceptual features. Thus, this section will primarily focus on findings regarding low level sensory and conceptual features as well as their interaction.

The interaction between conceptual features and sensory features makes it difficult to clearly associate experimental results to one feature category. Often it is possible that both feature categories contribute to certain experimental effects. Therefore, it is suggested that sensory features and conceptual features are integrated during event processing, as, for example, changes in the movement of a person are integrated with inferred knowledge on that person's goals (Zacks and Swallow, 2007). This hypothesis is further supported by neuroimaging studies that indicate a relation between goals and physical movement features. Motion features are frequently mentioned in the context of action segmentation but not further specified. For example, Zacks et al. (2001) summarize that unit boundaries correlate with peaks in biological motion. Thus, features corresponding to physical changes seem to be important. Especially motion features are also supported by neuroimaging studies. However, it is assumed that more features contribute to segmentation processes such as color and sound (Zacks and Swallow, 2007).

In the course of same study that has been summarized in Section 2.2.1, Hard (2006) performed an analysis that can be related to features relevant for action segmentation. In this analysis, a change index was calculated for subsequent pairs of slides that were previously classified into fine, intermediate, and coarse units using participants' video segmentation results. The change index was computed using differences of edge features for each pair of slides. Correlating this change index with different boundary categories revealed that breakpoints correspond to slides with a high change index. Additionally, the change index seems to be higher for coarse units than for finer units. These findings suggest that humans use physical cues such as the amount of visual change to identify hierarchical structure in action.

Meyer et al. (2010) correlated physical features from the speech home corpus (Roy et al., 2006) with human judgments, which were acquired similarly to the methods described by Hard (2006). According to their results movement features such as body and hand speeds correlate to event boundaries. Results by Zacks et al. (2009) support these findings. They showed significant correlations of head and hand movement, acceleration, as well as hand-hand and hand-head distances with fine unit boundaries. The correlations could be shown for videos of three different activities (assembling a video game, assembling building blocks, folding laundry) considering fine unit boundaries. For coarse unit boundaries the correlations were not significant for most of the movement and distance features except for one video which shows the laundry task. Since adults typically possess world knowledge, it is possible that for some videos primarily high level information was determining for segmentation at coarse unit boundaries instead of the features investigated.

Infants that only possess limited world knowledge provide additional reasons to assume that low level features, such as movement features, play a role in action segmentation. This idea has been considered by Baldwin et al. (2001) in the analysis of their study summarized in Section 2.2.1. They suggest that low level features contribute to infants' ability to detect intention boundaries. One argument is that the capability to detect structure using low level features would be a prerequisite for infants to develop intention understanding, since they initially do not have the necessary world knowledge.

Saylor et al. (2007) describe a study testing infants' ability to segment action at intention boundaries. Infants aged 9–11 months watched continuous human action shown in two windows on a stage. During the test condition different tones were played coinciding with intention boundaries of the action shown in one window. The main hypothesis of the study was that infants' look longer at the window where tones match the intention boundaries. The hypothesis was confirmed. This study cannot provide evidence that top down conceptual representations are used in infants actions segmentation or provide details on bottom-up motion cues possibly relevant for this process but it shows that the interaction between features from several modalities affect the infants' segmentation process.

In summary, both physical movement features and conceptual features such as goals seem to play a role in action segmentation. It is assumed that humans integrate these features when segmenting actions. Due to this tight integration it is difficult to associate experimental observations to solely physical features or conceptual features. However, since infants at about 6 months of age already seem to be capable of segmenting actions, it can be assumed that initially certain motion features help to establish event boundaries. Possibly visual changes are initially detected that become more object and body specific over time. By clustering typical situations and goals conceptual features could be formed that support the segmentation process additionally.

2.4. Perceptual Mechanisms in Event and Action Segmentation

Humans segment ongoing action into variably sized units and features suspected to be relevant in this process strongly correlate to motion. A body of evidence supporting these hypotheses has been reviewed in Sections 2.2 and 2.3. But what are the perceptual mechanisms that allow for action segmentation? How do they work? The answers to this are still subject to ongoing research. However, researchers mostly agree on a common high level hypothesis about this mechanism. According to the common view, the certainty of perceptual prediction decides where event boundaries are placed. This idea has been around for relatively long time, for example, Newtson (1973) suggests that the unit size depends on its utility of prediction for the perceiver. Conversely, Zacks et al. (2007) formulates that event perception depends on change, since a static world is easy to predict. This statement fits with the experimental evidence that movement features or other changing physical features correlate with action segment boundaries (see Section 2.3). Thus, a common event segmentation theory (Zacks et al., 2007; Kurby and Zacks, 2008) assumes that humans maintain a representation of the currently ongoing events that is reset on prediction errors and segment boundaries are perceived at these points.

Based on this theory Zacks et al. (2007) derives further properties of human event segmentation processes: Event segmentation is automatic and runs concurrently to other cognitive processes. This is consistent with neuroimaging studies (Zacks and Swallow, 2007) and dwell time measurements (Meyer et al., 2011a), where participants initially are

not instructed to perform a segmentation task. Furthermore, the segmentation process controls working memory updates. Dwell time measurements revealing that participants look longer at coarse unit boundaries than at fine units support this property, since they suggest that more information needs to be integrated at coarse unit boundaries (Hard, 2006). Another property is that event segmentation processes operate concurrently on different timescales. This is supported by experiments where people segment stimuli at different timescales (Zacks and Tversky, 2001). Moreover, humans integrate information from multiple senses when segmenting events. One example is an experiment reported by Saylor et al. (2007), where infants looked longer at actions when tones where played coinciding with intention boundaries compared to actions where the tones did not match the intention boundaries.

A more problematic topic is the integration of prior knowledge into action segmentation. According to Zacks et al. (2007) prior knowledge is incorporated in human event segmentation processes. However, the interaction between low level movement features and higher level conceptual knowledge does not directly seem to affect event boundary locations. For example Hard et al. (2006) conducted a study where participants segment video clips of schematic events. Manipulating the participants previous knowledge did not seem to affect perceived boundary locations but seemed to affect the granularity of segmentation. Based on these results they suggest that conceptual knowledge has an influence which movement features are considered relevant for certain goals and thus influences the segmentation granularity. Furthermore, it seems that physical features play a strong role for the placement of event boundaries while conceptual knowledge affects the interpretation of action. Hard et al. further suggest that conceptual knowledge is also build by parsing events, which they highlight as an explanation for the development of understanding goals and intentions. A recent study by Zacks et al. (2009) suggests certain effects of previous knowledge on event boundary placements. Here participants segmented both live-action movies and simplified animations of these movies. The simplified animations showed stronger correlation with movement features than the live-action movies. On the other hand, if the animation viewers where informed about the action they watch, they showed no significant difference to uninformed animation viewers. This suggests that high level conceptual features have less influence than other conceptual features the perceptual process operates on. Since in general fine segmentations were stronger correlated with movement features than coarse units Zacks et al. suggests that coarse segmentation utilizes conceptual information to a larger extent than fine segmentations. Results showing that familiar activities are segmented into coarser units compared to non-familiar activities further support this hypothesis (Kurby and Zacks, 2008).

In summary, a clear picture of the perceptual mechanisms responsible for action segmentation is not yet available. For example, details of how high level conceptual and low level movement features are integrated require more research. However, the related work did reveal certain properties of the segmentation processes. Humans seem to segment ongoing events into units based on feature changes that make prediction more difficult, which especially includes movement features. The underlying mechanism continuously parses

and integrates sensory information while considering different timescales. Furthermore, it controls memory updates and potentially integrates high level level information that is available for the current context.

2.5. Conclusion

The psychological work on event and action segmentation provides insights into how humans internally represent action, the relevant features, and properties of the perceptual mechanisms that operate on these features. Humans segment ongoing action into variably sized units that seem to follow a hierarchical structure. The body of evidence regarding this hypothesis is relatively large for two reasons: First, research on this topic has been carried out for a relatively long time. The second reason is that considering the experimental methods units size is either directly measured or has strong correlates such as dwell time. Thus, even if the main focus of a study deviates from this topic results on unit size are typically available.

The relevant features for action segmentation seem to stem from both physical as well as conceptual sources. What features humans really take into account cannot be observed directly, however, many experiments indicate correlations between movement features and event boundaries as well as goals and intentions. There is a tendency that a coarse segmentation level correlates stronger with goals and intentions compared to a finer segmentation granularity.

Regarding perceptual mechanisms, prediction and change detection are frequently hypothesized as central aspects. Additionally, the experimental evidence supports that sensory and contextual information is continuously parsed and integrated at different timescales. How exactly higher level and lower level information is integrated has to be further researched. Considering infant development, it is assumed that capabilities to segment low level features are available before conceptual knowledge structures are build. Currently the common experimental methods regarding event segmentation are reaching a limit of their explanatory power, due to their indirect nature.

In general, research results from the area of event and action segmentation provide a higher level view on how humans segment ongoing action (see Figure 2.1). However it has to be noted that most of the work reviewed yet, focuses on unimodal stimuli, specifically on how humans handle visual input with regard to action segmentation. Although acoustic stimuli play sometimes a role in experiments, as, for example, in Saylor et al. (2007), the main focus still remains on the visual modality. Some research has also been conducted on reading comprehension (see Kurby and Zacks, 2008, for references) but not in interaction with other concurrent modalities. Furthermore, these experiments require cognitively advanced participants and thus cannot be directly transfered to infants in order to study developmental effects. The critical point here is that the interaction between two modalities might provide additional cues that facilitate human event segmentation, especially if conceptual knowledge is not extensively available such

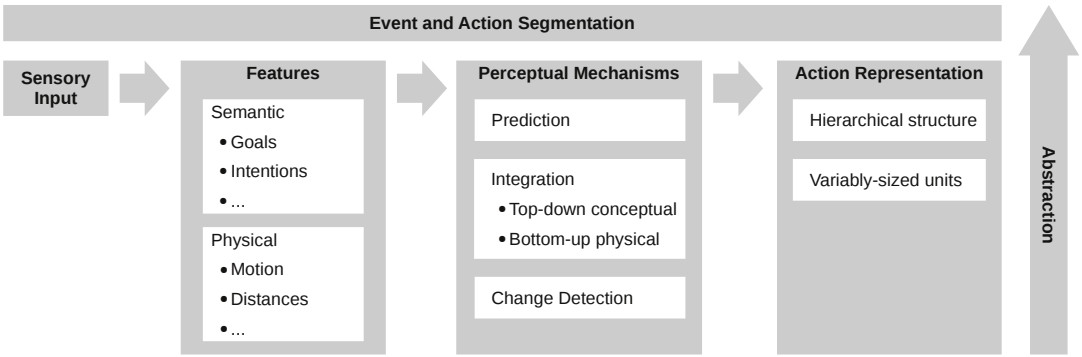


Figure 2.1.: Simplified schematic of the common view on event and action segmentation in humans including hypotheses from the publications reviewed.

as in infants. Furthermore, human event segmentation is primarily viewed as a model for perception without taking interaction with other humans into account as, for example, between a tutor and a learner. Especially during development this interaction could have an influence on action segmentation and the representations formed. The next chapter addresses some of these issues by reviewing insights on cross modal processing in humans.

3. Multimodal Processing and Acoustic Packaging

Infants perceive a continuous stream of multimodal sensory information and need to make sense of it although they have very little previous knowledge. Thus, it is a widely held view in developmental research that children use combinations of sensory cues to learn. These multimodal cues assist the learner, for example, in identifying relevant chunks of information in the sensory stream. While the work reviewed in Chapter 2 regarding event and action segmentation is typically focusing on one modality that is mainly the visual, in contrast here the primary focus lies on modality integration and its role during infant development.

This chapter will give a short introduction to modality integration during infant development. Closely related to modality integration is acoustic packaging — a concept specifically describing the interaction of language with events — that will be introduced. Acoustic packaging explains how infants are capable of segmenting the stream of multimodal information they perceive into meaningful chunks that form the first steps towards language learning.

3.1. Multimodal Processing and Integration

Human senses are specialized to perceive information from different modalities including acoustic and visual signals. The human brain does not process this information separately but is capable of fusing all these senses into one experience. The following sections will give a brief overview on the theories of how the sensory system develops and the effects on multimodal processing. For the present work the acoustic and visual modalities are primarily relevant. Thus, the integration of other senses such as touch is not further reviewed here.

3.1.1. Early and Late Integration

According to Robinson and Sloutsky (2010) two views on the initial state of the sensory system and the aspects that are subject to development are supported by empirical findings on child development. The *early integration* view assumes that sensory integration

is initially available. During their development, infants learn to separate modalities and how to identify specific details in the multimodal sensory stream. In contrast, the *late integration* view assumes that sensory integration is not initially available. During development infants learn to integrate different senses they perceive separately. For both views empirical evidence is available (Bahrick et al., 2002; Birch and Lefford, 1963). But each view has also difficulties in explaining certain effects which suggests that multiple factors have an impact on multimodal integration.

One class of effects that cannot be well explained by the late integration view are interference effects: Stimuli from one modality may hinder or facilitate stimuli from the other modality (Robinson and Sloutsky, 2010). Another class of effects cannot be explained by the early integration view alone: Infants seem to have problems to bind static visual stimuli with other modalities while they are able to form these links with dynamic visual stimuli (Robinson and Sloutsky, 2010). Thus, early integration may depend on special conditions. Robinson and Sloutsky propose two theories that model the effects of audio-visual input on children's attention that can provide an explanation for interference effects and binding problems. These theories will be summarized in the following sections.

3.1.2. The Intersensory Redundancy Hypothesis

Bahrick et al. (2004) propose the hypothesis that stimuli perceived redundantly between senses are preferred compared to stimuli present in only one modality. Redundant stimuli contain amodal information that is defined as not being specific to a single modality. Amodal information is redundantly conveyed across two or more modalities typically in a spatially and temporally coordinated way. This is only possible for modality unspecific information, namely tempo, rhythm, duration and intensity. A bouncing ball, for example, can exhibit a certain rhythm which is perceived both visually and acoustically. The Intersensory Redundancy Hypothesis (IRH) assumes that during early infancy redundant stimuli with amodal information lead to increased attention for the amodal part compared to the modality specific parts of the stimuli. The IRH predicts this interrelation as a mechanism that guides infants' attention to meaningful events. Therefore, multimodal stimulation facilitates processing and learning of amodal properties. Unimodal information on the other hand becomes less salient in presence of amodal stimuli which causes the interference effects described in the previous section. However, modality specific properties can still come into attention focus if no intersensory redundancy is available. Thus, a second prediction of the IRH is that unimodal stimuli facilitate processing and learning of modality-specific properties. Nevertheless, when infants' processing capabilities develop they learn to detect both modality specific and amodal properties regardless of an unimodal or a multimodal stimulus type.

The IRH is supported by a body of literature on experimental results both on animals and humans. For example, a study with infants revealed that a bimodal stimulus with intersensory redundancy can attenuate the infants capability to detect changes which

occur in a single modality of this stimulus (Bahrick et al., 2006). In a recent study (Flom and Bahrick, 2010), the effect of unimodal vs. multimodal stimulation on infants memory has been shown to be consistent with the prediction that modality specific properties are facilitated in unimodal stimuli. Furthermore, Bahrick et al. (2010) showed that although infants are able to detect amodal properties of a unimodal stimulus in one task, they require multimodal redundant stimulation if the task becomes more difficult. This suggests that the effects of IRH additionally depend on task difficulty.

3.1.3. Auditory Dominance

Robinson and Sloutsky (2010) describe a theory termed auditory dominance that provides an explanation for processing and binding problems when children are exposed to multimodal stimuli. According to this theory the corresponding visual information is less processed by the learner if auditory input is present (Robinson and Sloutsky, 2004). Therefore, auditory dominance is an asymmetrical effect that describes a priority difference for auditor information if auditory and visual information is perceived concurrently. However, the level of auditory dominance varies depending on the child's age, the length of the stimuli and the child's familiarity to the subject. With increasing age the auditory dominance effect decreases: Robinson and Sloutsky (2004) report that while younger children and infants show auditory dominance, children at four years of age switch between visual and auditory stimulus preferences depending on the content of the visual input.

Regarding input familiarity, Robinson and Sloutsky further analyze the relationship between stimulus length and the dominance effect. Short familiar input seems to interfere with unfamiliar input since processing time is short. Longer stimuli durations reduce this effect up to the point that children are able to handle both modalities. Based on these results the authors propose a mechanism describing the effects of modality dominance. The modality which gains attention first attenuates the other modalities but only for a short time. For longer durations the attention is released. Robinson and Sloutsky reason that allocating attention to modalities which disappear quicker is the more effective strategy to avoid missing transient auditory information and thus is the typical case. This explains why dynamic acoustic stimuli are dominant over static visual input.

The auditory dominance theory provides a sensible explanation for the complex processes in children that affect their attention and perceptual capabilities when processing multimodal stimuli. However, understanding the role of cross-modal processing in finding and learning meaningful units from the sensory stream requires a more overarching theory that will be reviewed in the next section.

3.2. Acoustic Packaging

The effects of multimodal stimuli on attention suggest that chunks of multimodal information play an important role in learning. The concept of acoustic packaging (Hirsh-Pasek and Golinkoff, 1996) includes the idea that acoustic information such as language provides infants with a bottom-up guide to attend to relevant parts in the sensory stream. This concept is consistent with the intersensory redundancy hypothesis and the findings on auditory dominance presented in the previous section.The theoretical foundations of acoustic packaging stem from research on language acquisition which allows to classify insights on multimodal perception along a developmental axis. Additionally, they provide first ideas how multimodal chunks in form of acoustic packages are stored, retrieved, and used for language learning.

In the following, two key publications that include and further develop the idea of acoustic packaging will be reviewed. Both focus on the idea that children learn language by using a coalition of sensory cues that include not only acoustic information. Although the second publication does not explicitly refer to acoustic packaging it continues to develop the initial work by putting a stronger weight on the role of social cues for language learning. The idea of acoustic packaging has been picked up by researchers that provided further evidence for this concept. These findings will be summarized subsequently.

3.2.1. A Coalition Model of Language Comprehension

Hirsh-Pasek and Golinkoff (1996) describe a theory of language learning, which especially focuses on the different cues that children take into account during their development. Their main message is that children use a coalition of cues when developing the ability to process language. Furthermore, they argue that language comprehension comes before language production, since comprehension plays a central role in building mental representations. This view is supported by findings that first comprehension of words seems to begin about four months before production of words starts (Hirsh-Pasek and Golinkoff, 1996, p. 172, par. 3). The authors describe three phases of development (Hirsh-Pasek and Golinkoff, 1996, pp. 163):

In the first development phase (0–9 months) the infant needs to make sense of the various inputs it perceives. Language production plays a secondary role in this phase, since children produce only few words if at all. However, during this first phase infants use acoustic packaging to internalize events. Acoustic packaging is a more primitive form of language comprehension where acoustic information is used to segment complex non-linguistic events. The authors define a minimal and a maximal role acoustic packaging can take (Hirsh-Pasek and Golinkoff, 1996, p. 168): 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. The result of this phase are acoustic and visual events which are linked by acoustic packages.

For this process to work infants already need to be capable of performing basic segmentation of their sensory input. Hirsh-Pasek and Golinkoff (1996, p. 166) describe three preconditions of which the first two address this issue: The first refers to the infants visual processing of world's events to image-schemas. In summary, they must be able to perceive basic spatial relations between events and recognize categories of movements as, for example, self-initiated movements. The second prerequisite states that infants must be able to extract acoustic correlates of linguistic units, such as phrasal and clausal units, from the speech stream. The third explains that the language the infant hears describes ongoing events and thus requires a temporal relationship between language and ongoing events. According to the authors this accompanying narration is at least valid for western societies.

In the second phase of development (9–24 months) infants refine the acoustic packages they have formed in the first phase. They perform a more fine grained analysis of acoustic packages and associate them with specific objects, events or actions. The result of this process is a linguistic mapping between acoustic units and linguistic units (i.e. phrasal and clausal units) as well as their meaning.

The third phase of development (24–36 months) focuses on syntactic aspects. In this phase the child is able to discover syntactic relationships within sentences and between them. From these the child is able to understand complex meanings. At the the end of this phase the child can rely on syntactic cues and understands more difficult linguistic constructions such as passives.

In summary, children use a coalition of cues to acquire the ability to comprehend language. During their development these cues are weighted differently. At the beginning, prosodic cues play an important role while syntactic cues get a stronger weight at the end. Regarding language production, findings are ambiguous, namely some report comprehension precedes production and others report production precedes comprehension (see Hirsh-Pasek and Golinkoff, 1996, p. 191, for references). The authors theorize that the order depends on how well these capabilities have been established at the respective point in development. Namely, comprehension and production develop in parallel if comprehension is resilient but comprehension precedes production if it is fragile. The latter is the case especially at the beginning of the child's development.

Hirsh-Pasek and Golinkoff strongly focus on linguistic aspects when defining the coalition model. Thus, they refer to prosodic and syntactic cues from the acoustic and linguistic domain but do not go into detail regarding visual cues. For example they do not provide details on how action is visually segmented but refer to image schemas. These focus on the relation of objects while actions are considered less. Since these cues are required as a prerequisite to acoustic packaging, a system realizing acoustic packaging already must possess these capabilities. The results of the packaging process are used by the second

development phase but the selection and further segmentation processes remain unclear. Additionally, cues such as social factors and their integration are not further specified. The latter is addressed by a further development of the coalition model summarized in the following.

3.2.2. The Emergentist Coalition Model

In Hollich et al. (2000b) a variant of the model in Hirsh-Pasek and Golinkoff (1996) is described. The central question is how children break the word barrier. Their hypothesis is that children's lexical development results from the interaction of multiple cues. This position results in three main points, which are described in the following.

The first point is that children are sensitive to multiple cues when learning words: Attentional cues, social cues, and linguistic cues. Concerning *attentional cues* the authors mention perceptual salience, temporal contiguity, and novelty. Children are able to follow the eye gaze of adults and they are sensitive to pointing, which are *social cues*. Furthermore, infants are able to attend to social information although the ability to actively detect this information emerges later. *Linguistic cues* are detailed as the ability of detecting language, segmenting speech and spotting words. Specifically prosodic information is used by infants to find words. In this context the authors highlight the role of exaggerated pitch and intonation in child-directed speech, which seems to modulate the infant's attention. Moreover, children also exploit grammatical information, for example, to identify what is labeled by a novel word.

The second point is that children weight cues differently in the course of their development. For word learning, certain cues such as saliency, for example, can play a more emphasized or less emphasized role based on the child's experience. The weighting of these input cues changes over time and is adapted to the learning situation. Furthermore, children can correlate cues to form categories.

The third point is that the principles children use for word learning are emergent and change from an immature to a mature state during development. This is reflected in more sophisticated heuristics children use for word learning. The authors hypothesize that immature children detect referents in a domain-general way by using perceptual saliency. Mature learners prefer social cues, which allow them to interpret the speakers focus of attention to find referents and learn new words. Thus, children move from domain general principles to domain specific principles (see Figure 3.1).

The main differences of the emergentist coalition model compared to the coalition model described in the previous section are the following: In the present version the role of social and attentional cues is more elaborated. Especially the role of eye gaze as word learning cue that allows children to assign words to referents is included here in contrast to the previous version. Furthermore, while the coalition model describes several phases of language development, the present model views the language acquisition process in a more continuous way in which the weighting of cues dynamically changes. Additionally,

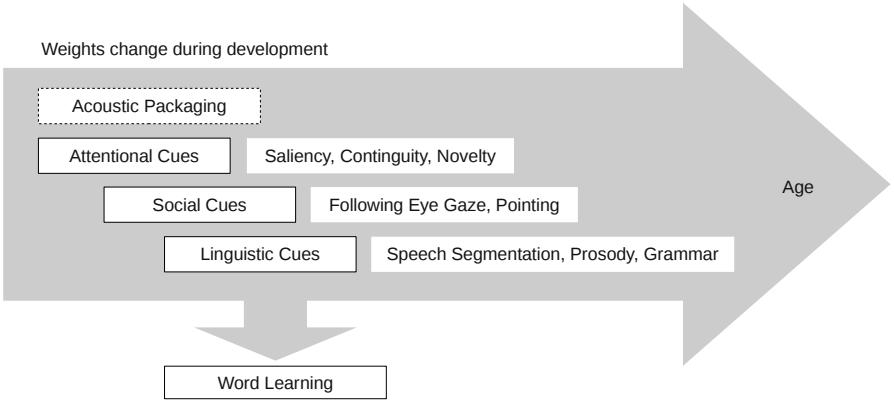


Figure 3.1.: According to the emergentist coalition model multiple cues contribute to word learning. The shifted depiction visualizes their changing weights during development.

specific mechanisms and representations in the language acquisition process such as acoustic packaging are not discussed in this variant. Nevertheless, these findings are relevant to the acoustic packaging concept since they suggest that social cues are also packaged providing, for example, information on the relevance of this multimodal chunk. Furthermore, acoustic packaging could itself be seen as a cue with changing weight during development whereas the highest weight lies at the beginning of development (see Figure 3.1).

3.2.3. Evidence for Acoustic Packaging

In Section 3.2.1 a minimal and a maximal role of acoustic packaging were defined. While in the minimal role, acoustic packages associate an acoustic chunk with a particular event, in the maximal role acoustic packages can fuse separate events into meaningful macroevents. In this section, evidence for both views will be summarized. Since the minimal role of acoustic packaging is noncontroversial (Hirsh-Pasek and Golinkoff, 1996, p. 168), findings in this direction will be summarized first. Regarding the maximal role of acoustic packaging, very little has been reported. However, the available findings strengthen the maximal role of acoustic packaging towards a concept how infants identify action structure by fusing smaller actions with accompanying speech. These results will be summarized subsequently.

The minimal role of acoustic packaging can be understood in terms of the intersensory redundancy hypothesis (see Section 3.1.2). The synchronous presentation of acoustic and visual activity can be seen as an amodal cue that heightens attention according to the IRH and is thus better perceived. Already seven month-old infants who were presented a syllable with a synchronous movement of the labeled objects could remember this syllable and link it to the presented objects more easily than their peers receiving an asynchronous

presentation (Gogate and Bahrick, 1998, 2001). Furthermore, object motion seems to be an important cue for young infants. Werker et al. (1998) showed that while 14 month-old infants were able to learn object word pairings with static visual stimuli, 8–12 month-old infants required moving stimuli. Even children at the age of 24 months are still sensitive to sensory redundancy of speech and motion when associating a label to one of multiple moving objects (Jesse and Johnson, 2008). The role of synchronous speech and motion is also supported by research on parent-infant interaction indicating that parents use synchronous speech and motion to teach their children certain words (Gogate et al., 2000; Zukow-Goldring, 1996). On a low level, Rolf et al. (2009) showed synchrony between motion and speech in adult-infant interaction. Furthermore, Meyer et al. (2011b) carried out a detailed analysis on infant direct speech and actions by discriminating action and non-action describing utterances. The level of synchrony was assessed by measuring the difference of onsets and offsets between manually annotated speech and action segments. Their analysis revealed higher synchrony and higher overlap between actions and action describing utterances compared to non-action describing utterances.

The maximal role of acoustic packaging extends the minimal role by stating that accompanying speech facilitates binding multiple visual events to larger meaningful chunks. In this case, acoustic packaging influences both the association of acoustic and visual events and the segmentation of ongoing action into acoustic packages. Little research has been conducted regarding this role of acoustic packaging.

Brand and Tapscott (2007) for example, pushed this topic forward and conducted a study which explicitly focused on the effects of acoustic packaging. They investigated whether infant directed speech influenced infants' segmentation of actions sequences. In this study two groups of infants aged 7.5–9.5 months and 9.5–11.5 months were exposed to video clips that were accompanied by speech in one condition (packaged) and silence in another condition (non-packaged). Each clip consisted of two smaller clips presenting an action demonstrated by a tutor. After a familiarization phase the infants were exposed to the same clips with reversed order of the smaller clips within. In this case the clips were played muted for both conditions. Infants' looking times were compared between the two conditions. The results show that the non-packaged clips were preferred by the older infant group in terms of longer looking time. Based on these results the authors infer that infants perceived the packaged clip as a unit. Thus, the infants considered the clips a single familiar event if they have been packaged before which explains the shorter looking times in this condition. The non-packaged pairs in contrast were perceived as individual units resulting in longer looking times. The authors conclude that these results confirm the influence of co-occurring acoustic input on action processing in infants. The results presented by Brand and Tapscott strengthen the acoustic packaging concept towards the role of an intermodal cue which facilitates structuring of ongoing action into larger units.

3.3. Conclusion

The potential of multimodal integration theories lies in their capability to explain how first steps in making sense of events in the world can be made with little or no previous knowledge by combining multiple sources of information. However, these theories are still subject to ongoing research. At the current point the findings reviewed in this chapter show that following a single paradigm such as early or late integration does not seem to be able to explain all effects observed in experiments. Therefore, it is suggested that multiple factors contribute to how audio-visual information is processed (see Section 3.1.1).

The *intermodal redundancy hypothesis* (IRH) assumes that amodal input that is perceived redundantly between senses is preferred compared to stimuli present in only one modality. Regarding infants, it is therefore hypothesized that amodal stimuli affect learning and memory, since they impact infants' attention. A body of literature elaborates on the effects of amodal input supporting the IRH. For example, depending on the infants age, unimodal information is suppressed when amodal stimuli are present (see Section 3.1.2).

While the IRH does not consider different weights or priorities of input modalities, the *auditory dominance* theory assumes that acoustic input differs from visual input in this regard. It is reasoned that acoustic stimuli are transient and thus by initially prioritizing acoustic input, infants avoid loosing information. The experimental findings on auditory dominance suggest that processing of these stimuli depends on the infants' state of development, their familiarity with the input as well as the length and content of the input (see Section 3.1.3).

The impact of concurrent acoustic and visual input has previously been recognized by researchers in a related context. In research on language learning it has been theorized that integration of multiple cues contributes to the ability of infants to learn language. The coalition model of language learning looks at three phases of infant development ranging from 0–9, 9–24, and 24–36 months of age. Especially in the first phase, before — according to this theory — any language mapping is available to the child, *acoustic packaging* binds visual and acoustic events together and forms acoustic packages based on the temporal co-occurrence of these events (see Section 3.2.1).

In its maximal extent infants use acoustic packaging to find structure in the perceptual stream by combining multiple visual events with accompanying narration. For this, speech is used by the caretaker to highlight structures. In the next developmental phase acoustic packages are subject to an analysis that assigns them to specific objects and actions resulting in first linguistic capabilities. The interaction of acoustic and visual stimuli in acoustic packaging is consistent with the IRH. Acoustic packages can be seen as an amodal cue that not only leads to increased attention on multimodal events but also supports binding of these events. Furthermore, the elevated role of speech can be related to auditory dominance that assigns an initial attentional priority to acoustic input. The relation of the different multimodal processing and integration theories to audio-visual events is depicted in Figure 3.2.

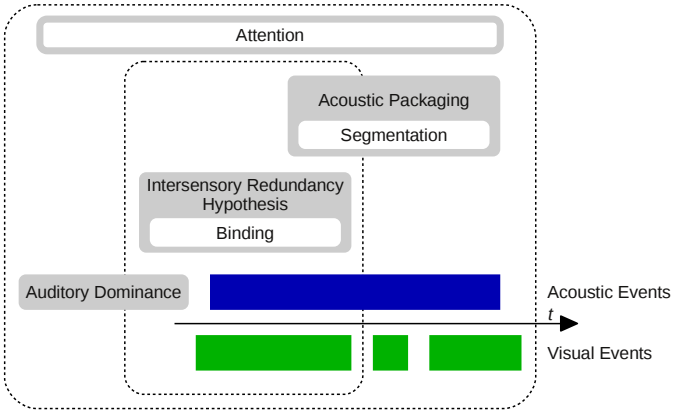


Figure 3.2.: Multimodal processing and integration theories that affect attention, segmentation, and learning in early infant development and their relation of to audio-visual input and different temporal horizons.

The initial definition of acoustic packaging considers only few cues that guide the learner in the process of learning first words. The emergentist coalition model extends this view by including social cues and attentional cues such as saliency in the word learning process. Thus, it is plausible acoustic packages also contain social and attentional cues, which change their weight during development (see Section 3.2.2).

Support for acoustic packaging can primarily be found in studies showing that at about seven months of age infants are able to associate speech with moving objects if these stimuli are presented concurrently. Static visual stimuli can be associated at a later stage of development suggesting that motion is an important attentional factor. Furthermore, synchronous speech and motion is utilized by adults when teaching certain words and in demonstrating actions to their children (see Section 3.2.3). The study by Brand and Tapscott (2007) also provides support for the maximal role of acoustic packaging, in which multiple visual action events are structured by accompanying speech.

The number of studies concerning the maximal role of acoustic packaging shows that this role has been little researched. Furthermore, the current knowledge on how the visual and acoustic modalities are processed before they are combined to acoustic packages is limited. There is also no information available on the association process that forms acoustic packages and how they are represented. Additionally, it remains unclear how exactly acoustic packages are rated and processed in the subsequent developmental steps. Therefore, in the next chapter a computational model of acoustic packaging is elaborated that integrates the currently available knowledge from linguistic and psychological research into an implementation of acoustic packaging.

4. A Computational Model of Acoustic Packaging

While the previous chapters focused on psychological and linguistic findings related to acoustic packaging this chapter will primarily focus on the technical realization of a model that implements acoustic packaging. This requires to substantiate the current findings on acoustic packaging where they lack sufficient detail an implementation requires. Thus, in the following acoustic packaging will be conceptualized. This step is especially necessary to clarify how the system should process and represent the input modalities and how acoustic packages are created and structured. Additional related work regarding the implementation will be reviewed subsequently. Finally, the resulting system architecture and evaluation is described.

4.1. Scenario and Task Overview

The development of robots, that are able to interact with humans and learn action from them, requires methods to segment actions into meaningful parts. The idea of acoustic packaging is transferred to this context to fulfill two important tasks in 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 could further be processed by other system components that infer models about the actions currently presented. Another use case is the analysis of human-human and human-robot interaction structure.

An example of a human-human tutoring situation is depicted in Figure 4.1. The child observes a multimodal action demonstration where the tutor shows how to stack cups and verbally comments his actions. According to the psychological findings on multimodal processing, it is assumed that children are able to segment and learn from these multimodal action demonstrations. Acoustic packaging follows the idea that redundant stimuli in tutoring situations provide cues on how action demonstrations can initially be segmented. Hirsh-Pasek and Golinkoff (1996) defined two roles acoustic packaging can take (see Chapter 3). 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. The approach described in this chapter

Figure 4.1.: Example of a tutoring situation where a participant demonstrates how to stack cups to an infant Rohlfing et al. (2006).

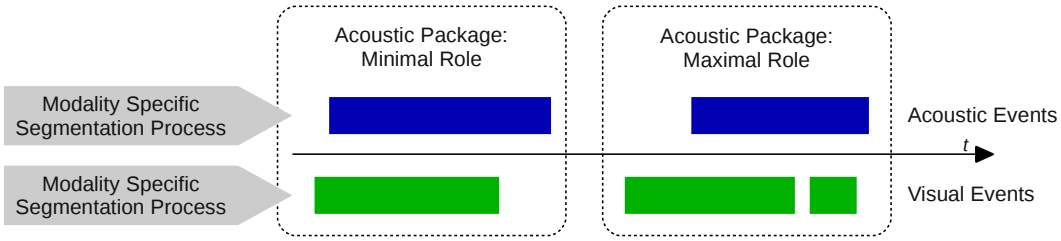


Figure 4.2.: Illustration of an audio-visual stream of events that is segmented into acoustic packages based on their temporal overlap.

aims at the maximal role of acoustic packaging. Thus, the action structure represented by acoustic packages is influenced by the tutors way of structuring the presentation to the child.

In pursuing a concept of acoustic packaging the terms event and action need to be further clarified. Here, both events and actions are used synonymously and defined as temporal intervals that span chunks of sensory input. This input typically consists of acoustic and visual information. Regarding semantics an event or action may contain a variably sized segment of ongoing visual activity or acoustic activity. The details of segmentation depend on sensory cues that provide information on how a modality is segmented. Thus, acoustic packaging requires a segmentation process for each type of event it associates.

Acoustic packaging merges these events by associating visual events to acoustic events based on their temporal overlap. In its minimal version an acoustic package consists of a single speech segment and a single visual event. In case of the maximal role acoustic packaging can take, multiple visual events are associated to a package, which is the typical case (see Figure 4.2). This initial concept using two event sources does not exclude events from further visual and acoustic cues from being associated to acoustic packages. In a second development step of acoustic packaging this case will be described (see Chapter 6).

By operating on temporal segments acoustic packaging does not require modality processing that depends on complex conceptual knowledge such as object classes or lexical information. This bottom-up design of acoustic packaging is consistent with being used in a preverbal developmental phase where complex conceptual knowledge is not available (see Section 3.2).

In summary, the core acoustic packaging system needs to perform the following tasks:

- Acoustic segmentation
- Temporal visual segmentation
- Temporal association

4.2. Related Work

Temporally segmenting continuous sensory information is required as a first step in systems that subsequently operate on chunks of sensory information. Typically these approaches are specialized to certain classes of information. Due to this specialization a high number of different applications are available and the spectrum of methods is broad. When considering acoustic and visual temporal segmentation strategies for acoustic packaging, methods and features that are potentially available to a preverbal child should be selected to avoid pre-learned knowledge. Therefore, the following review aims at identifying common methods from different domains that are consistent with these constraints. First, segmentation methods for both the visual and the acoustic modality will be reviewed. Since one of acoustic packaging's key tasks is to segment action, findings related to action demonstration towards robots and segmentation of human interaction will be reviewed additionally.

4.2.1. Acoustic Segmentation

Concerning the maximal role of the acoustic packaging system, accompanying speech fuses multiple visual events to acoustic packages. This poses the question how speech can be identified and segmented into semantic units. One option is to use prosodic features, such as intonation, rhythm, and stress for speech segmentation. Theoretical models as, for example, the prosodic hierarchy (Selkirk, 1986) consider an utterance the topmost semantic level. Utterances, however, can be segmented by using simple features such as pauses. For example Wang et al. (2003) showed that pause duration features alone allow to identify 80% of utterance boundaries in broadcast news. Since infant-directed speech is typically more structured by pauses than broadcast news, an initial segmentation of the acoustic signal into utterances provides reasonable semantic units, which can be further segmented based on prosodic features in subsequent processing steps.

Thus, first a method for segmenting speech into utterances and pauses is required. Voice activity detection (VAD) methods are able to segment the acoustic signal into speech and non-speech (pause) segments. The most basic VAD method is to use a threshold on the acoustic signal energy for segmentation. However, in real world situations such as adult-child interactions speech does not always occur in silence. Sometimes human breathing noise or noises of toys might interfere with this approach. Thus, common VAD techniques try to be noise robust (Ramírez et al., 2007).

From a technical point of view a VAD based segmentation approach is sufficient to provide an acoustic segmentation for the acoustic packaging system, as it is only required to detect the presence of accompanying speech. However, infants at 7.5 months are already capable of reacting to familiar words in fluent speech (Jusczyk, 1999). On the one hand, these findings show that acoustic segmentation based on VAD lies well within the capabilities of preverbal infants. On the other hand, only using VAD for acoustic segmentation reaches its limits when acoustic packages need to be further analyzed to identify linguistic units in a subsequent development process as described in Section 3.2.1. Therefore, acoustic processing has to be extensible to support further development of the acoustic packaging system. Extensions to the acoustic segmentation methods described in this chapter are introduced in Section 6.2.1.

4.2.2. Temporal Visual Segmentation

In computer vision systems temporal segmentation of human actions is often implicitly defined by the activation of specialized classes as, for example, the start- and endpoint of a hand trajectory. Regarding acoustic packaging a method is required that does not rely on the detection of classes depending on a high interpretation level of the sensory information. This way acoustic packaging is consistent with the limited amount of world knowledge that is available to preverbal infants. The related work on event and action segmentation reviewed in Chapter 2 shows that motion features and the capability to detect visual changes seem to be important cues for action segmentation. These features are also commonly used in systems that have a slightly different aim, namely the segmentation of video sequences. However, the problem is comparable on a technical level. 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 processing or the capability of handling multimodal input (see Table 4.1).

Scene Cut Detection

The problem of finding scene cuts 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 (Gargi et al., 1998; Janvier et al., 2006). Other work is focusing on selecting key frames within shots marked by scene boundaries (Wolf, 2002). 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 (Wolf, 2002), others are designed for offline processing (Janvier et al., 2006). The commonality is that all approaches use the visual modality only.

Action and Activity Segmentation

In many approaches, developments on action segmentation are motivated by recognizing predefined classes as, for example, in Davis and Bobick (1997); Schuldt et al. (2004). Even if generic features are used, these systems need to be trained on human labeled data (Hunter, 2009). 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 pre-trained classes a more complex approach than scene cut detection but with a similar basis is presented in Rui and Anandan (2000). 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. Recently, Buchsbaum et al. (2011) described a different approach. Here, videos displaying human activity are segmented into spatio-temporal features called visual words and clustered subsequently. The authors were able to show that changes in the distribution of these clusters correlate to human boundary judgments. Additionally, their model is able to identify further structure based on the statistical occurrence of visual words in similar contexts which is especially the case for repeated actions. Systems to find action structure have also been used to analyze parent infant interaction. In Nagai and Rohlfing (2009), a visual saliency model is used to detect structural information in parent-infant interaction. With a view on designing developmental capabilities in action learning on robots, Nagai and Rohlfing showed that their model is able to detect the initial and final states of actions as well as highlighting properties of objects.

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 4.1, most of the work focuses on one modality exclusively

Reference	Segmentation Goal	Online	Predefined Classes	Temporal Representation
Buchsbaum et al. (2011)	Human actions from multiple corpora including everyday actions	no	no	intervals
Davis and Bobick (1997)	Classes of aerobic actions	yes	yes	intervals
Janvier et al. (2006)	Scene cuts in broadcast video	no	no	boundaries
Nagai and Rohlfing (2009)	Initial and final states of a manipulation task	yes	no	boundaries
Rui and Anandan (2000)	Key poses of household chores	?	no	boundaries
Schuldt et al. (2004)	Classes of human actions (e.g. running, boxing)	?	yes	intervals
Wolf (2002)	Key frames in movies	yes	no	boundaries

Table 4.1.: Overview of visual motion segmentation approaches.

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.

4.2.3. Multimodal Event Detection and Segmentation

The benefit of processing more than the visual modality has been shown for a variety of tasks. For example Rapantzikos et al. (2007) described a saliency model processing both the visual and acoustic modality to find keyframes relevant for video summarization purposes. The resulting acoustic and visual saliency ratings are weighted and combined. The ratings are functions over time that are combined locally. Temporal relations between events are not considered in this model. Thus, the segmentation strategy is very similar to visual only video segmentation.

Since human communication is inherently multimodal, similar methods have also been applied to automatically segment meeting recordings. Here, the segmentation goal is to identify coarse grained categories. In Zhang et al. (2004), 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 HMM based approaches for automatic clustering of group actions. 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.

In contrast, some audiovisual saliency approaches try to explicitly model synchronous acoustic and visual events. The idea is that many physical events have both acoustic and visual correlates which are highly synchronous. For example a cup that is tapped on a table correlates with visual change and an acoustic energy peak. Models for audiovisual

saliency exploit this relationship to perform both spatial and temporal segmentation of audio-visual signals. In Hershey and Movellan (1999) this low-level approach is used to locate sound sources in a video signal by identifying areas that correlate with the acoustic signal. Using a method inspired by the work of Hershey and Movellan, more recently Rolf et al. (2009) found differences in synchrony between infant- and adult-directed interactions. Therefore, synchrony might be a useful cue for analyzing demonstrated actions in a tutoring scenario. However, this method focuses on acoustic and visual events that coincide on a low level, while the concept of acoustic packaging is more flexible regarding modality specific segments that overlap but differ in their on- and offsets.

4.2.4. Insights from Human-Robot Teaching Scenarios

In interactive systems such as robotic scenarios multimodal information is typically not processed in a bottom-up manner as acoustic packaging is designed. However, systems that interact with humans need to be able to segment actions depending on the scenario.

In the following, systems categorized in two different types of scenarios will be reviewed. Systems designed for programming by demonstration scenarios usually learn to repeat and adapt actions that are demonstrated towards the system. The question, how these actions are segmented is relevant for the design of the acoustic packaging system. A second very broad spectrum of interactive systems is defined by their commonality to learn from multimodal cues in interaction with humans. These systems need to segment and associate different modalities to achieve their learning goal and can thus be related to acoustic packagings tasks.

Programming by Demonstration

When human users teach actions towards a robot their interaction with the system and their demonstrations need to be segmented. The difference to temporal visual segmentation (see Section 4.2.2) is that here action sequences are typically demonstrated towards the system. The system is mostly not a distant observer and is expected to respond by imitating the demonstrations. Thus, in many systems local features such as hand trajectories or subgoals are of more interest than global features covering the whole interaction partner.

In general, programming by demonstration systems can be divided based on their action representation (Dillmann et al., 2010). One group of systems represents actions on a more symbolic level such as subactions that can be bound to visually perceived goals. Kuniyoshi et al. (1994) described one of the early systems learning the structure of block assembly tasks demonstrated by humans. In their approach a temporal segmentation method of action sequences is described that does not require explicit signals by the

human user. Initially segmentation is realized by detecting changes in the visual scene. The changes are then further analyzed and propagated to a symbolic level by classifying action primitives in the context of manipulating blocks in the scene.

Another group of systems represents actions on a trajectory level. Initially these systems primarily recorded and replayed single human demonstrations while more recent implementations are more flexible (Schaal, 1999). Many of these system require specialized — typically unimodal — features that can be translated to robot movements. Such movements are generally tracked either visually, for example, by tracking markers or by a sensor glove. The corresponding systems try to generalize from this demonstration data, which requires appropriate segments that can be grouped. Pardowitz et al. (2008) use visual cues related to the hand and object movements in order to derive a gestalt-based action segmentation. In other approaches, different kinds of inherent movement structure and implicitly coded world knowledge is used allowing for a meaningful action segmentation (Ekvall and Kragic, 2005; Kang and Ikeuchi, 1993).

In addition to the segmentation and learning methods used in programming by demonstration systems, an interaction strategy is also important since a single demonstration is usually not enough to learn a skill (Calinon and Billard, 2007). Thus, Calinon and Billard propose a system that keeps the human in the loop by incrementally learning a skill on trajectory level from multiple demonstration modalities. The robot learns an initial approximation of the task by observing tracking markers. Incorporating kinesthetic features from subsequent interactions with the tutor refines the skill.

While programming by demonstration systems mainly use specialized features such as marker tracking and scenario constraints for action segmentation the related work shows that an incremental approach is important for interactive systems to keep the tutor in the loop by providing feedback. Although these systems typically do not operate on concurrent multimodal data different modalities were sequentially exploited to support different stages in acquiring a skill. Furthermore, the related work shows that detecting visual change can be used to segment actions in a programming by demonstration scenario, although in the end a classification based on preprogrammed knowledge was performed.

Learning using Multimodal Information in Robotics

The interaction in human robot teaching scenarios is more complex than the aspects illuminated by programming by demonstration systems. Therefore, a broad spectrum of system exists that also learn from humans but their learning or interaction strategy has different aims although their learning goals overlap to a certain extent with the system described above. In the following, systems will be introduced that use multimodal input to facilitate learning using different strategies.

Similar to programming by demonstration systems, one branch of them interprets verbal and gestural instructions to execute and learn tasks that can be constructed of preprogrammed primitives. Their interaction strategy is to incorporate different inputs

such as gesture and speech to identify the relevant items and action primitives that need to be applied. In these systems interaction is mainly seen as a high level control method for robots (Biggs and Macdonald, 2003).

However, natural interaction with humans is usually more bidirectional as, for example, in collaborative tasks. Breazeal et al. (2004) describe a system that learns a task hierarchy in collaboration with a human partner. Here the robot uses speech, visual scene information, as well as gestures to recognize human intentions. Instead of solely acquiring task knowledge by observation, the robot uses facial expressions and gestures to communicate its knowledge to facilitate collaboration. Furthermore, the system communicates turns by shifting gaze from the scene to the interaction partner. However, in this approach, implicit knowledge about the action is used for the process of action segmentation, which limits the system's capabilities in this regard. Nevertheless, the findings of Breazeal et al. show that in collaborative situations, interaction needs to be processed online to provide the necessary feedback. Additionally, social feedback facilitates collaboration and structures social interaction with the system.

Similarly to the previously introduced approaches that consider task knowledge, in the domain of object learning, systems exist that associate modalities: Zhang and Weng (2003) described a system, 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 correct 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. However, during learning only very synchronous visual and acoustic features are associated. Thus, the performance of the system depends on when precisely the tutor names the object while moving it in front of the robot, showing it, and moving it out of sight again. If the object is named at the beginning or end of this action the performance is degraded. A segmentation strategy that segments the action towards the robot and exploits synchrony on a segment level could yield better results, since it does not enforce learning on highly synchronous but less optimal visual features. In contrast to the system described by Breazeal et al., here the tutor needs to provide the correct input to the system and the system's response is limited to reacting on queries. Thus, temporal segmentation of the interaction is provided by the tutor.

The latter point — providing the right input — can be an issue for inexperienced users interacting with a robot, since they have very limited information on the robot's capabilities. A passive system that only responds to queries can even generate false expectations and users may have to determine the right interaction strategy with the robot in an iterative and frustrating process. This problem can be addressed by robots that actively communicate towards humans and thus provide information on how to interact with them. An implementation of this idea is described by Lütkebohle et al. (2009). In an interactive object learning scenario the system initiates the dialog by pointing to an object and asking for its label. Thus, the system provides interaction structure to the

user by communicating which information it is interested in and, therefore, probably able to process. Furthermore, associating the linguistic information provided by the human interaction partner to objects is simplified. Although the learning process uses symbolic information from a dialog system these findings suggest that further developed systems which can take the initiative and provide feedback benefit from receiving information they require.

The previously described systems mainly use visual and acoustic cues and rely on synchrony for segmenting sensory information and learning. Only few systems exist that make use of other amodal cues. In Fitzpatrick et al. (2006) a system is described that uses rhythm to identify repeated information from several modalities and synchrony to detect their causal relation. Their approach uses visual input, acoustic input, and proprioceptive joint data available on the underlying robotic platform. On each modality a period estimation and segmentation step is performed to extract repeating units from the signal. The resulting units are then associated across modalities if they temporally coincide within a certain tolerance. The system has been tested with different rhythmic actions that span modalities as, for example, tambourine shaking (visual–acoustic) and shaking the robot's own arm (proprioceptive–visual–acoustic). The authors argue that including proprioceptive information allows the robot to acquire knowledge about its own body characteristics. Furthermore, the segmentation method facilitates identifying generalizable visual and acoustic information since it tends to strip context dependent information. However, when a robots needs to segment and learn goal oriented actions, this method is not able to segment the actions or the relevant objects since goal oriented actions are typically not rhythmically repeated. In certain situations, where, for example, a tutor tries to get the attention of the system by tapping an object this method is able to segments these parts.

4.2.5. Summary

Acoustic packaging can be related to findings in acoustic segmentation, temporal visual segmentation, and multimodal event detection. Further relevant findings are provided by models for multimodal processing in robotic systems since they are not only tuned towards perception but additionally consider interactive aspects. The related work in this domain provides insights on how systems for multimodal action segmentation can be designed.

The acoustic modality can be segmented into utterances using voice activity detection methods. This assumes that utterances are bounded by non-speech or silence, which is in accordance with its phonetic definition. Considering the high level of structuring in child directed speech, this seems to be a valid approach. However, during further development a more detailed segmentation might become necessary.

For the visual modality various segmentation methods are available. In many cases they are tuned towards their segmentation goal as, for example, scene changes in movies or human activities (see Section 4.2.2). The general idea behind visual segmentation is detecting temporal discontinuities. Visual segmentation methods use features that represent the change in visual input either globally or more locally by analyzing movement features such as human body or hand motion.

Methods operating on multimodal data can be found in segmentation of meeting recordings and in multimodal saliency models. In this context multimodal segmentation is realized by combining features from the visual and acoustic modality. How these features are combined depends on the relationship between the modalities and the segmentation goal. The group of multimodal saliency models assumes physical relationship between the visual and acoustic information they process. Therefore, they tend to combine features locally assuming a high level of synchrony between the relevant features.

On a more abstract level such as the classification of group actions in meeting recordings, speech is not necessarily highly synchronous with actions (see Section 4.2.3). Thus, the temporal relationship between the visual and acoustic modality is statistically modeled to allow for these deviations. This relationship also impacts the temporal representations used. Most methods focus on detecting single events leading to representations which are based on boundaries rather than interval based temporal representations. The latter is typically the case when the input data is classified after segmentation.

In the context of robotics further unimodal and multimodal action segmentation methods have been developed. In programming by demonstration systems, action is typically segmented based on goals of the ongoing action. The problem is simplified by tracking specific features such as hands or markers. Many systems represent action on a symbolic level to enable linking them to preprogrammed action primitives. Non symbolic action learning uses trajectory level representations. Action segmentation is usually performed by detecting goals or by other preprogrammed world knowledge. In addition to visual input a number of systems incorporates speech to allow for naming action primitives or objects in interaction with humans. Only few systems in robotics use more cues for segmentation than synchrony such as detecting repetitions to find structure in the multimodal sensory stream (see Section 4.2.4).

One important difference in robotics compared to solely perceptual models is that interaction with humans is regarded. Keeping the human in an interactive loop supports learning algorithms that need to acquire a sufficient amount of data to be effective. Additional reasons for interaction are that systems need to acquire the right data. Thus, the user needs to know how to interact with the system and which information should be provided. Designing proactive systems has been shown to be effective in solving this problem, since they can actively communicate to the user in a way that structures the interaction. These aspects will be considered in the design of the acoustic packaging system, which will be described in the next section.

4.3. The Acoustic Packaging System

The previous summary offers a number of methods on how to segment acoustic and visual sensory input, that can be considered in designing a system for acoustic packaging. The requirements for the underlying model of acoustic packaging result from selecting methods for action segmentation that are sensible from an engineering viewpoint as well as with regard to the psychological findings described in the Chapters 2 and 3. Further design issues concern properties of acoustic packaging necessary for its application in interactive robotic scenarios. These aspects are discussed in the following section. Subsequently the implementation of the acoustic packaging system will be described. Portions of this section were previously published by the author (Schillingmann et al., 2009b).

4.3.1. Requirements

As a first step towards the development of a computational model of acoustic packaging the *segmentation problem* has to be solved. The related work reviewed previously, shows that detecting change is a common concept in systems that segment visual and acoustic input. This approach is consistent with the psychological findings on action segmentation and acoustic packaging (see Chapters 2 and 3). The general role of acoustic packaging is to segment multimodal action demonstrations into learning units that can further be processed in a developmental action and language learning context. Hence, the segmentation methods should not require pre-trained classes which makes the idea of change detection for segmentation a sensible candidate. Since the model has to make use of at least one visual and one acoustic cue, a *temporal segmentation* for both cues is required. The segmentation problem is addressed for both cues in detail in Sections 4.3.3 and 4.3.4.

A second problem is the *temporal synchronization* of these sensory cues. The difficulty here is, that hypotheses from audio and vision processing are typically generated neither at the same time nor in the same rate. One reason is the different temporal resolutions necessary for visual and acoustic processing. Furthermore, visual and acoustic processing might introduce different processing latencies. 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. Furthermore, the system should not only look for events that are aligned with events from the other modality in a strictly synchronous way but allow for tolerance when associating events by looking at their temporal overlap. Although accompanying speech overlaps with action demonstrations, it does not imply highly synchronous on- or offsets as, for example, in physical causal events.

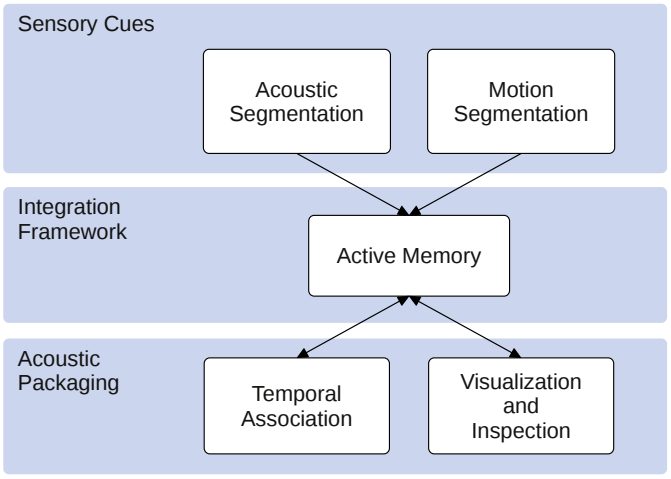


Figure 4.3.: System overview with highlighted layers and their relation to the acoustic packaging system.

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 has to be *online* usable and able to cope with updating hypotheses. However, the system should not be limited online sensory sources but also support offline processing to produce repeatable analyses of interaction data.

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.

4.3.2. System Overview

The system for acoustic packaging proposed here consists of four modules (see Figure 4.3). These modules communicate events through a central memory, the so called Active Memory (Fritsch and Wrede, 2007). The Active Memory notifies components about event types they have subscribed to and is able to store these events persistently. It establishes 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. Furthermore, the Active Memory can be queried for events that are persistently stored to allow for inspection tools that access and analyze events processed in the past.

All signal processing modules are connected to the Active Memory. In this configuration of the system two modules process the acoustic and visual modality and insert the resulting events into the memory. Acoustic packaging is performed by a temporal association

module that subscribes to modality specific events, forms acoustic packages, and inserts acoustic packaging event types into the memory. The visualization and inspection module is able to listen to all event types within the acoustic packaging system as well as to retrieve events persistently stored.

All events are modeled as temporal intervals that contain basic information about their temporal location within the sensory stream the system processes. Thus, all events have a `begin` and an `end` timestamp in common. For online processing these timestamp values refer to the time the signal was acquired as opposed to the time the event is generated which is affected by processing delays. During offline processing, the timestamps are determined on the basis of the current sample or frame number while reading a file. This method ensures that offline processing generates repeatable results where relations between events are temporally consistent regardless of the order they are processed.

During processing the system has to handle possibly unstable hypotheses that are subject to change in the short future. Thus, all events possess a `stable` attribute to inform further processing modules about the state of each event. The rationale here is to support different latency classes. A future module processing events for long-term learning might require only stable events while a module processing events to provide feedback requires very recent but not necessarily stable information.

4.3.3. Acoustic Segmentation

As described in Section 4.2.1 infant-directed actions exhibit more, and more structured pauses than adult-directed actions, it seems appropriate to segment the acoustic signal simply into speech and non-speech (pause) segments. Related to the perceptual mechanism of change detection (see Section 4.3.1) the idea is to segment the acoustic modality based on changes in voice activity. Yet in a relatively noisy environment such as the described experimental setting (see Section 4.1), the separation of speech from non-speech is a difficult task. Therefore, instead of a simple voice activity detection algorithm based solely on a signal energy threshold, a more sophisticated approach is used: The audio signal is processed using the ESMERALDA speech recognizer (Fink, 1999), which is configured to use an acoustic model for monophoneme recognition derived from a model based on the Verbmobil corpus (Kohler et al., 1994). Phonotactics are modeled statistically via an n-gram model. An acoustic segment is defined as speech framed by non-speech. Since the acoustic model contains noise models in addition to the phoneme models, non-speech is more robustly recognized as if only the signal energy was used as a criterion for voice activity. As a consequence, a continuous chain of phoneme hypotheses generated by the speech recognizer is considered as a speech segment. The 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 until the speech segment ends and is marked

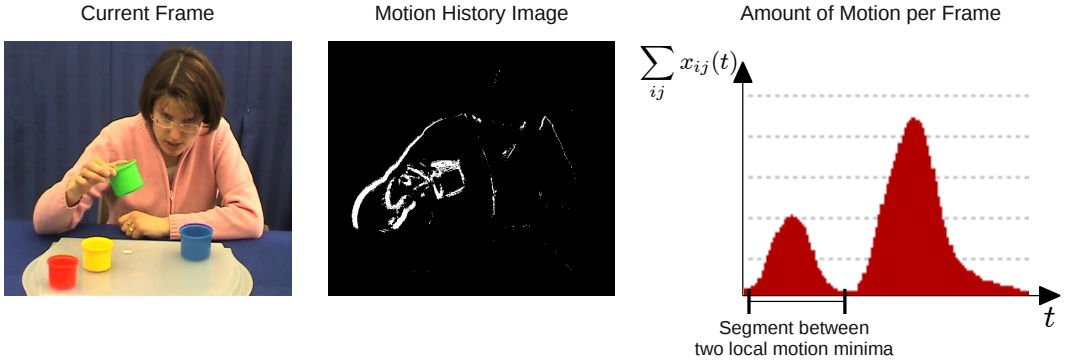


Figure 4.4.: The left image depicts a person showing a cup. The middle image displays the corresponding motion history image. The right image illustrates the approach to visually segment actions via the amount of motion per frame.

as stable. For the case of online speech segmentation a typical configuration introduces a delay of 300 ms until the speech recognizer assumes that the corresponding phoneme hypotheses have become stable during incremental speech recognition.

4.3.4. Visual Action Segmentation

The design decisions regarding visual action segmentation follow the idea to use simple features that do not require to include a large amount of previous knowledge into the system. Furthermore, these features need to be consistent with the existing psychological findings on event and action segmentation (see Chapter 2). Taking the existing findings on automatic visual temporal segmentation and event segmentation into account (see Sections 4.2.2 and 4.2.3), movement features have been successfully used as a common cue for segmentation. Another cue that is closely related is visual change which correlates to motion. Change detection and segmentation based on these cues can be realized by finding discontinuities in the visual signal. Thus, the visual signal is segmented at minima in local motion into motion peaks. Each peak ranges between two local minima 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 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 this heuristic approach to segment actions into motion peaks.

The segmentation into motion peaks is technically realized by an approach based on motion history images (Davis and Bobick, 1997). A graphical plugin environment (Lömker et al., 2006) has been used as framework to implement the visual segmentation method

Parameter Name	Value
Motion history size	10 frames (for 25 fps video input)
Peak detection window size	14 frames (for 25 fps video input)
Minimal motion threshold	0.001% of the image's pixel count
Minimal relative motion peak height	0.005% of the image's pixel count

Table 4.2.: Values of relevant parameters for the motion segmentation module in a typical configuration.

as a reusable plugin. The plugin reads input from a plugin developed by Ingo Lütkebohle, which provides motion history images. In the following, the visual segmentation method is described.

The amount of motion is calculated per frame by summing up the motion history image (see Figure 4.4). 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 yet small enough to not affect human movements. In addition, the peak height relative to the amount of motion at the position of the local minima is calculated. This relative peak height can optionally be used to filter small local peaks. In a typical configuration the parameters for the visual action segmentation module are set as depicted in Table 4.2.

The 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, this approach is designed to limit prior knowledge with respect to space and content of visual information to be consistent with the bottom-up strategy the acoustic packaging approach follows. It has to be noted that children possess certain prior knowledge as, for example, a preference for attending faces, which is suspected to be innate (Rosa Salva et al., 2011).

When a local minimum is detected, an event describing the motion peak between the previous and the current motion minimum is inserted into the Active Memory. The description contains the peak's time interval and the frames at the minima from the beginning and end of the motion peak. Furthermore, the position of the maximum as well as the absolute and relative peak height are included in the description. However, this method would insert the most recent peak only if the next minimum has been detected, which would introduce a delay. With respect to the requirement of online processing a partial description of the most recent peak is continuously reflected into the Active Memory as long as the next local minimum has not been determined. Afterwards the current description is marked as stable.

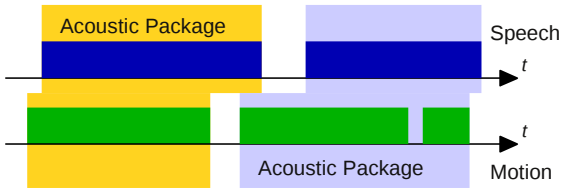


Figure 4.5.: Motion and speech intervals are assigned to an acoustic package if they overlap. The middle motion interval has been assigned to the second acoustic package due to greater overlap.

4.3.5. 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. The temporal association module subscribes to events communicated through the Active Memory and maintains a timeline for different types of time intervals. In the following, the processing of motion peaks and speech segments is considered. When a new event arrives, the segment is aligned to its modality-specific 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 Allen (1983) 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 4.5 for the association process). Thus, a motion segment cannot bind multiple speech segments together. However, multiple motion segments can be associated to one speech segment to form an acoustic package. Therefore, the length of an acoustic packages is in general larger than a single utterance. An example of multiple motion peaks, which were associated to one acoustic package, is depicted in Figure 4.6. A typical acoustic package has an average length of three seconds and contains 1.5 motion peaks. In contrast, a typical utterance length from adult-child interaction segmented by the system is about one second long. Thus, the combination of visual and the acoustic segments provides a higher level segmentation than considering the individual modalities. In Chapter 5, more examples of segmentation into acoustic packages will be discussed, and a detailed evaluation on the properties of acoustic packages will be presented.

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 the 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 asynchrony between the incoming events of the different modalities. This requires handling potential processing delays on the one hand. On the other hand, it eases

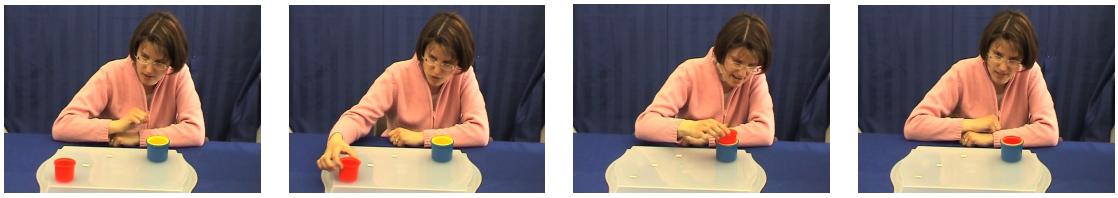


Figure 4.6.: Frames at the beginning and end of three motion peaks which were associated to one acoustic package because of temporal overlap with the utterance “the red one into the yellow one”.

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 (see Figure 4.7). 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.

The temporal association module is not only a key component of the acoustic packaging system because it associates events and forms acoustic packages. It provides extension points to the system as, for example, if further segmentation modules are integrated the segments can be associated to acoustic packages by adding a timeline for the respective event types. The resulting acoustic packages can be processed by modules that learn or generate feedback in robotic scenarios which is another extension point. Both temporal processing and accessing synchronized information that spans modalities is simplified for modules operating on acoustic packages. Since acoustic packages are also persistently stored in the Active Memory they can be recalled later to analyze past segmentation results. This property is also important for the visualization and inspection module described in the next section.

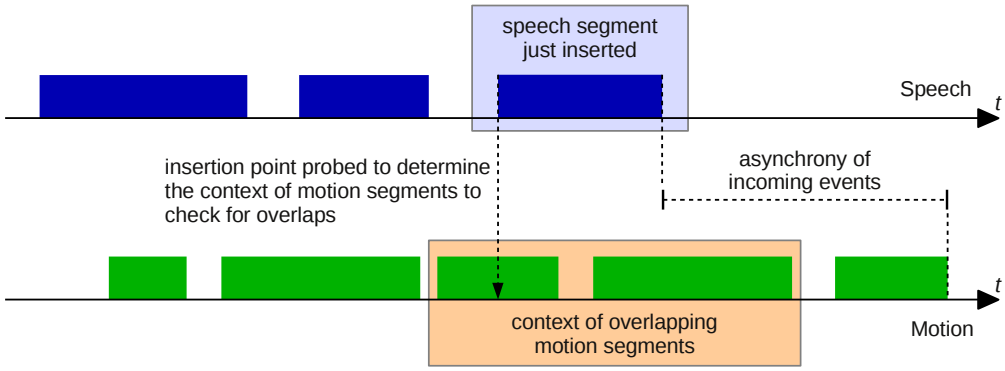


Figure 4.7.: The temporal association module can handle any asynchrony between input cues by maintaining timelines for each modality. The Figure illustrates how the arrival of a new speech segment is efficiently handled.

4.3.6. 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 4.8 shows the 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 but empty row displays further cues that will be introduced in Chapter 6. The third row shows the signal energy that gives an estimate about speech activity. The forth 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 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 4.9 shows the inspection tool, which is able to query all segmentation hypotheses from the Active Memory. The inspection tool can be combined with the tool for visualization of the cues (Figure 4.8) 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. Figure 4.9 shows the inspection tool in a state where it displays details of acoustic packages namely the temporal extents of acoustic packages and the segmentation hypotheses associated to it. The corresponding interval is automatically highlighted in the cue visualization window (see Figure 4.8, rows 4–6). The modality specific segmentation hypotheses can also be inspected by selecting the respective modality in the tab view to provide means of analyzing the segmentation results. Speech, for example, can be replayed to assess the speech segmentation results.

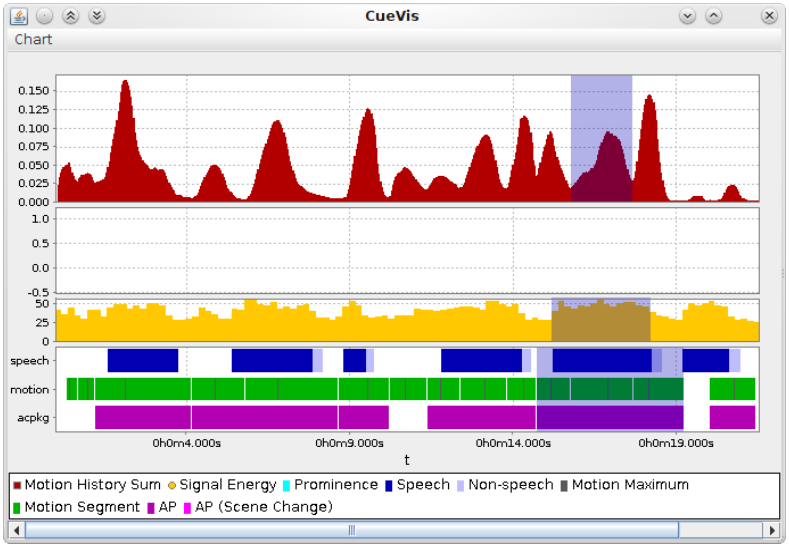


Figure 4.8.: Cue visualization tool showing motion peaks (row 1), acoustic signal energy (row 3), speech segmentation (row 4), visual segmentation (row 5), and acoustic packages (row 6).

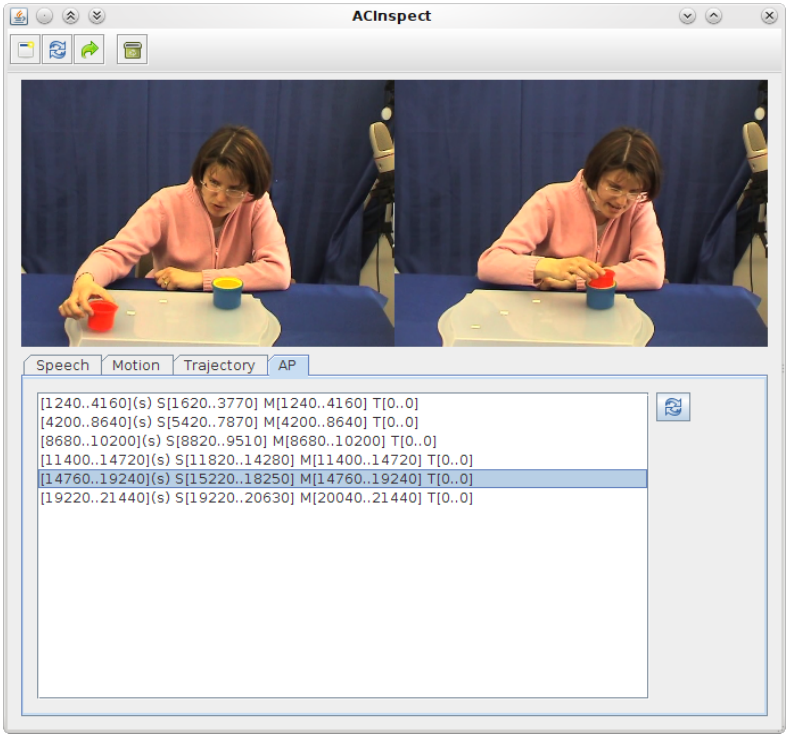


Figure 4.9.: Inspection tool showing a list of acoustic packages with details on each package's temporal extent and its associated segmentation hypotheses.

Furthermore, if a motion peak is selected, the inspection tool displays the frames at the beginning and the end of the selected peak. Taken together, these features of the inspection tool help to rate, optimize and debug the acoustic packaging system and its parameters.

4.4. Conclusion

Acoustic packaging can be viewed as perceptual mechanism that assists infants in binding sequences of observed actions in order to form one meaningful unit. In this chapter a computational approach towards modeling acoustic packaging in a tutoring scenario has been described. The present approach was developed with scenarios for human-robot interaction and the analysis of interaction data in mind. The implementation uses simple bottom-up cues to segment the signal streams of each modality. Speech is segmented into utterances and the visual signal is segmented into motion peaks. The segments derived are then bound together based on their temporal overlap relations defining a new segmentation of the multimodal input stream into acoustic packages.

The modality specific segmentation methods have been selected with respect to psychological findings on event and action segmentation as well as multimodal processing (see Chapters 2 and 3). In accordance with these findings, the review of existing systems and methods (see Section 4.2) identified motion features as promising candidates for visual segmentation. More complex methods are typically tuned towards specific tasks and features and thus are not compatible with the view of acoustic packaging as an early segmentation and binding process in preverbal children with limited world knowledge. For speech segmentation voice activity detection has been selected as method, since it reflects structuring with pauses. The binding process maintains a timeline for each modality to handle asynchrony of speech and motion hypotheses and to efficiently determine the necessary overlap relations to form acoustic packages.

The system design follows a modular concept and is capable of online processing multimodal input both offline and online. Furthermore, the architecture facilitates the integration of further processing modules. To this point, modules for acoustic segmentation, visual action segmentation and temporal association have been presented. A visualization and inspection module helps to debug and further optimize the system. As a whole, the systems capabilities fulfill the prerequisites necessary for being integrated in robotic platforms. The presented approach is a solid basis for supporting action learning in human-robot interactions. On the one hand, the derived acoustic packages form multimodal units that can be used for further learning processes, such as simple mapping processes between words and objects. On the other hand, acoustic packages binding visual and acoustic information into a sequence promise to be a valuable resource for human-robot interaction in a tutoring scenario as they provide cues for a possible feedback.

Acoustic packaging should thus reflect how tutors structure their action demonstrations, which has to be verified. This point will be taken up in the following chapter, where evaluations of the system system on interaction data will be analyzed.

5. Acoustic Packaging as Analysis Tool for Multimodal Interaction

In the previous chapter the basis implementation of the acoustic packaging system was described but no evaluation results were reported. Therefore, this chapter focuses on results generated by processing multimodal interaction data with the acoustic packaging system. First, this chapter will deal with the evaluation of the acoustic packaging system. The evaluation methodology used here compares differences between adult child and adult adult interaction. First, the reasons for this procedure will be discussed. Subsequently, the evaluation results will be reported and discussed. Second, the acoustic packaging system will be used to analyze adult-child interactions with respect to different child age groups and compare them to adult-adult interaction. Third, adult-robot interaction will be analyzed and related to the previous results.

5.1. How can Acoustic Packaging be Evaluated?

It is important to emphasize that the acoustic packaging system delivers bottom-up hypotheses for the segmentation of action demonstrations in tutoring situations 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 used in interaction frequently can verify and refine these hypotheses. This requires an evaluation of acoustic packaging within interaction, which makes it problematic to evaluate it as a subsystem alone. Since acoustic packaging is based on a process early in infant development it should have no access to complex conceptual information that allows for evaluation in terms of classification accuracy.

This problem persists when agreement of segmentation boundaries with adults is used as an evaluation method. On the one hand, this evaluation method has already identified that human event segmentation correlates with motion features. The corresponding findings from the area of event and action segmentation are summarized in Chapter 2. On the other hand, if adults segment action, their existing conceptual knowledge likely interacts with their boundary placement decisions. Thus, at a lower level several sensible

Figure 5.1.: An adult demonstrates how to stack cups to a child.

action segmentations are possible. In the following, two examples are presented, which illustrate how action is segmented into acoustic packages. Based on these examples, we will discuss issues concerning the evaluation of segmentation correctness.

The first example stems from a tutoring situation (see Figure 5.1). In this situation, a mother is 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 “into 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 moving the cup 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 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 may 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

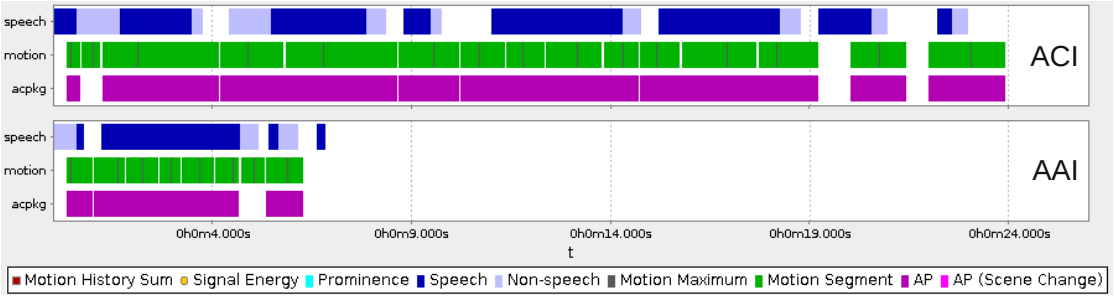


Figure 5.2.: Segmentation of a stacking cups task into acoustic packages in two conditions. The adult's interaction with a child (ACI) is compared to the acoustic packaging results of the same adult's interaction with another adult (AAI).

level on which acoustic packaging operates. Therefore, a different approach to provide an initial evaluation of the acoustic packaging system is chosen here. The idea is to compare tutoring behavior between two interaction conditions which differ in the tutor's interaction partner. In one condition the tutor interacts with a child, while in the other condition the learner is another adult.

5.2. Evaluation of Acoustic Packaging on Adult-Adult and Adult-Child Interaction Data

With reference to previous research (Zukow-Goldring, 1996; Rohlfing et al., 2006), it can be hypothesized, that parents structure their actions more when interacting with their children. Parents use shorter utterances and synchronize demonstrations, as, for example, showing an object more frequently with speech as compared towards adults (see Section 3.2.3 for more findings in this direction). Therefore, the acoustic packaging system is expected to generate more packages in an adult-child condition than in an adult-adult condition. Another expectation is that adult-adult interaction (AAI) is less structured when compared to adult-child interaction (ACI). Since adults perform their actions and narrations more fluently when interacting with each other, a larger amount of motion segments per package is expected than compared to the adult-child condition. Both effects can be observed in Figure 5.2 which depicts a segmentation of two interactions into acoustic packages. In the first interaction, an adult demonstrated a stacking cups task to a child, while in the second interaction, the same adult demonstrated this task to another adult. Figure 5.2 shows a higher number of acoustic packages in ACI compared to AAI. Furthermore, a higher number of motion segments per acoustic packages in AAI compared to ACI can be observed. Both hypotheses seem to be supported by this example. However, to verify these hypotheses they must be evaluated on a larger sample. This process is described in the following. Portions of this section were previously published by the author (Schillingmann et al., 2009b).

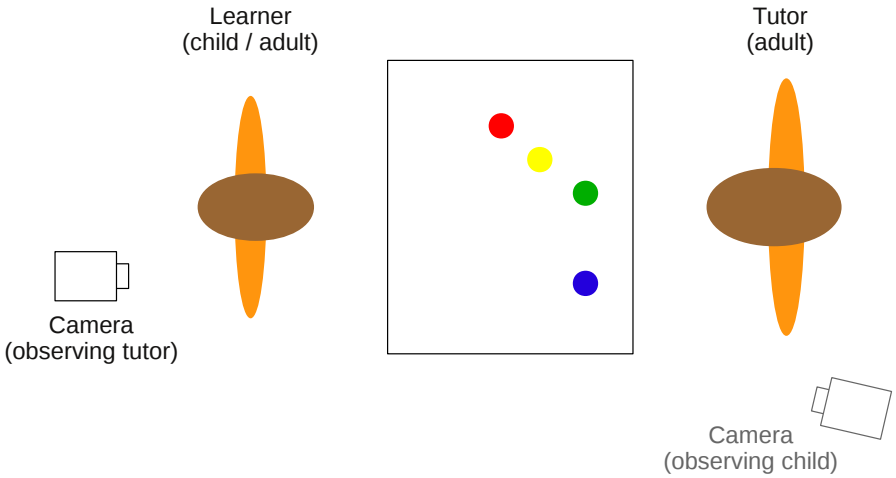


Figure 5.3.: Adult-Child / Adult-Adult interaction setting. The interaction partners are seated at a table facing each other. In this evaluation, recordings from the camera observing the tutor are used.

5.2.1. Corpus Overview

For this evaluation a subset of a corpus, containing video and audio data on adult- and infant-directed interactions in a tutoring situation, is used (Rohlfing et al., 2006). From this corpus, 11 participants interacting with their 8 to 11 month-old children were selected. The participants were asked to demonstrate functions of 10 different objects to their children as well as to another adult (partner or experimenter). The evaluation reported below focuses on one task, namely the stacking cups task. The setting is depicted in Figure 5.3. The following description refers to the data from the camera that recorded all actions of the adult tutor in the cup stacking task. A view from the camera's perspective is illustrated in Figure 5.1.

5.2.2. Procedure

The acoustic packaging system was exposed to the multimodal data described in the previous section. The audio data was normalized beforehand due to highly variable gain and noise levels. After processing each interaction the Active Memory is queried for acoustic packages, the statistics described in the next section are calculated, and the system is reset to process the next interaction. Statistical tests were calculated using the R software package (R Development Core Team, 2011).

	Adult-Adult-Interaction			Adult-Child-Interaction		
Pair	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

Table 5.1.: Counts of acoustic packages (AP) and motion peaks (M) on participants in adult-adult interaction compared to the same adults interacting with children.

5.2.3. Evaluation Results

The first hypothesis predicts a higher number of acoustic packages in ACI compared to AAI. For this purpose 11 videos with adults demonstrating the stacking of cups to children were compared with 11 videos of the same adults demonstrating the same task to an adult (see Table 5.1). A Wilcoxon signed rank test revealed a significant difference in the amount of acoustic packages between these groups: $W = 0,\ Z = -2.900,\ p = 0.002$. This result strongly suggests that more acoustic packages can be found in an interaction with a child.

The second hypothesis expects a larger amount of motion segments per package in AAI compared to the ACI. This hypothesis was tested by applying a Wilcoxon signed rank test on the ratio of motion peaks to acoustic packages in both conditions. A significant difference was found: $W = 66,\ Z = 2.934,\ p = 0.001$. This result strongly suggests that more motion segments are packaged together in an interaction with an adult. Table 5.1 shows the motion peak counts per participant.

What is also noticeable, is that in adult-adult interaction, the variance of motion peaks per acoustic package is higher than in adult-child interaction. This is due to the fact the participants displayed highly individual communication styles: For example, some participants 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 participants behaved 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 to note that in adult-child interaction, the variance is lower. This suggests that adult-child interaction is not affected by the participant's specific communication style to the same extent as it is in an adult-adult interaction.

5.2.4. Discussion

The results show that when comparing the same participants 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 interaction, and this is in line with previous findings (Brand et al., 2002; Brand and Tapscott, 2007; Rohlfing et al., 2006; Zukow-Goldring, 2006).

Based on these results, one 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.

As exemplified in Section 5.1, this evaluation argues for showing that acoustic packaging is able to reflect the differences between adult-child and adult-adult tutoring behavior. The main reason for this comparative method is that an assessment of segmentation correctness is difficult since multiple sensible action segmentations are possible. Another reason why it is neither desirable nor 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 system is packaged although it is not related to object manipulation in 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. However, additional cues might help to filter acoustic packages respectively. One idea is that such cues allow for discrimination of packages containing communication cues from packages that are related to scene changes. In Chapter 6, further cues are introduced, providing first steps in this direction.

The method proposed here has been applied to interactions containing tutoring situations, in which the tutor performed manipulative actions. This specific situation thus 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 relies on this assumption: Acoustic packaging segments motion by finding discontinuities in the visual signal as a 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 could still 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 (i.e. a rather 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 by including social information. This course of development is supported by the Emergentist Coalition Model (Hollich et al. (2000b), see Section 3.2.2), which makes a statement about the cues that children take into account when learning words: Initially, higher weights are given to perceptual cues. During further development, social cues play an increasingly important role. In sum, the present choice of cues in the acoustic packaging system is sensible concerning first developmental steps. To support more complex social interaction during further development, the set of cues can be extended. This will be done in Chapter 6, which also includes an analysis of the semantic content and representational capabilities of acoustic packages.

5.3. Analysis of Adult-Adult and Adult-Child Interaction

In the previous section acoustic packaging was assessed based on statistical properties of adult-adult and adult-child interaction data. Here, acoustic packaging will be used to further compare adult-adult and adult-child interaction, by including additional age groups from the corpus (Rohlfing et al., 2006). Furthermore, it will be analyzed if statistical properties of acoustic packages change during development.

5.3.1. Corpus Overview

The scenario used for the evaluation is the same as described in Section 5.2.1. But here the acoustic packaging system processed the data from all age groups. For the analysis of adult-child interaction data 64 participants were processed in total. The 64 adult participants are divided into four groups according to the age of their child. The children's age ranges from 8 to 30 months. An overview of the individual groups can be found in Table 5.3. For the analysis of adult-adult interaction, data of 66 participants was processed. In contrast to the previous analysis no pairwise comparison of the adult-child and adult-adult condition is performed which allows the inclusion of more participants from the first group. The reason is that pairwise comparison only allows to include participants that complete both conditions without problems, such as, for example, a crying child.

5.3.2. Procedure and Design

The acoustic packaging system was exposed to the data described above as described in Section 5.2.2. The difference here is that manual annotation is used for acoustic segmentation into utterances. The reason for this lies in the fact that children become more verbal with increasing age. Since the acoustic data is not recorded with a close-talking microphone, voice-activity-based acoustic segmentation would also segment the child's voice which is not desired in this evaluation.

To provide an overview of the relation between the content of acoustic packages and the interaction structure within tutoring situations several measurements were calculated. The results are divided into two tables: In Table 5.2 statistics on adult-adult and adult-child interactions are presented. For each item a Wilcoxon Mann-Whitney rank sum test has been calculated to assess if a significant difference can be assumed comparing the AAI and the ACI condition. Additionally, on the adult-child data, the same measurements have been calculated for the four separate age groups described previously (see Table 5.3). Tables comparing AAI and ACI for individual age groups are provided in appendix A. The measurements are structured into conceptual groups, which are described in the following.

The first group of measurements concerns properties of acoustic packages themselves; it consists of their total number per trial, their total length per trial as well as their average lengths. Furthermore, statistics of motion peaks which have been associated to acoustic packages are included (see Table 5.3, rows 2–7). The second group refers to the individual modalities which are used to form acoustic packages (see Table 5.3, rows 8–17). Here, measurements of motion peaks, utterances, and pauses are included. The third group summarizes different ratios between acoustic packages and modality specific segmentations (see Table 5.3, rows 18–23). The different groups are used in the following to analyze differences between adult-adult and adult-child interaction as well as finding developmental trends. For this purpose the children's age and the relevant measurements from each group were correlated, while assuming monotonically raising or falling values with increasing age. Therefore, Spearman's rank correlation coefficient ϱ was calculated to verify this assumption.

5.3.3. Results on Individual Modalities

In the following, modality specific segmentation results on adult-child interaction (ACI) and adult-adult interaction (AAI) will be presented. These segmentations are used by the temporal association process to form acoustic packages (see Section 4.3.5). Acoustic packaging results will be presented in the subsequent sections.

Both the number of utterances and motion peaks differ significantly between adult-child interaction and adult-adult interaction. ACI contains a higher number of utterances compared to AAI (see Table 5.2, row 11). Also the number of motion peaks is higher

		ACI M (SD)	AAI M (SD)	ACI-AAI Z	p
1	Number of participants	64	66		
2	Total number of APs	10.33 (6.17)	4.11 (2.06)	7.3	0.000
3	Total length of APs [s]	30.33 (19.79)	14.85 (8.57)	5.8	0.000
4	Average length of APs [s]	2.90 (0.70)	3.70 (1.28)	−4.1	0.000
5	Total number of MPs (in APs)	15.44 (8.88)	8.62 (4.79)	5.5	0.000
6	Total length of MPs (in APs) [s]	18.23 (10.37)	8.40 (4.63)	6.6	0.000
7	Average length of MPs (in APs) [s]	1.19 (0.21)	0.99 (0.19)	5.4	0.000
8	Total number of MPs	21.67 (10.56)	11.67 (4.82)	6.5	0.000
9	Total length of MPs [s]	24.36 (12.47)	10.71 (4.52)	7.6	0.000
10	Average length of MPs [s]	1.13 (0.18)	0.93 (0.16)	5.8	0.000
11	Total number of utterances	11.97 (7.80)	4.47 (2.19)	7.6	0.000
12	Total length of utterances [s]	9.72 (5.87)	6.08 (3.94)	4.2	0.000
13	Average utterance length [s]	0.87 (0.39)	1.48 (1.04)	−5.1	0.000
14	Average utterance length (in APs) [s]	0.93 (0.43)	1.55 (1.03)	−5.0	0.000
15	Total number of pauses in speech	10.97 (7.80)	3.47 (2.19)	7.6	0.000
16	Total length of pauses in speech [s]	14.78 (9.79)	3.29 (2.42)	8.4	0.000
17	Average length of pauses in speech [s]	1.42 (0.55)	1.01 (0.95)	5.0	0.000
18	Average number of MPs per AP	1.54 (0.34)	2.25 (0.96)	−5.4	0.000
19	Ratio of interaction length to speech length	3.68 (4.13)	2.36 (1.31)	4.5	0.000
20	Ratio of AP length to speech length (in APs)	3.67 (2.05)	2.92 (1.77)	3.6	0.000
21	Ratio of AP count to speech length (in APs) 1/[s]	1.26 (0.81)	0.83 (0.51)	4.7	0.000
22	Ratio of all MPs to MPs assigned to APs	1.60 (1.02)	1.53 (0.72)	0.7	0.466
23	Ratio of interaction length to AP length	1.07 (0.80)	0.91 (0.42)	2.0	0.048

Table 5.2.: Results from the comparison of child-directed versus adult-directed interaction (all age groups together). The right columns show the results of Wilcoxon Mann-Whitney rank sum tests between ACI and AAI.

		Group 1 8–12 months M (SD)	Group 2a 12–18 months M (SD)	Group 2b 18–24 months M (SD)	Group 3 25–30 months M (SD)
1	Number of participants	24	12	10	18
2	Age of children [months]	10.06 (1.08)	16.52 (1.43)	20.44 (1.75)	26.15 (1.63)
3	Total number of APs	13.25 (7.33)	6.58 (4.91)	11.70 (5.79)	8.17 (2.66)
4	Total length of APs [s]	39.30 (25.57)	17.55 (13.09)	31.50 (15.45)	26.24 (9.06)
5	Average length of APs [s]	2.88 (0.55)	2.58 (0.75)	2.68 (0.46)	3.28 (0.82)
6	Total number of MPs (in APs)	18.25 (10.94)	10.58 (6.92)	18.30 (9.06)	13.33 (4.33)
7	Total length of MPs (in APs) [s]	22.48 (12.22)	11.05 (7.54)	21.26 (10.21)	15.68 (5.49)
8	Average length of MPs (in APs) [s]	1.27 (0.25)	1.04 (0.16)	1.19 (0.16)	1.17 (0.14)
9	Total number of MPs	26.79 (12.59)	15.75 (5.40)	24.10 (11.70)	17.44 (4.85)
10	Total length of MPs [s]	31.21 (14.12)	15.73 (6.05)	27.33 (13.67)	19.34 (5.78)
11	Average length of MPs [s]	1.20 (0.23)	0.99 (0.11)	1.15 (0.16)	1.11 (0.10)
12	Total number of utterances	15.38 (9.84)	7.42 (6.01)	13.60 (6.40)	9.56 (3.29)
13	Total length of utterances [s]	10.52 (6.33)	6.28 (4.92)	12.56 (6.51)	9.36 (4.58)
14	Average utterance length [s]	0.72 (0.23)	0.88 (0.44)	0.92 (0.20)	1.02 (0.54)
15	Average utterance length (in APs) [s]	0.75 (0.23)	0.91 (0.44)	1.02 (0.24)	1.13 (0.60)
16	Total number of pauses in speech	14.38 (9.84)	6.42 (6.01)	12.60 (6.40)	8.56 (3.29)
17	Total length of pauses in speech [s]	21.65 (10.84)	7.17 (5.25)	15.56 (8.21)	10.28 (3.62)
18	Average length of pauses in speech [s]	1.68 (0.68)	1.27 (0.46)	1.26 (0.36)	1.24 (0.32)
19	Average number of MPs per AP	1.37 (0.20)	1.65 (0.45)	1.56 (0.21)	1.67 (0.38)
20	Ratio of interaction length to speech length	3.63 (1.46)	5.78 (8.80)	2.51 (0.48)	2.99 (2.36)
21	Ratio of AP length to speech length (in APs)	4.08 (1.63)	3.14 (1.00)	2.76 (0.85)	3.99 (3.15)
22	Ratio of AP count to speech length (in APs) 1/[s]	1.34 (0.44)	1.35 (0.96)	0.96 (0.24)	1.26 (1.22)
23	Ratio of all MPs to MPs assigned to APs	1.55 (0.45)	2.30 (2.17)	1.34 (0.17)	1.35 (0.27)
24	Ratio of interaction length to AP length	0.97 (0.31)	1.65 (1.68)	1.04 (0.31)	0.84 (0.20)

Table 5.3.: Acoustic packaging statistics on adult-child interaction by age groups.

in ACI than in AAI (see Table 5.2, row 8). For the number of utterances in ACI there is no significant trend concerning the infants' age (see Table 5.3, row 12; $\varrho = -0.17$, $df = 62$, $p = 0.174$). The number of motion peaks shows a significant trend over age (see Table 5.3, row 9; $\varrho = -0.28$, $df = 62$, $p = 0.025$). The number of motion peaks tends to fall over age, which is consistent with their lower number in AAI than in ACI. Note, that this result refers to the total number of motion peaks within an action demonstration.

Regarding the average length of utterances, a significant difference between ACI and AAI can be observed (see Table 5.2, row 13). In ACI utterances are shorter compared to AAI. Furthermore, a trend over the infants' age was found (see Table 5.3 row 14; $\varrho = 0.31$, $df = 62$, $p = 0.014$). Utterances become longer with increasing age which is consistent with the previous result. Furthermore, pauses exhibit a significant difference in their average length (see Table 5.2, row 17). Pauses are longer in ACI compared to AAI suggesting they become shorter during development. The correlation of the average pause length with the infants' age is also significant and confirms this hypothesis (see Table 5.3 row 18; $\varrho = -0.30$, $df = 62$, $p = 0.015$). In the visual modality the average length of motion peaks is significantly different between ACI and AAI (see Table 5.2, row 10). However, there is no significant trend over age (see Table 5.3, row 11; $\varrho = -0.11$, $df = 62$, $p = 0.392$).

In summary, individual modalities exhibit strong structural differences between ACI and AAI both in the number of events segmented and the length of these events. A developmental trend can be shown both for the number of utterances and motion peaks. Concerning the length of visual and acoustic events a developmental trend is only visible for the length of utterances, but not for the length of motion peaks.

5.3.4. Results on the Number of Acoustic Packages per Interaction

The hypothesis evaluated here is that more acoustic packages per participant are segmented in child-directed interaction compared to adult-directed interaction. The aim of this analysis is to explore the fine-tuning in the parental use of acoustic packaging as a teaching strategy. For this, the previous analyses (see Section 5.2) are extended to all ages of the adult-child and adult-adult interaction corpus (see Table 5.3 for the results). Wilcoxon Mann-Whitney rank sum tests on the number of acoustic packages between AAI and ACI for each group have shown that with the exception of group 2a (12 to 17 month-olds), more acoustic packaging can be found in a child-directed interaction than in an interaction with another adult.

Furthermore, a Kruskal-Wallis chi-squared test ($H = 13.81$, $p = 0.003$) was conducted suggesting that significant differences in the number of acoustic packages can be found between age groups. Since differences between groups seem to be significant, a monotonic relationship between age and the number of acoustic packages was assumed. However, according to spearman's test a negative correlation of children's age with respect to the total number of acoustic packages can not be shown with a high level of significance

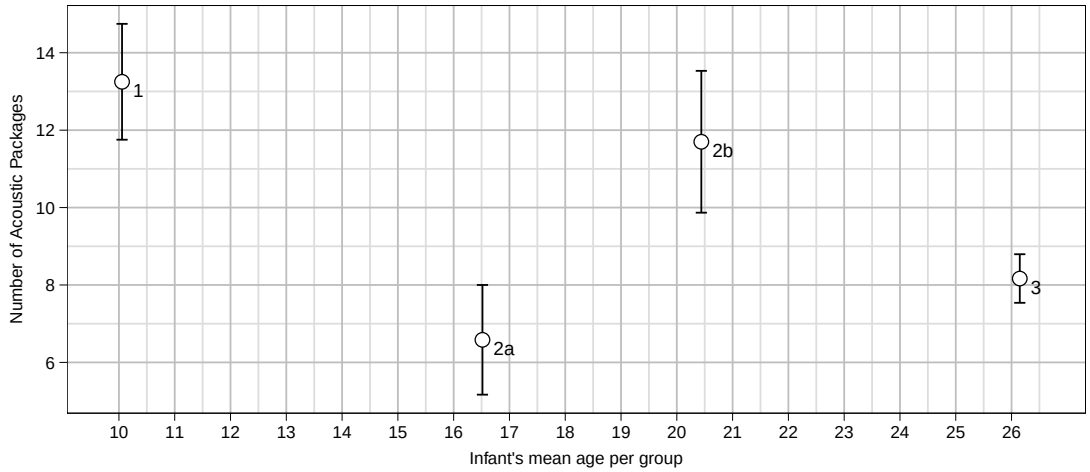


Figure 5.4.: Plot of the average number of acoustic packages per participant for each age group. Error bars display the standard error.

($\varrho = -0.22$, $df = 62$, $p = 0.084$). This weakly suggests that less acoustic packaging in the interaction can be found with growing age of children. A reason for this result could be the results of group 2a which deviate from this trend (see Figure 5.4).

Why do the results in group 2a deviate from the results in other groups? An explanation for this is motivated by observations on the videos in group 2a. Group 2a involves children that are 12 to 17 month-old and at this age, most of the children learn to walk. It could be that this locomotor task changes the interaction with the child as the child moves around. In the literature, such changes in motor skills have already been described as changing the social interaction (Bertenthal and Campos, 1990). Taking this into consideration, the results can be interpreted as follows: Acoustic packaging is present in interactions with younger and older children. It seems to be a teaching strategy that persists even though children's linguistic and cognitive capabilities increase.

5.3.5. Results on the Amount of Motion Peaks per Acoustic Package

The first hypothesis evaluated in the following assumes that the amount of motion peaks per acoustic package will be greater in adult-directed interaction than in child-directed interaction. It addresses the fact that acoustic packages are formed with less content (less motion peaks) when addressing children. In contrast to the previous previous analyses in Section 5.2, which considers one age group, here all age groups will be considered. The results for all age groups of the adult-adult and adult-child interaction corpus are provided in Table 5.2. For all age groups together, adults perform more motion peaks per acoustic package when interacting with a child than in an interaction with another

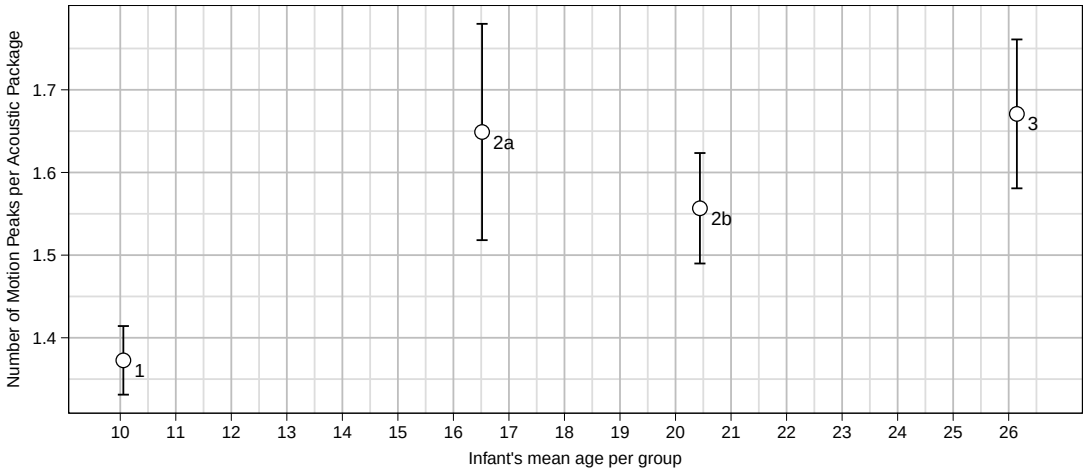


Figure 5.5.: Plot of the average number of motion peaks per acoustic package for each age group. Error bars display the standard error.

adult, which supports the initial hypothesis. These differences between ACI and AAI are significant according to a one-tailed Wilcoxon Mann-Whitney rank sum test, except for group 2a. This result is consistent with the results reported in the previous section.

The last results raise the question if there is a developmental course regarding the number of motion peaks per acoustic package. Therefore, the second hypotheses assumes there is an increase of number of motion peaks per acoustic package with respect to children's age. According to a Kruskal Wallis test, the number of motion peaks per acoustic package differs significantly between age groups ($H = 8.96$, $p = 0.03$). Furthermore, the hypothesis of a monotonic relationship between age and the number of motion peaks per acoustic packages could be confirmed: A significant correlation between the number of motion peaks per acoustic package and the age of children was found ($\varrho = 0.29$, $df = 62$, $p = 0.019$). In sum, acoustic packaging seems to be a teaching strategy that is used towards children of different age groups as more acoustic packages with a lesser content are formulated to children than to adults. Furthermore, it seems that the strategy is adaptive and converges towards adult-adult interaction patterns as infants become older.

5.3.6. Discussion

Multimodal recordings of adult-adult and adult-child interaction with infants from four different age groups ranging from 9 to 30 months were processed by the acoustic packaging system. In the previous sections the resulting data was analyzed concerning differences between adult-child (ACI) and adult-adult interaction (AAI) as well as possible

developmental trends. Specifically, statistics on individual modalities that contribute to acoustic packaging, statistics on the number of acoustic packages segmented, and statistics on the number of motion peaks per acoustic package were reported.

The analysis of the acoustic modality revealed that in AAI less utterances can be found compared to ACI, which is in line with previous findings (Fischer et al., 2011). Additionally, their length is smaller in ACI than in AAI. Concerning the visual modality the number of motion peaks is larger in ACI compared to AAI including their length. A developmental trend was shown for the average length of utterances that increase with age and the total number of motion peaks, which decrease with age. Results focusing on acoustic packages exhibit a significant difference in number between the ACI and AAI condition. Namely, more acoustic packages were found in ACI compared to AAI. However, no significant developmental trend could be shown but there is a tendency towards a decreasing number of acoustic packages with increasing age of the children. Another structural property of acoustic packages is represented by the number of motion peaks per acoustic package. Here, less motion peaks per acoustic package were found in ACI compared to AAI. Furthermore, with rising age the number of motion peaks per acoustic package increases.

The results show that acoustic packaging is able to statistically reflect structural differences between adult-child and adult-adult interaction. Furthermore, it was shown that changes in the interaction within the course of development are manifested in properties of acoustic packages. The analysis of individual modalities revealed that both utterances and motion peaks contribute to these differences. This is in line with findings showing that child-directed communication manifests itself visually and acoustically (Brand et al., 2002). The resulting packages thus provide an integrated measure that is able to distinguish ACI from AAI in different age groups. Furthermore, these measures can directly be derived from the information that is already present in each acoustic package which allows for an initial rating of its level of structuring. Therefore, a robotic system that uses acoustic packaging as an initial segmentation process for analyzing human action demonstrations may exploit this information to focus on information which is appropriate for the system's state of development. This idea is further supported by findings suggesting that both the acoustic and visual modifications in child directed action demonstrations are beneficial for learning actions and language (Brand and Shallcross, 2008; Ma et al., 2011).

One question that might arise is how selective acoustic packaging is compared to the segmentation of individual modalities. Acoustic packaging is a bottom up process, thus its aim does not lie in strong filtering of the multimodal input, but in its segmentation. Still, comparing the number of modality specific segments with the segments that were associated to acoustic packages reveals small differences. For motion peaks it makes sense that the number associated to acoustic packages is lower than for all motion peaks, since not all motion peaks overlap with acoustic segments, and thus, do not form an acoustic package (see Table 5.2, rows 5 and 8). The acoustic segmentation also exhibits this difference (see Table 5.2, rows 2 and 11). Here, the reason is that each valid acoustic package contains one utterance. Thus, the number of acoustic packages is slightly lower

than the number of utterances, since in case multiple short utterances overlap with the same motion peak, only the one with the longest overlap forms an acoustic package. This is more likely the case if there is a higher density of multiple short utterances. Thus, utterances associated to acoustic packages tend to be slightly longer in average than all utterances (see Table 5.2, rows 13 and 14). Due to its low impact on the results this behavior is currently neglected but might require modification in a system with further developed capabilities. At this point acoustic packaging does not provide further means to filter specific information from multimodal action demonstrations as, for example, differentiating gestures from motions that manipulate objects. This issue will be addressed in the Chapter 6.

Another question concerns the impact of interaction length on the present results. In ACI parents tend to repeat their demonstrations towards the child but not in AAI. On the one hand, one can argue that interaction length and repetitions are part of ACI-AAI differences. On the other hand, statistics on the structure of acoustic packages as the number of motion peaks per acoustic package are not affected by the interaction length (see Section 5.3.5). Accordingly, the evaluation in the following section describes results on task demonstrations limited to one presentation.

Until this point only human interaction has been analyzed. Interaction between humans and robots might exhibit different characteristics, which could reduce the usefulness of acoustic packaging in these systems. This question will be addressed in the following section.

5.4. Analysis of Human Robot Interaction

This section addresses the question how communication between humans and robots in a tutoring situation is structured, and how it relates to adult-adult and adult-child interaction. The results help to decide whether it is possible to use acoustic packaging as an analysis tool for other types of interaction such as human-robot interaction. For this purpose, adult-adult and adult-child interaction will be compared to adult-robot interaction to examine how acoustic packages reflect the characteristics of this type of interaction. For the adult-robot interaction, a simulated robot was used that reacted to the environment using a saliency based attention model (Nagai et al., 2008). Portions of this section were previously published by the author (Schillingmann et al., 2009a).

5.4.1. Corpus Overview

The data used in this evaluation consists of two corpora. One is a corpus containing video and audio data with adult- and infant-directed interactions (Rohlfing et al., 2006). This corpus has also been used for the analyses in the previous sections. From this corpus, 26

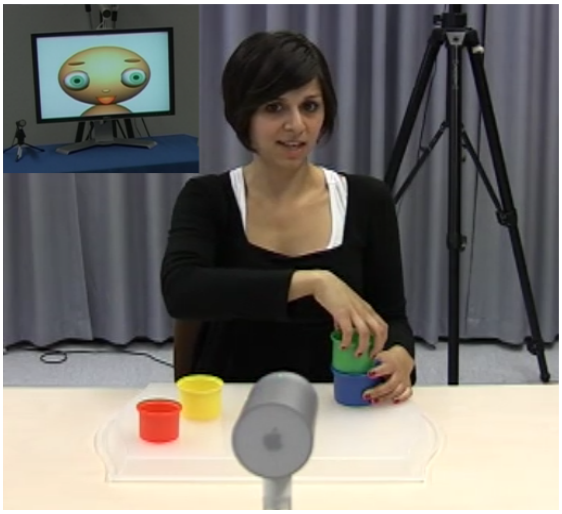


Figure 5.6.: An adult demonstrates how to stack cups to a robot simulation. The robot simulation is shown in the top left for illustration purposes.

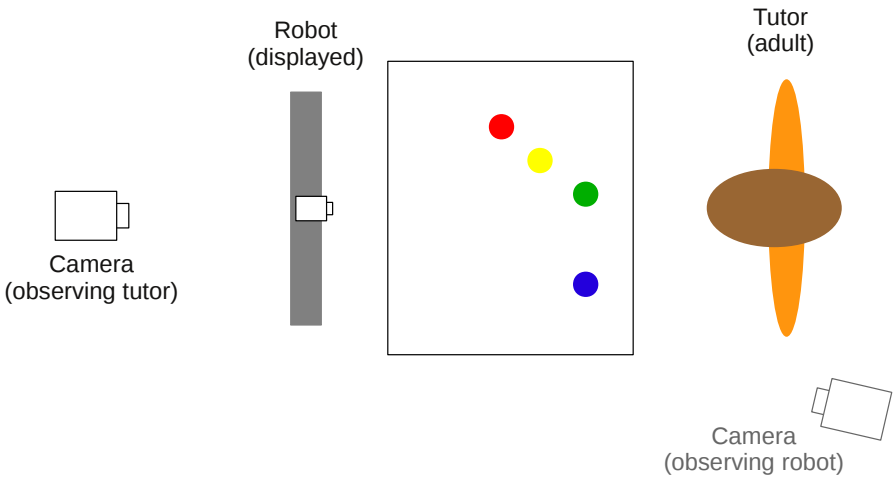


Figure 5.7.: Adult-robot interaction setting. The participant is facing the robot simulation, which is displayed on a screen. In this evaluation, recordings from the camera observing the tutor are used.

participants with 8 to 11 month-old children were selected. This corresponds to the first age group in the previous analysis (see Section 5.3.1). An overview of the experimental setting can be found in Section 5.2.1.

The second corpus (Vollmer et al., 2009) contains 31 German interactions between human participants and a simulated robot using the same tasks as in the corpus with adult-adult (AAI) and adult-child interactions (AAI). The view from the robots perspective is illustrated in Figure 5.6. An overview of the experimental setting is displayed in Figure 5.7. 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 (Nagai et al., 2008). Thus, the eyes focus on salient points, like moving or colorful objects. Here, 25 participants who performed the stacking cups task comparable to the participants in the corpus with ACI and AAI were chosen.

For better comparison of the actions across participants and corpora, a single task presentation was extracted from each video. The criterion for the extraction interval was defined as two seconds before the first cup was lifted until two seconds after the last cup has been stacked by the participant. This method was applied to both copora used in this evaluation.

5.4.2. Procedure and Design

The acoustic packaging system was used to segment the data as in the previous evaluation described in Section 5.3.2. One difference is that for single task demonstrations the acoustic quality was acceptable and thus speech was segmented automatically according to the initial system design (see Section 4.3.3). The same measurements as in the previous evaluation were calculated on the acoustic packages (see Table 5.4). Thus, each row contains measurements calculated on the three types of interaction available from the corpora, namely adult-adult (AAI), adult-child (ACI) and adult-robot (ARI) interaction. The results are averaged over the number of participants for each interaction type. An asymptotic Wilcoxon Mann-Whitney rank sum test was performed to assess which measurements show significant differences between the interaction types.

In the following, acoustic packaging results will be analyzed with focus on the differences and similarities of AAI and ACI towards ARI. The results in Table 5.4 will be analyzed according to the groups described in Section 5.3.2. First, results concerning the individual modalities are presented. Second, acoustic packages and their structural properties will be reviewed.

		ARI	ACI	AAI	ACI-AAI		ACI-ARI		AAI-ARI	
		M (SD)	M (SD)	M (SD)	Z	p	Z	p	Z	p
1	Number of participants	25	23	23						
2	Total number of APs	5.40(3.11)	4.35(2.10)	2.48(0.95)	3.4	0.00	−1.1	0.29	−3.8	0.00
3	Total length of APs [s]	12.98(6.08)	10.22(4.66)	7.07(2.04)	2.4	0.02	−1.7	0.09	−3.8	0.00
4	Average length of APs [s]	2.59(0.63)	2.45(0.77)	3.19(1.35)	−2.1	0.03	−1.1	0.27	1.4	0.15
5	Total number of MPs (in APs)	9.80(4.37)	8.22(3.70)	7.00(2.20)	0.9	0.38	−1.3	0.19	−2.5	0.01
6	Total length of MPs (in APs) [s]	11.81(5.52)	9.62(4.52)	6.43(1.95)	2.7	0.01	−1.5	0.15	−3.9	0.00
7	Average length of MPs (in APs) [s]	1.21(0.24)	1.17(0.27)	0.95(0.18)	3.1	0.00	−0.6	0.57	−3.7	0.00
8	Total number of MPs	11.80(4.71)	11.65(4.28)	7.74(1.89)	3.6	0.00	−0.0	0.97	−3.7	0.00
9	Total length of MPs [s]	13.66(5.39)	13.03(4.59)	6.94(1.67)	4.9	0.00	−0.2	0.85	−5.2	0.00
10	Average length of MPs [s]	1.17(0.23)	1.14(0.23)	0.92(0.18)	3.6	0.00	−0.5	0.63	−4.1	0.00
11	Total number of utterances	6.04(3.46)	4.65(2.29)	2.65(0.98)	3.2	0.00	−1.4	0.17	−4.1	0.00
12	Total length of utterances [s]	8.57(4.09)	6.15(2.88)	5.51(1.78)	0.7	0.47	−2.0	0.04	−3.0	0.00
13	Average utterance length [s]	1.53(0.49)	1.43(0.68)	2.38(1.29)	−3.4	0.00	−1.3	0.18	2.8	0.01
14	Average utterance length (in APs) [s]	1.68(0.61)	1.51(0.76)	2.54(1.38)	−3.4	0.00	−1.5	0.15	2.5	0.01
15	Total number of pauses in speech	5.04(3.46)	3.65(2.29)	1.65(0.98)	3.2	0.00	−1.4	0.17	−4.1	0.00
16	Total length of pauses in speech [s]	5.13(3.15)	4.86(3.56)	1.51(1.13)	3.8	0.00	−0.6	0.52	−4.7	0.00
17	Average length of pauses in speech [s]	1.22(0.92)	1.31(0.76)	0.87(0.58)	2.2	0.03	1.5	0.14	−1.5	0.13
18	Average number of MPs per AP	2.04(0.68)	2.04(0.76)	3.06(1.06)	−3.6	0.00	−0.1	0.93	3.6	0.00
19	Ratio of interaction length to speech length	2.08(1.30)	3.70(6.55)	1.39(0.37)	3.9	0.00	1.7	0.10	−3.3	0.00
20	Ratio of AP length to speech length (in APs)	1.62(0.29)	1.76(0.33)	1.33(0.21)	4.3	0.00	1.8	0.08	−3.8	0.00
21	Ratio of AP count to speech length (in APs) 1/[s]	0.65(0.22)	0.79(0.33)	0.48(0.21)	3.3	0.00	1.5	0.13	−2.4	0.02
22	Ratio of all MPs to MPs assigned to APs	1.35(0.67)	1.84(1.92)	1.16(0.36)	2.5	0.01	1.3	0.21	−1.5	0.13
23	Ratio of interaction length to AP length	1.29(0.64)	1.92(2.75)	1.05(0.19)	3.1	0.00	1.0	0.30	−2.7	0.01

Table 5.4.: Acoustic packaging statistics calculated on results of adult-robot (ARI), adult-child (ACI) and adult-adult (AAI) interaction (AP: Acoustic Package; MP: Motion Peak; Z, p: Results of asymptotic Wilcoxon Mann-Whitney rank sum tests).

5.4.3. Results on Individual Modalities

To simplify the comparison of the interaction types $>$ and $<$ will be used to indicate a significant difference in the corresponding measurement while $\gtrsim$ and $\lesssim$ will be used in case there is a tendency but no significant test result. In the subsequent paragraph, row numbers refer to the results in Table 5.4.

The analyses revealed a significant difference in the length of utterances (see row 12: ARI $>$ ACI $>$ AAI). The number of utterances is significantly different between ARI and AAI but not between ACI and ARI (see row 11: ARI $\gtrsim$ ACI $>$ AAI). This suggests that the verbosity tends to be higher in ARI than in ACI and is significantly higher than in AAI while the structure tends to be similar. However, the average length of utterances was found not to be significantly different between ACI and ARI although a tendency can be assumed (see row 13: ARI $\gtrsim$ ACI). The average length of pauses was found not to be significantly different between ACI and ARI but between AAI and ARI (see row 17: ARI $\gtrsim$ ACI $>$ AAI). However, a tendency was found that pauses are longer in ARI than in ACI, which is in line with findings on foreigner-directed speech (Biersack et al., 2005). In foreigner-directed speech participants tend to lengthen pauses. Concerning the visual modality, the analysis of the number of motion peaks and their average length provided similar results. Here, the number and length shows a significant difference between ARI and AAI but no significant difference to ACI (see rows 8 and 9: ARI $>$ AAI).

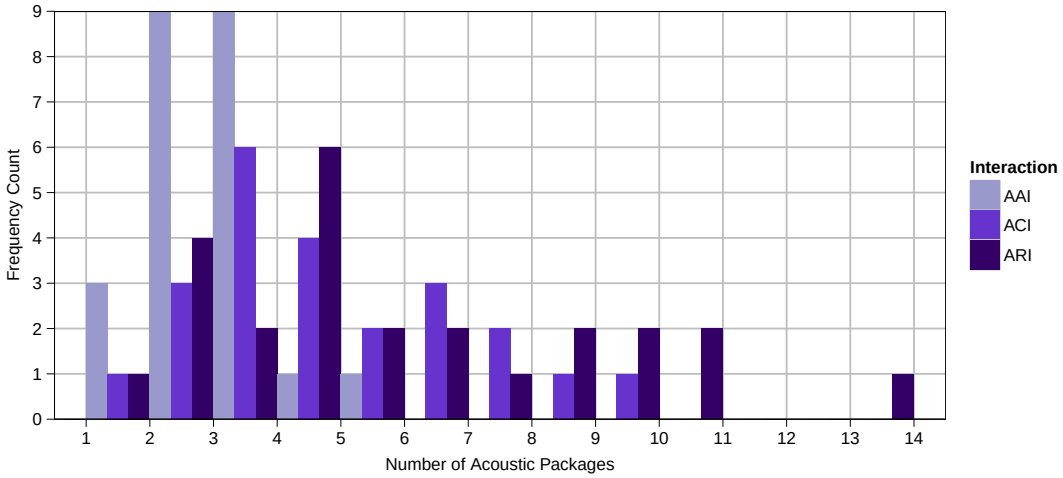


Figure 5.8.: Combined histogram over the number of acoustic packages per presentation for adult-adult (AAI), adult-child (ACI), and adult-robot (ARI) interaction. The histogram was created using a bin width of one.

In sum, results on the segmentation suggest differences between ARI and AAI. Significant differences between ARI and ACI are limited to the total utterance length per presentation. Together the results suggest a similarity between ARI and ACI with a higher verbosity in ARI.

5.4.4. Results on the Number and Total Length of Acoustic Packages

As described in Section 5.4.1 in this evaluation, one trial corresponds to one presentation of the stacking cups task. Thus, a significant difference in the total number of acoustic packages per presentation can be shown (see Table 5.4, row 2: ARI > AAI). However, ARI and ACI do not show a significant difference in this regard, since their distributions are similar (see Figure 5.8). Furthermore, the total length of acoustic packages exhibits a significant difference between ACI, ARI, and AAI (see row 3: ARI > ACI > AAI). The hypothesis ARI > ACI is not rejected by a one-tailed Wilcoxon Mann-Whitney rank sum test ($Z = -1.68$, $p = 0.046$), which suggests that segments of tutoring in ARI are in general longer than in ACI but consist of a similar structure. The latter is also reflected in the average length of acoustic packages, which exhibit no significant difference (see row 4). Therefore, a "unit" in the interactions seems to be temporally the same regardless of the interaction type.

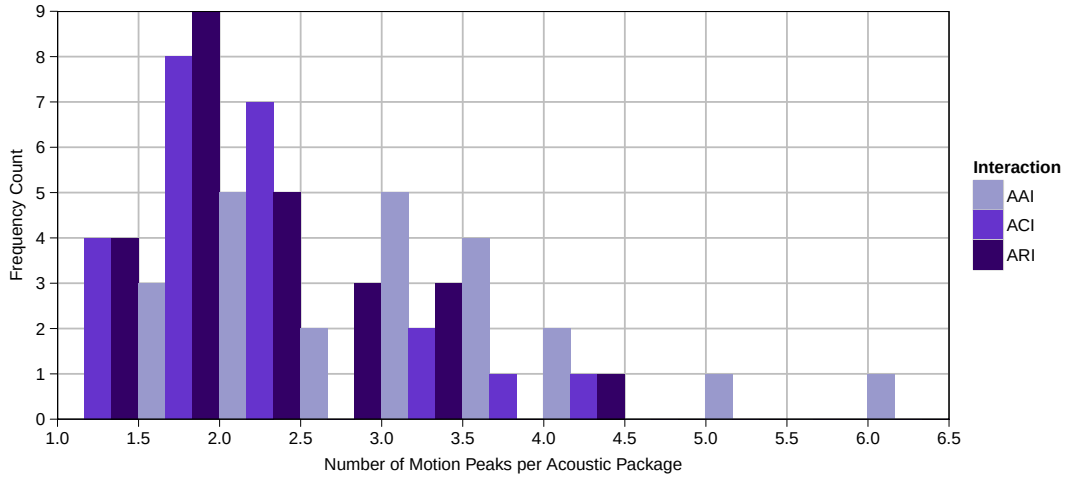


Figure 5.9.: Combined histogram over the number of motion peaks per acoustic package for adult-adult (AAI), adult-child (ACI), and adult-robot (ARI) interaction. The histogram was created using a bin width of 0.5.

5.4.5. Results on the Amount of Motion Peaks per Acoustic Package

Looking at the average number of motion peaks per acoustic package (see Table 5.4, row 18), the results show a significant difference between ACI-AAI and ARI-AAI (ACI < AAI). However, there is no difference between ACI and ARI which is also confirmed by the distribution of the number of motion peaks per acoustic package (see Figure 5.9). 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 and confirms the previous evaluation in Section 5.3 for single presentations. The results reveal that structuring in ARI is on the same level as in ACI. The ratio in row 20 suggests that acoustic packages in ARI have a higher proportion of speech compared to ACI due to the increased verbosity in ARI.

5.4.6. Discussion

The acoustic packaging system was used 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 (Brand and Tapscott, 2007). The previous evaluations (see Section 5.2 and 5.3) showed that the acoustic packaging model is able to reflect the structural differences between tutoring in adult-adult and adult-child interactions. Additionally, adult-robot interaction

was analyzed in comparison with adult-adult and adult-infant interaction. According to the analysis of acoustic packages, 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, participants 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 including the tendency to lengthen pauses. The differences between ACI and ARI regarding verbosity and pauses are supported by findings on manual annotated motion features on the ARI corpus, in which ARI exhibited longer pauses compared to ACI and AAI (Vollmer et al., 2009). They were attributed to the limited feedback the robot simulation is able to provide, causing participants to wait for the system's response. Another reason might lie in the unfamiliarity of the studies' participants with the robot interaction partner, causing similar effects as in foreigner directed speech (Biersack et al., 2005).

The data in this evaluation 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 the goal to develop strategies that enable robots to react to and learn from tutoring situations. One aspect is to detect the presence of tutoring behavior to select chunks of multimodal input which facilitate learning.

In summary, participants exhibited a similar tutoring behavior to children as to robots. However, differences between ARI and ACI suggest that a robot learning from multimodal action demonstrations should provide appropriate feedback.

5.5. Conclusion

In this chapter, three evaluations have been described. The first evaluation provided a pairwise comparison between adult-child (ACI) and adult-adult (AAI) interaction showing that acoustic packaging is able to reflect differences between both interaction types. The second evaluation aimed at identifying properties of acoustic packages that change in the course of infant development. Furthermore, the results confirmed the differences towards AAI for all age groups. This suggests properties, such as the number of acoustic packages in interactions and the number of motion peaks per acoustic package, as measurements for structure in tutoring interaction. The third evaluation focused on the differences and similarities of adult-robot interaction (ARI) towards ACI and AAI. The results show that ARI has very similar properties compared to ACI, but participants tend to exhibit a higher verbosity, which is reflected in the length of acoustic packages.

In general, the results support acoustic packaging as a method for segmenting action demonstrations in tutoring situations and assessing their structure to determine the presence of tutoring behavior in interaction with humans as well as robots. Acoustic packaging describes temporal properties of segments in a multimodal sensory stream.

In the present version, global properties, such as speech activity and motion, are part of this representation. However, more local properties such as emphasis in speech are not included. Furthermore, spatial properties as, for example, object locations are not represented, either. In the present form, acoustic packaging is able to find learning units in human robot interaction, but it is not further specified how these units can be used for learning actions or properties of the ongoing action. Extracting more detailed information on the ongoing interaction is not only important for learning but also for providing feedback, since it needs to convey parts of the robot's understanding of the interaction to the human tutor. Therefore, additional cues addressing these issues were integrated into the acoustic packaging system. The corresponding modules are described in the following chapter.

6. Acoustic Packaging as a Basis for Feedback on the iCub Robot

This chapter describes modules which extend the acoustic packaging system to run acoustic packaging on the iCub robot (Metta et al., 2010). The focus is to acquire additional detailed information about the ongoing interaction that allows a robot to provide feedback to a human interaction partner. Thus, features beyond solely temporal segmentation are required by the system to pick up contents of the multimodal action presentations it perceives. These features will be provided by two additional modules that were integrated to the acoustic packaging system. The first is a color saliency based tracking module which allows tracking and feature extraction of moving colored regions typical for child toys. The second is a prominence detection module which allows to identify emphasized syllables in the tutor's speech. Both cues are linked by acoustic packages and therefore associate visual and acoustic information, which can be used to respond to a human interaction partner. This was tested by integrating the acoustic packaging system on the iCub robot. For example, if the human emphasizes a color term while moving a cup with the corresponding color, the system is able to exploit this information for feedback by replaying the emphasized syllable when the cup is moved again. A plausibility test of this form of feedback was performed by analyzing synchrony of object characteristics and stressed words in acoustic packages using a corpus (Rohlfing et al., 2006) with adult-child interactions in a tutoring situation.

6.1. Color Saliency Based Tracking

In general, saliency models try to model visual attention by identifying outstanding regions in the input image by calculating saliency maps (Itti and Koch, 2001). Saliency maps are mainly discussed on static scenes where the sequence of focal points a saliency model provides is evaluated in comparison to biological vision. The main differences between saliency models depend on the features they use and how they are combined to saliency maps. For example, the model described by Itti et al. (1998) uses several feature maps such as color, intensity, and orientation to calculate a bottom-up saliency map. Modifications of this approach include maps with additional temporal features such as changes of motion and brightness to support the processing of video data instead of static scenes (Nagai et al., 2008). However, the saliency maps resulting from these

models tend to have a much lower resolution than the input image, and thus, they do not define exact regions of interest. Furthermore, a saliency model applied to human robot interaction in a tutoring scenario should support tracking regions over time, but the models described above do not provide additional tracking functionality. One approach to this problem is to tune the saliency map in an initialization step towards features which should be tracked (Frintrop and Kessel, 2009). But then, the method depends on manual initialization and might not be able to automatically recover from tracking errors. Another solution is to identify salient regions, compare them, and track them over time to acquire object trajectories (Zhang and Stentiford, 2008). However, the distance measure used for tracking is task specific. Additionally, interactive systems require the saliency module to be capable of real-time processing images since otherwise it is impossible to provide feedback in time. Saliency algorithms which combine multiple independent feature maps can still create high computational demands even on recent systems. Therefore, the color saliency module used in this work is inspired by color preference in infants and uses properties of a task demonstration to efficiently track color salient regions.

6.1.1. Color Vision in Infants

The color saliency module in the acoustic packaging system will process video data from tutoring situations where objects — typically colored toys — are presented to the system. Thus, the saliency module will be tuned towards this situation. Findings on color perception in infants indicate that children prefer chromatic colors as present in toys, since saturation but not brightness controls infants looking preferences among chromatic stimuli (Teller et al., 2004). Consistently, infants prefer brown and grey significantly less than basic colors (Pitchford and Mullen, 2005). Infants' looking behavior does not seem to be static but is influenced by surrounding colors. Results by Pereverzeva and Teller (2004) indicate that infants' looking preferences are influenced by colors which differ maximally in purity from surrounding colors. Both aspects will be integrated in this module.

6.1.2. Design Rationale and Requirements

The design of the color saliency module will be tuned towards tutoring scenarios that involve the presentation of colored toys toward an infant or a robot. The reason behind choosing a saliency based approach is to avoid including a high amount of task specific knowledge into the system. Using, for example, a pre-trained object tracker would make the system inflexible towards new objects. Even if these constraints are accepted the system could be hard to use in varying light conditions, since they might require retraining the object tracker. Still, some assumptions on the scenario are included into the module. Since the module should process data from action demonstrations, the saliency module exploits this information and will be limited to parts of the visual input which changes

over time. Thus, sensitivity to motion is a central aspect of the method developed here, because it allows to optimize the amount of information processed from the visual input. This is beneficial for designing a color saliency module that needs to be able to process the input data online.

Rather than requiring any manual initialization, the tracking method should identify relevant information by using motion cues provided by the tutor. Furthermore, it is sensible to exploit color information in this type of tutoring scenario where colored toys are present. As discussed in the previous section, the design of color based attention should include sensitivity to chromatic colors and incorporate surrounding color information. All processing steps require an efficient implementation that allows for online processing video data in adequate resolutions as, for example, 640x480 pixels. One reason is that objects can be relatively small in the camera image depending on the distance. Another problem is the relatively high possible velocity of human hand motions in tutoring situations. Simple tracking methods which depend on regions that overlap from one frame to the next are not reliable in this case, since their spatial distance can be higher than their dimension. In cases of fast movements, this occurs even for framerates between 20 and 30 Hz. This aspect needs to be respected in the tracker's implementation.

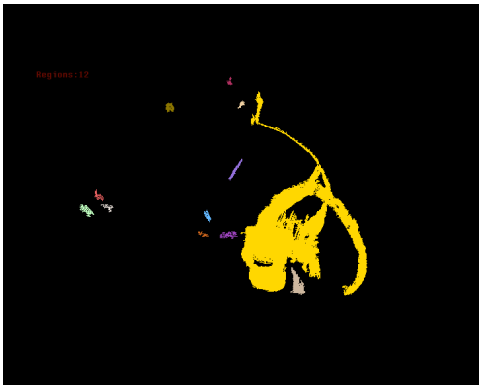
6.1.3. The Color Saliency Based Tracking Module

The main task of the color saliency based tracking module is to provide features on salient moving regions in the visual input of the system. The module provides trajectories including color properties of the moving regions to the acoustic packaging system. The approach is based on the assumptions that during action demonstrations the objects are typically moved, and uniformly colored objects are used. For the implementation, the same framework by Lömker et al. (2006) as described in Section 4.3.4 is used. The module performs several processing steps starting with the input image and ending with trajectory estimation which will be described in the following.

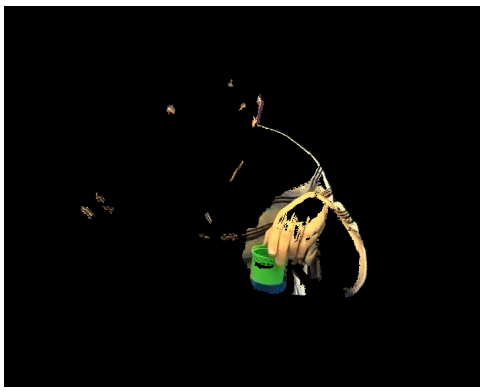
Detecting Changing Regions

First, temporal changes in the visual input are detected using an approach based on motion history images (Davis and Bobick, 1997). Motion history images are reused, since they are already calculated for measuring the amount of motion within the initial acoustic packaging system (see Section 4.3.4). Here, the difference is that motion history images will be used to locate and separate changes in the visual input instead of only measuring the amount of change. An example of a motion history image is displayed in Figure 6.1a.

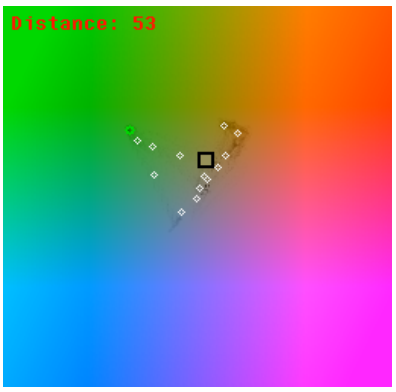
(a) Motion history image.



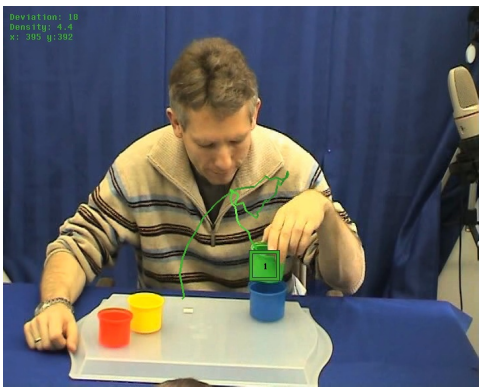
(b) Labeled motion history image (Each label is highlighted in a different color for reasons of discernibility).



(c) Original image masked by the motion history image.



(d) Projection of masked pixels into the YUV color space.



(e) Original image with overlaid trajectory and top ranked salient region.

Figure 6.1.: Snapshots of processing steps within the color saliency tracking module.

Masking

Once changing regions are available, the corresponding parts of the input image need to be selected. Therefore, the motion history image and the input image are combined leaving only those pixels that match areas with non-decayed history. Since the history extents over a certain time $t_{\max}$, an input image delayed by $t_{\max}/2$ is used (see Figure 6.1c). This approach ensures that actual moving parts are in the center of the history image, which allows for better segmentation of homogeneous areas in the input.

Labeling

The previous step selects changing parts in the input image but does not identify which pixels belong to the same region or a region which is spatially not connected. Therefore, a labeling algorithm is used to identify connected regions (Soille, 2002). Additionally, this method allows for suppressing noise by rejecting regions that do not contain a sufficient amount of pixels. An example with labels highlighted in different colors is displayed in Figure 6.1b.

Projection into a UV Color Histogram

In this step, the result of the previous masking operation is projected into a UV color histogram. The YUV color space was chosen since it roughly approximates human color perception. The idea is to prepare for efficient color clustering. Using a histogram has the advantage that the distance function of the clustering algorithm only needs to be calculated once for each pair of histogram bins processed. Otherwise, every pair of pixels needs to be considered separately. In addition to U and V, the region label provided by the labeling step is used as a third key which separates colors by spatially different regions. Furthermore, each histogram bin does not only contain the number of pixels accumulated, but also their coordinates which allows for backprojection. In summary, all changing regions are represented within a three dimensional histogram using the region label as well as U and V as indices. Figure 6.1d shows a UV representation of the histogram where darker areas represent a larger bin size.

Clustering

The histogram bins are now clustered using the k-means algorithm (MacQueen, 1967). Clustering is performed separately for each region label. The Euclidean distance in the UV space is used to compare color values. Figure 6.1d shows the UV space where cluster means are displayed as white circles. The black square shows the centroid of all clusters, which is required for the next processing step. After the histogram bins are clustered by comparing their colors, a merging step is performed where similar clusters are combined to a single cluster. In contrast to the previous clustering method, here not only color

information but also each clusters mean position in the input image is exploited. The mean position is determined based on the pixel coordinates associated to the histogram bins (backprojection). In this step, the region label is not further maintained, since spatial distance is part of the comparison. The idea behind this step is to merge neighboring clusters with similar colors, but to avoid clustering colors from separately moving regions.

Ranking

The ranking algorithm has a key role in the color saliency module, since it decides about each cluster's level of saliency. For this purpose, all color clusters are sorted according to their distance to the centroid of all clusters. The cluster with the largest distance to the global centroid is the most salient cluster. Figure 6.1d displays the most salient cluster as a green circle. In this processing step, two aspects concerning infant color vision are realized (see Section 6.1.1). First, the position of the global centroid — and thus, the ranking — depends on the surrounding colors, since it averages over the area that is uncovered by the motion history image. Second, chromatic colors are preferred, since clusters with high distance to the centroid are likely in the outer areas of the UV plane.

Heuristic Filtering

Several heuristic filtering steps are applied to the ranked clusters. The first filter removes clusters which are below a saliency threshold, which means a cluster with no sufficient distance to the centroid. Typically, this is the case when only a hand or parts of a body are moving. Furthermore, each cluster's density is checked, which is implemented as the ratio of pixel count to the area they are distributed over. By rejecting clusters with low density, noise is suppressed. Very large clusters are removed by checking the standard deviation of the pixels belonging to that cluster. The most important step in this process is removing clusters that correspond to uncovered background. Especially, background that contains chromatic colors would lead to clusters which do not correspond to an object. To overcome this problem, seeded region growing is used to test each cluster. The idea is that if region growing would expand a cluster further than a threshold factor, it is likely that this cluster belongs to uncovered background. However, this method is limited to uniformly colored backgrounds. Depending on the lighting condition, highlighted areas with skin color such as hands are detected as salient, especially if no object is moved. Thus, a skin color detection algorithm implemented by Ingo Lütkebohle based on the method by Peer et al. (2003) was adopted to reject these clusters.

Trajectory Accumulation

The surviving clusters from the previous filtering steps are tracked over time to build a trajectory. The tracking algorithm first compares all new clusters to the existing trajectory hypotheses. If, according to a distance function, no existing trajectory matches

the cluster, a new trajectory is initialized. The distance function uses both the distance in color space as well as the spatial distance, which allows to track clusters even if they move quickly. If a newly initialized trajectory is shorter than a certain time frame or converges to another trajectory during this time frame, the trajectory is rejected. A trajectory is considered as complete if no new clusters are added within a certain time frame. The trajectories are ranked according to their mean saliency level. The output of the color saliency module is the most salient trajectory with a sufficient minimal length. An exemplary tracking result is depicted in Figure 6.1e. The trajectory accumulation and all previous steps work incrementally. For each frame the current trajectory hypotheses are updated and the most salient trajectory is made available to the other modules in the acoustic packaging system by inserting it into the Active Memory (see Section 4.3.2).

6.1.4. Evaluation

A small evaluation on the performance and the accuracy of the color saliency module was carried out. Timings were measured on a system using an Intel Core Duo CPU (3 GHz). On typical input video data with dimensions of 720x576 pixels the implementation requires 3–5 ms processing time per frame. Including the timings for the motion history images, which are required as a preprocessing step, results in 6–8 ms processing time. Thus, the approach is well capable of online processing video data even for frame rates greater than 30 Hz. To estimate the accuracy of the saliency module, a sample of 441 trajectories was generated by processing 15 videos showing a cup stacking task from a corpus with adult-adult and adult-child interactions (Rohlfing et al., 2006). Subsequently, the trajectories were overlaid with the underlying video data. The trajectories were categorized by a human coder into correct and incorrect trajectories. Since the evaluation focuses on false positives, all trajectories that did not origin from moving objects or did not have the correct color were classified as incorrect. It was found that 93.87% of the trajectories were classified as correct, although the light conditions in the evaluation data vary. The most common cause for the low level of false positives are small colored regions, as, for example, objects which were not subject to manipulation in that moment but they were covered and then uncovered by a hand or arm. In sum, the results show that the module is able to reliably track uniformly colored regions.

6.1.5. Summary

The color saliency module was developed and integrated into the acoustic packaging system to efficiently track colored objects such as toys used in action demonstrations towards infants and robots. The module provides positional information in form of trajectories and color features to the acoustic packaging system. The implementation does not require pretrained models and it does not need to be initialized before tracking. The module can be configured with several parameters, which have been mentioned in the previous sections. An overview of these parameters can be found in Table 6.1. One

Parameter Name	Value
Motion history size	10 frames (for 25 fps video input)
Labeling: minimal region size	50 pixels
k-means: starting points	10
k-means: iterations	4
k-means: minimal cluster size	50 pixels
Cluster merging: maximal cluster distance	8 pixels
Cluster merging: maximal color distance	2.2% of the maximal distance value
Heuristic filtering: saliency threshold	1.6% of max color distance
Heuristic filtering: maximal cluster deviation	50 pixels
Heuristic filtering: minimal cluster density	1% of the current clusters variance
Heuristic filtering: maximal region grow factor	4 times the initial number of pixels
Heuristic filtering: maximal region growing color distance	3.3% of the maximal distance value
Trajectory accumulation: maximal cluster distance	99 pixels
Trajectory accumulation: maximal color distance	6,6% of the maximal distance value
Trajectory accumulation: minimal trajectory length	500 ms

Table 6.1.: Values of relevant parameters for the color saliency tracking module in a typical configuration. For the skin color filter parameters see Peer et al. (2003).

can argue that these parameters encode a certain level of previous knowledge. However, this knowledge is very generic and, thus, does not need to be frequently adapted. An important factor is that the module operates in a wide range of light conditions without changing its parameters.

6.2. Prominence Detection

The role of the prominence detection module in the acoustic packaging system is to identify highlighted parts in the tutor's speech. This way, the system is able to perceive words or expressions for an action or another term the tutor has focused on. For example, if the tutor shows a cup and focuses on the cup's color in the tutoring situation, s/he will probably emphasize the color term. Portions of this section were previously published by the author (Schillingmann et al., 2011).

6.2.1. Perceptual Prominence

Perceptual prominence of linguistic units is defined as the unit's degree of standing out of its environment (Tamburini and Wagner, 2007). This definition results in the following aspects which need to be modeled. First, it is necessary to define which type of linguistic unit the module should operate on. Syllables are typically used in prominence detection methods (Tamburini and Wagner, 2007). They have the advantage that speech can be segmented into syllables without using models that require a known lexicon. Second, an environment has to be defined, which is analyzed to compare linguistic units. In this

work, syllables will be ranked on a per utterance basis, which is a common approach and integrates well with the speech segmentation the acoustic packaging system already performs. Third, features to rate the level of prominence of each unit and a method to rank the results are required. These features have to be chosen carefully according to their robustness in noisy acoustic environments. Additionally, possible features vary depending on the language. For German, a possible set of features consists of nucleus duration, spectral emphasis, pitch (F0) movements, and the overall intensity (Tamburini and Wagner, 2007).

However, for the scenario in this work, the set of features needs to be reduced. One reason is the relatively noisy environment of a tutoring situation. Especially if the module is used for analysis of adult-child interaction, the feature must be noise robust. This environment makes pitch estimation and nucleus duration features less reliable. A nucleus duration feature would additionally depend on the accuracy of the syllable segmentation. Furthermore, findings on prosodic event detection show that the difference between combining several features compared to single features is relatively small (Rosenberg et al., 2012). Tamburini and Wagner (2007) achieved optimal results using high weighting factors for spectral emphasis and nucleus duration features. Considering these results as well as the finding that in adult-child interaction prosodic features are more exaggerated (Brand et al., 2002), the prominence detection module will rely on spectral emphasis in its implementation.

6.2.2. The Prominence Detection Module

According to the model described in the previous section, the prominence detection module operates on utterance level. If a new utterance hypothesis is completed, the prominence detection module retrieves the acoustic signal from active memory and performs the subsequent steps. First, the speech stream is segmented into linguistic units, which in the present case are syllables. The second step rates these linguistic units according to the acoustic parameters which correlate to the perceived prominence. The result is a syllable segmentation which includes a prominence rating for each syllable. The utterance hypothesis is extended with this information and made available to other modules by inserting the updated hypothesis into the active memory. The syllable segmentation method and the implementation of prominence detection are described in more detail in the following.

Syllable Segmentation

A modified version of the Mermelstein algorithm (Mermelstein, 1975) is used to segment utterances into syllables. In a first step, the signal is filtered using an equal loudness filter (Robinson, 2011). The filtered signal is further bandpass filtered using a 4th order Butterworth bandpass filter with cut-off frequencies at 500 Hz and 4000 Hz. Then, the signal is full wave rectified and low-pass filtered with a second order Butterworth filter at

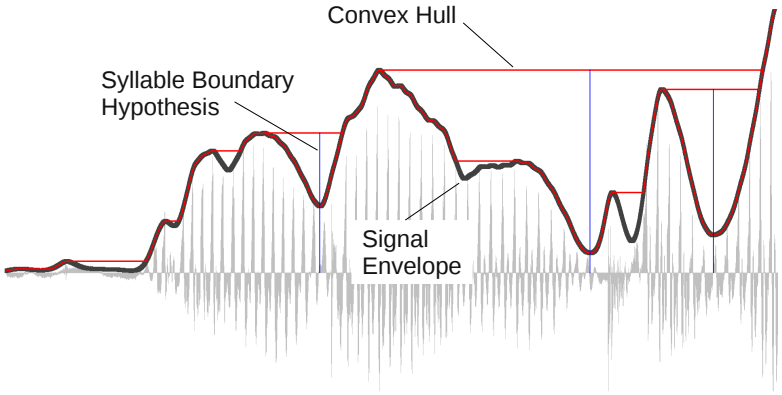


Figure 6.2.: Visualization of the Mermelstein convex hull based syllable segmentation algorithm. The convex hull is drawn at multiple iterations to visualize its approximation of the energy envelope.

Parameter Name	Value
Difference threshold between hull and envelope	1.39 dB
Minimal segment duration	80 ms

Table 6.2.: Values of relevant parameters for the syllable segmentation method in a typical configuration.

40 Hz to obtain an estimation of the signal's envelope. The basic idea of the Mermelstein algorithm is to detect minima in the signal's energy envelope. The locations of these minima are the desired syllable boundaries. The minima detection is described in the following: The signal's envelope is approximated using a convex hull. A syllable boundary is identified at the maximum difference between the convex hull and the signal's envelope (see Figure 6.2). The algorithm is carried out recursively for the intervals left and right to the syllable boundary. The recursion is terminated if the maximum distance drops below a certain threshold or the interval between two boundaries falls below a minimal length. Table 6.2 gives an overview of these parameters and common values. Values were determined by a parameter optimization on a subset of the Verbmobil corpus (Kohler et al., 1994). The general idea behind this approach is to prioritize the most significant minima in the signal's envelope.

Prominence Rating

As discussed in Section 6.2.1, spectral emphasis is used to rate the syllable segments. The syllable segment with the highest spectral emphasis rating is considered the most prominent syllable in the utterance. The spectral emphasis feature is calculated by

Matches	Deletions	Insertions
68.65%	31.35%	31.35%

Table 6.3.: Evaluation results of our syllable detection method on utterances from the Verbmobil corpus.

Matches	Utterances	*Words
59.71%	139	4.45

Table 6.4.: Evaluation results of prominence detection approaches on utterances from adult-infant interactions. The results are 2.7 times better than chance. (*Average number of words per utterance)

bandpass filtering the signal with a 4th order Butterworth filter in the band 500 Hz to 4000 Hz. Then, RMS energy is computed for each syllable segment and normalized per utterance.

6.2.3. Evaluation

Both the syllable segmentation approach and the prominence rating method were evaluated. Syllable segmentation was evaluated on a subset of the Verbmobil corpus (Kohler et al., 1994), since an accurate syllable segmentation is available. The subset consists of 2,000 randomly selected utterances containing 68,276 syllables in total. A syllable boundary is considered a match if a boundary hypothesis is within 50ms distance. Table 6.3 shows results with balanced insertion and deletion rates.

The prominence rating algorithm has been evaluated on a corpus with adult-infant interactions (Rohlfing et al., 2006). For the evaluation, a subset where adults explain children how to stack cups was used. The acoustic channel has been recorded from a distant microphone and thus contains environmental noise e.g. from the cup stacking task and in some cases from the child. Word boundaries were automatically determined from a transcription by performing a forced alignment. A human annotator has marked the most prominent word in each utterance. If the center of the syllable with the highest prominence ranking lies within the word boundaries, a match is counted. Utterances with very bad acoustic conditions where even the forced alignment failed were not taken into account. 139 utterances have been used in the evaluation. The results are presented in Table 6.4.

6.2.4. Summary

A prominence detection module was described including its integration in the acoustic packaging system, where it detects semantically relevant information linguistically highlighted by a tutor. Evaluation results on speech data from adult-infant interactions

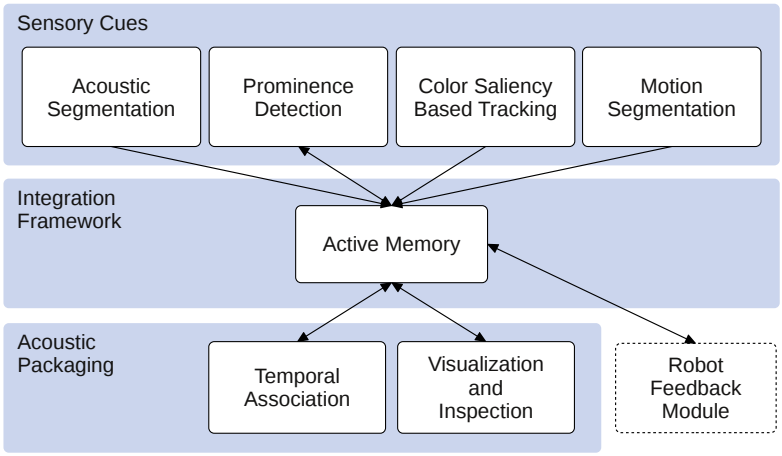


Figure 6.3.: System overview with highlighted layers and their relation to the acoustic packaging system.

show a 59.7% agreement with human raters. This means that through a fully automated approach of syllable segmentation and prominence detection more than half of the stressed words can be obtained. While this may seem a low recognition rate, it should be noted that the results were achieved on highly realistic data which include much noise from toy playing and children's interruptions. Although the prominence module's agreement with a human rater is not perfect, the method works in the more difficult acoustic conditions of tutoring scenarios. Furthermore, the method definitely works considerably better than chance. By using more complex acoustic features, the results could possibly be improved. For German, including nucleus duration would likely lead to an improvement as long as it is estimated robustly.

6.3. Integration of Color Saliency and Prominence Detection into the Acoustic Packaging System

In the previous section, a module for tracking color salient regions and a module for detection prominent syllables in speech were described. Both modules communicate their hypotheses to the active memory, and thus, other modules within the system already have access to this information. However, some additions to the existing components are necessary to visualize the data and utilize the new information in the acoustic packaging system. In the following, these changes will be described first. Subsequently, the initial integration with the iCub platform (Metta et al., 2010) will be reported. Furthermore, findings on speech and object properties based on their local synchrony will be presented.

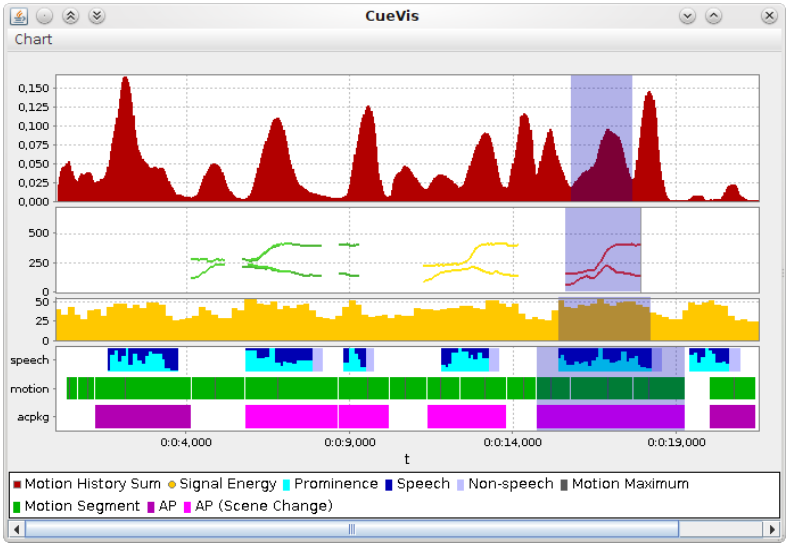


Figure 6.4.: Cue visualization tool showing motion peaks (row 1), trajectory coordinates (row 2 shows $x(t)$ and $y(t)$) acoustic signal energy (row 3), speech segmentation (row 4), visual segmentation (row 5), and acoustic packages (row 6).

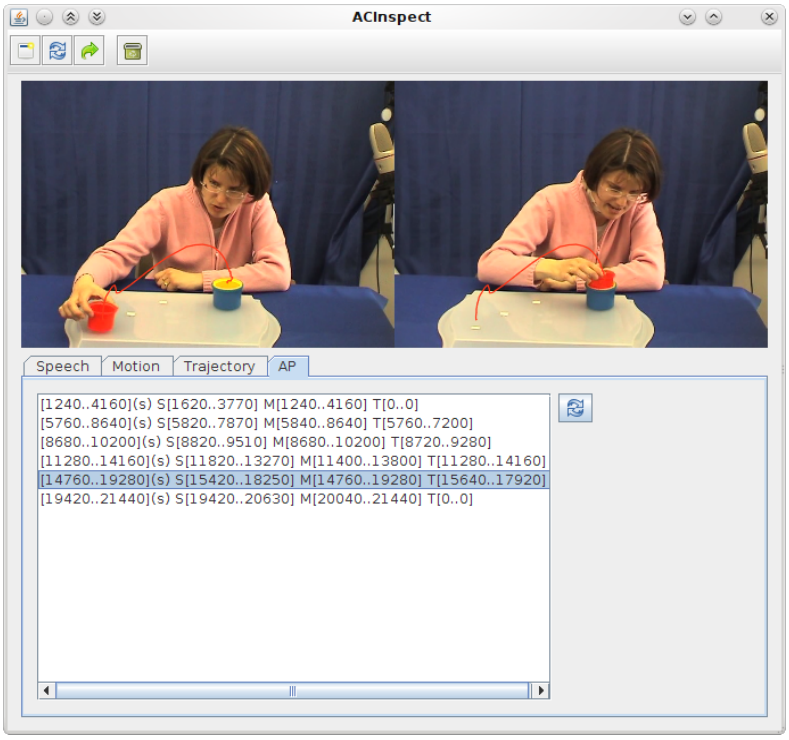


Figure 6.5.: Inspection tool showing a list of acoustic packages with details on each package's temporal extent and its associated segmentation hypotheses.

6.3.1. Additions to the Existing System Components

Including the new modules, the acoustic packaging system consists of five modules connected to the Active Memory (see Figure 6.3). The cue visualization and the inspection tool were extended to display trajectory data provided by the color saliency module. The cue visualization tool shows the temporal development of the x and y coordinates of the trajectory including the average color of each trajectory (see Figure 6.4). The inspection tool was extended to display the x and y coordinates as an overlay over the frames that mark the beginning and end of the current motion peak. This way, both the temporal development of each trajectory as well as its spatial accuracy and the relation to motion peaks can be analyzed. The segmentation of speech into syllables with their prominence rating is also displayed by the cue visualization tool as nested segments for each speech segment (see Figure 6.4). The inspection tool allows to replay prominent syllables for each speech segment.

Besides the changes related to visualization and inspection, the temporal association module was extended to additionally associate trajectories to acoustic packages. The association method follows a similar concept as for motion and speech. Overlapping trajectory and speech segments are associated to an acoustic package by the temporal association module. This step directly allows for an additional interpretation of the interaction based on the content of acoustic packages. Acoustic packages with no associated trajectory likely do not contain significant changes to the items in the interaction. These packages are generated by communication, which does not involve moving items, as, for example, showing an item or talking to the interaction partner. The cue visualization tool highlights these packages in a different color (see Figure 6.4).

6.3.2. Acoustic Packaging as a Basis for Feedback on the iCub Robot

The acoustic packaging system was tested on the iCub robot (see Figure 6.6). For this purpose a feedback module was implemented which uses information from acoustic packages to provide feedback to the tutor (see Figure 6.3). Currently, the feedback module focuses on extrapolating word-meaning pairs out of running interaction. This is realized by a two step process: During an action demonstration with the caregivers' verbal comments, the feedback module clusters acoustic packages by their trajectory color. For example, if the tutor — in the second step — shows a cup, but does not verbally comment his action, the robot provides feedback by replaying the most prominent syllable from one of the acoustic packages where the trajectory color matches the current one. This way the system communicates which information it has identified as relevant from the caregivers' demonstration. Furthermore, the feedback module replays the trajectory using the right arm of the iCub by mapping the trajectory coordinates into a two dimensional plane in front of the robot. The mapping process acts as a bridge component between the Active Memory and the cartesian controller (Pattacini et al., 2010) of the iCub. On the one hand, the component monitors the Active Memory for trajectories which have

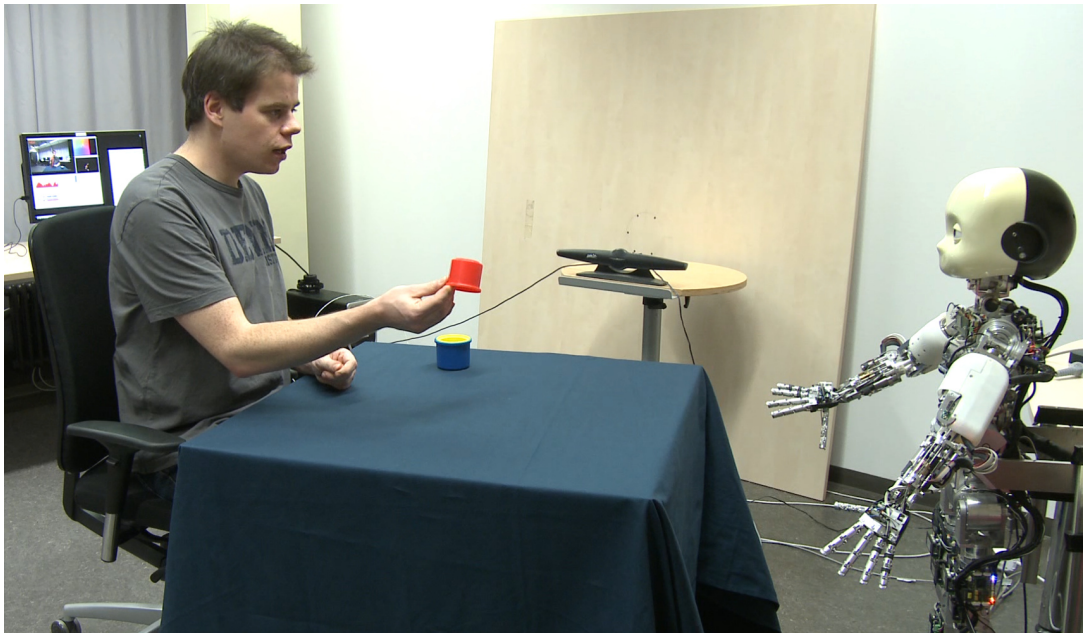


Figure 6.6.: A human user demonstrates cup stacking to the iCub robot. The iCub observes the scene through one of its eye cameras. Speech is recorded using an external microphone (middle). Visualization tools of the acoustic packaging systems are displayed on the screen in the background.

been selected for replay. On the other hand, the mapped coordinates are sequentially communicated to the cartesian controller. Note, that this aspect of feedback is in an experimental stage.

First tests of the feedback module on the iCub robot showed that the robots response referred to semantically relevant parts of the utterance. However, to close the loop between tutor and robot strategies for handling corrections or other types of feedback regarding the quality of the acoustic packages have to be implemented. Such methods would allow for developing the system to adapt to the tutor and keep only those packages which maintain information also considered as relevant by the tutor.

6.3.3. Summary

The prominence detection module and the color saliency tracking module was added to the acoustic packaging system. The temporal association module includes these additional cues when forming acoustic packages. Furthermore, the system was tested on the iCub robot where it provides initial feedback using the prominent syllables and trajectory information that was linked to acoustic packages.

6.4. Analysis of Local Synchrony within Acoustic Packages

The previous analyses of acoustic packaging were centered around more global properties based on the synchrony between action and speech as, for example, the number of motion peaks per acoustic package (see Chapter 5). The additional modules added to the acoustic packaging system (see Section 6.1 and 6.2) allow to take the meaning of the language uttered into consideration. Some parts of the speech are more prominent as they emphasize the corresponding meaning, which can link to the visual action demonstrated. This idea was used in the previous sections to develop a feedback module. Here, an analysis of words the tutor highlights in speech as well as the link of these words to object colors is carried out.

6.4.1. Procedure

This analysis focuses again on the cup nesting task from a corpus of adult-child interaction (Rohlfing et al., 2006). It consists of four differently sized and colored cups which have to be nested into each other (see Section 5.3.1). The acoustic packaging system was exposed to the multimodal corpus. In this analysis, a segmentation into words based on manual transcription of the tutors' utterances was used by the acoustic packaging system to segment speech. Manual transcription was necessary for the following analysis of prominent words. The prominence detection module automatically adds a syllable segmentation including the prominence rating to all speech segments processed by the system. Acoustic packages were acquired by querying the Active Memory the same way as in the previous experiments (see Section 5.2.2). The results now additionally contain the prominence ranking and the trajectories both associated to acoustic packages.

6.4.2. Prominent Words in Acoustic Packages

For this analysis, the most prominent word in each acoustic package was identified by matching the most prominent syllable with the word segmentation in that package. Based on these results, a count of all stressed words over all age groups was performed. Table 6.5 shows the 21 most frequent stressed words in the infant-directed speech within the acoustic packages from all parent infant interactions that were processed automatically.

As can be seen in Table 6.5, color terms (red, yellow, green) in different inflexions are among the most frequent prominent words. Only attention getters (look etc.), structuring devices (and, then), spatial markers (here, there, into) and the cup manipulated ("becher") are emphasized more frequently. Viewed from a level of categories color terms provide the largest group. No systematic change of these frequencies could be found over the four age groups in the corpus. However, this might be due to the small number of occurrences that resulted from selecting such a specific subset. Although especially emphasized spatial markers would be interesting to investigate in more detail, this analysis focuses on the

Rank	Frequency	Word
1	27	und (and)
2	21	den (the, ACC)
3	18	mal (modal particle)
4	17	becher (cup)
5	14	so (like this)
6	14	der (the)
7	**13**	**rote (red)**
8	**12**	**gelbe (yellow)**
9	11	ja (modal particle)
10	10	guck (look)
11	**10**	**grüne (green)**
12	10	dann (then)
13	10	da (there)
14	9	rein (into)
15	**9**	**grünen (green, ACC)**
16	8	auch (too)
17	7	in (in)
18	7	hier (here)
19	6	zack (onomatopoeisis)
20	5	hm (hm)
21	**5**	**gelben (yellow, ACC)**

Table 6.5.: The 21 most frequent stressed words as detected by the prominence detection module in infant-directed speech within Acoustic Packages. Translation of the German words are given in parentheses (ACC: accusative form)

color terms, since they can be related to color information the saliency module extracts from the visual modality. Therefore, do the color terms, that are often emphasized, correspond to the color properties of objects manipulated during action execution? This question will be analyzed in the following.

6.4.3. Relationship Color Adjectives with Motion Trajectories

The acoustic packaging system does not only provide information about the timing and the xy-coordinates of object movements, but also about the color of the moving object. Acoustic packages link this information with the tutor's speech based on its temporal overlap. Thus, the color of the trajectories can be linked with the stressed color term, which is present in that package. This result is visualized in Figure 6.7. The x-axis represents the relative position of the most stressed syllable within the corresponding trajectory, while the y-axis represents the relative position within the utterance. Plotted within this space are the stressed color terms and the color of the trajectory that has been observed at the same time. Note, that the depicted colors are real RGB-values acquired by the color saliency module.

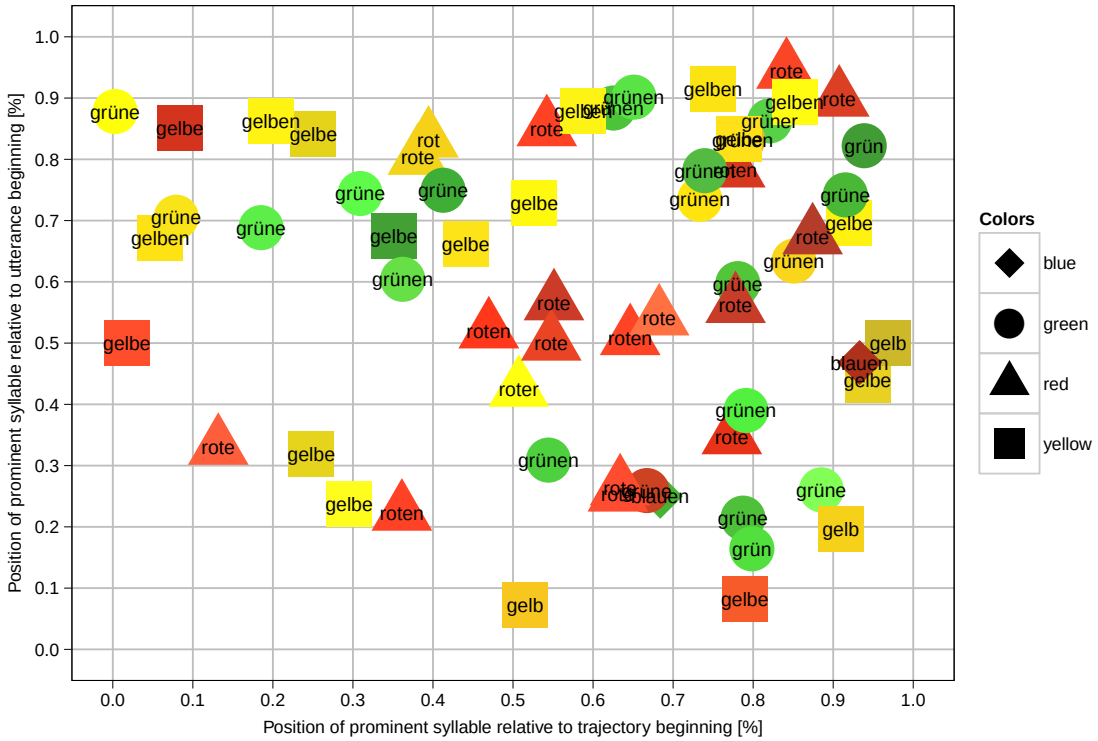


Figure 6.7.: Position of color terms within utterance and motion trajectory. The depicted colors are the same RGB values as observed in the videos along the trajectories of the moved objects. The shape of each item indicates the translation of the corresponding color term (see legend).

Age Group	Color Terms [%]
1	15.9
2a	16.1
2b	10.5
3	14.1
all	14.6

Table 6.6.: Percentage of color terms of stressed words for different age groups.

As can be seen, the stressed color words coincide frequently with the color of the object being moved at that time. Hand-coding of the number of cases where the word coincides with the corresponding trajectory color yields a result of 79% of all stressed color terms matching exactly the object color. This is an exciting result as it suggests a strong relationship between stressed word and actual color information, which makes it possible to automatically learn the association between color clusters and the word. However, this is only possible by focusing on the stressed color terms, thus filtering out stressed non-color terms. This is analogous to applying a top-down bias towards color words. Although color terms are by far the most frequent ones, they present only about 14.6% of all stressed words.

In current models of word learning such as the Emergentist Coalition Model, it is assumed that infants differentially weigh the input they receive from their caregivers depending on their stage of development (see Section 3.2.2; Hollich et al., 2000a). On the other hand, it has been pointed out that caregivers are very sensitive to the development of their infant and provide specifically designed input towards them (Pitsch et al., 2009; Vollmer et al., 2010). It could, thus, be possible that the caregivers provide such a top-down bias towards color terms depending on the perceived level of development of their infant. Therefore, the relation of the frequency of color terms provided by the parents to the age of their children is analyzed. For this purpose, the relative amount of emphasized color terms was calculated for each age group in the corpus of adult-child interactions (see Section 5.3.1). Table 6.6 shows the resulting percentage of color terms of all stressed words over all four age groups.

Again, no systematic pattern can be found by this analysis alone. The amount of color words remains relatively stable around 15%. However, the cup nesting task is not primarily about color learning, but rather about which cup goes into which. Thus, there may be other cues provided by the tutor to educate the infant's attention towards what s/he thinks most relevant for the infant at that time.

6.4.4. Conclusion

In the last sections, prominent words within acoustic packages were analyzed and compared to color properties of object trajectories, which coincide with these words and are, thus, associated to these packages based on their temporal synchrony. The

analysis based on a corpus of adult-child interactions in the context of a cup stacking task revealed that color adjectives are frequently emphasized by the tutor. Subsequently, the comparison with trajectory properties showed that acoustic packaging can be used to acquire semantic knowledge such as color names from action demonstrations.

However, this analysis only provides evidence for few cues such as the synchrony of emphasized syllables and object color. Caregivers provide more cues as, for example, spatial markers or temporal segmentation markers. While some of these cues can be related to objects or their relations, others must be related to the structure of the ongoing action as, for example, the order of objects. These cues also need to be analyzed but due to their lower number of occurrence they might require larger corpora. An additional aspect is that corpora cannot reflect the interaction of a system with a human. Realizing a scenario with interactions over a longer period of time requires new modules which provide feedback about what the system has understood about the ongoing action demonstration to maintain the interaction. This context raises the question about the representational capabilities and limitations of acoustic packages which will be discussed in the next chapter along with possible further development steps of acoustic packaging.

6.5. Summary

This chapter described two additional modules which were integrated in the acoustic packaging system. The color saliency based tracking module tracks and extracts features of moving colored regions typical for children's toys. The prominence detection module identifies emphasized syllables in human speech. Both cues are linked by acoustic packages and therefore associate visual and acoustic information, which can be used to respond to an interaction partner. This was evaluated by testing the acoustic packaging system on the iCub robot. Using the additional cues delivered by the new modules, a feedback module was implemented to provide feedback on the color of cups used in an action demonstration towards the robot. By responding with syllables which were emphasized during the demonstration, the system is able to communicate which associations it has made from observing the action demonstration. The association between the color of objects and emphasized syllables was analyzed in detail on the corpus with adult-child interactions again (see Chapter 5). The results showed that this semantic knowledge is represented in acoustic packages. The representational capabilities of acoustic packages and future development steps of acoustic packaging will be discussed in the following chapter.

7. A Roadmap to Multimodal Action and Language Learning in Interaction

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. On the other hand, it can be used as an analysis tool for tutoring interactions. In this thesis it is argued that acoustic packaging provides the first steps towards learning action in interaction. In this section, the further development of acoustic packaging will be envisioned. To achieve this goal a discussion of the representational capabilities of acoustic packages will provide insights which future development topics need to be addressed. Portions of this chapter were previously published by the author (Schillingmann et al., 2009b).

7.1. Representation of Action Perception and Action Production in Acoustic Packages

In general, acoustic packaging fuses separate events from vision and speech into macro-events, which provide an initial segmentation of action without previous training. This is accomplished by looking at the temporal relationship of events in both signals and combining temporally overlapping events into acoustic packages. Acoustic packages already contain information about objects manipulated and related parts in the tutors speech. An overview of these cues is depicted in Figure 7.1.

Speech segments and motion peaks provide basic cues for an initial segmentation of multimodal action demonstrations (see Sections 4.3.3 and 4.3.4). Cues such as trajectories based on color saliency and acoustic prominence allow the system to identify information that goes beyond temporal segmentation. Trajectories contain information on the path of the action (see Sections 6.1 and 6.2). Their beginning and endpoints provide information on potential goals. The information about prominence delivers the parts of utterances that are highlighted by the tutor. Thus, acoustic packages provide a link between linguistic and visual cues. Collecting samples of speech that relate to certain visual cues can be achieved by clustering acoustic packages. For example, speech samples that tend to relate to color can be acquired by using trajectory color to cluster acoustic packages. The speech samples can then be acquired by extracting prominent syllables from the acoustic

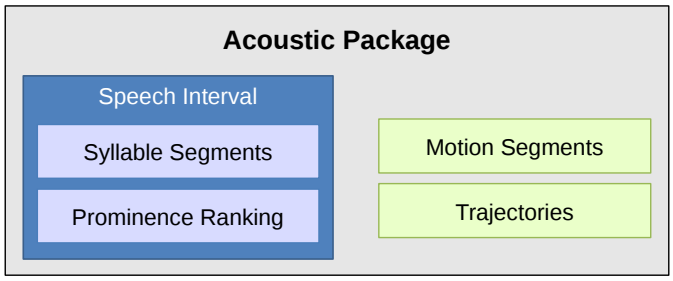


Figure 7.1.: Overview of the sensory cues which are associated to an acoustic package

segmentation that each acoustic package provides. In a demonstration system on the iCub robot, acoustic feedback was successfully provided within an interactive tutoring situation. Due to the link between speech and visual cues, the transition from visual to linguistic information can be made quickly. Furthermore, the different cues within an acoustic package represent action on different levels of granularity. The motion segmentation is of a fine granularity while acoustic packages represent action on a more coarse grained level. This can be viewed as a first step towards a hierarchical representation of the observed action.

Although acoustic packages represent actions in several aspects, yet not every aspect of action can be modeled. This concerns higher level representations of action, which require accumulating, filtering and linking acoustic packages over a longer period of time to form them. First steps in this direction are made by persistently storing acoustic packages within the Active Memory. However, development and changes in a high level representation possibly affect the action segmentation process. For example, certain known parts of actions could lead to an instantiation of new cues which then can be packaged again. As a result, a more complex *hierarchical action structure* develops over time. In this structuring process, further cues like *rhythm and long term repetitions* could help to separate different types of action as, for example, separating path and manner oriented actions (Wagner and Lakusta, 2009). A first basis for such processes is provided by the timeline where all acoustic packages are aligned to. Further related to long term analysis are aspects of language development. While, for example, prominence cues provide immediate information on highlighted parts within speech, long term analysis is required to identify *linguistic structures* such as words. Furthermore, the differentiation between linguistic classes like nouns and verbs is important. Finding linguistic and visual cues to distinguish and relate these units to actions is important to develop linguistic capabilities that go beyond immediate feedback in response to tutoring interaction. Another challenge is to build a representation based on acoustic packages that facilitates prediction of action such as anticipating goals. A combination of both dynamical and conceptual information possibly contributes to this ability. Additional *social cues* such as eye gaze might support acquiring these capabilities.

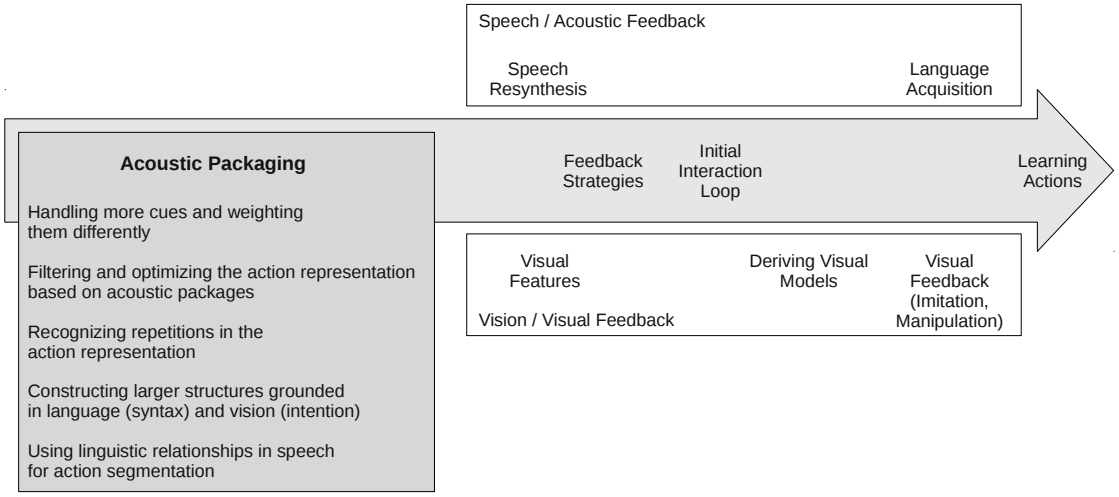


Figure 7.2.: Roadmap showing future improvements of acoustic packaging in one dimension and next steps towards action learning in the other dimension.

Regarding *action production*, the motion trajectory information in acoustic packages can be used to imitate movements that have been previously segmented. This has been tested with the iCub robot as possible feedback in the case that a tutor shows objects that have been previously used in an action demonstration. However, currently, this imitation is limited to a 2D plane and no adaption to goals is made: This demonstrates that acoustic packages can be used for very basic imitation tasks. For a more generic approach, the system must be able to create action primitives by transforming observations and exploring his own body dynamics as well as the environment.

7.2. Roadmap Overview

The previous analysis shows that a system learning and representing action in tutoring situations requires a complex set of capabilities, which exceeds the scope of this work. Based on the current results, the further development of acoustic packaging can be pushed forward along two dimensions (see Figure 7.2). One dimension spans across the improvements of acoustic packaging as a method and acoustic packages as a representation format. The other dimension extents when acoustic packaging is put into human-robot interaction scenarios. Here, the general goal is that a robot will learn actions by interaction for which the necessary step is the further development of feedback strategies (Wrede et al., 2010; Asada et al., 2009) in the robot and the initiation of interaction loops (Pitsch et al., 2009). The improving capabilities of the robot in interactive scenarios is linked to further developmental steps concerning speech and vision processing but also acoustic and visual feedback. In the following section, the items on the roadmap will be described in detail.

7.3. Handling More Cues

Concerning the further development of acoustic packaging, more cues might be helpful to bootstrap action representations which are grounded both visually and acoustically. First cues supporting segmentation were integrated into acoustic packaging to analyze manipulation in the environment. However, in addition to research about what and when to imitate (Breazeal and Scassellati, 2001, 2002; Carpenter and Call, 2007) it might be important for the robot to distinguish between human motion on the one hand and objects in the environment manipulated by humans on the other hand. 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 three months of age, but when tested with faces this capability can be observed significantly earlier (Boyer and Bertenthal, 2008). 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. The color saliency module already provides information on the situational change to a certain extent. However, more cues might be required to detect object relations. 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 literature review, the Emergentist Coalition Model suggests to take a different nature of cues into consideration (see Section 3.2.2; Hollich et al., 2000b). 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 in their perceptual system. More specifically, in the first stage of development, predominantly perceptual cues are taken into consideration. Starting from the 10th month of age, 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.

Furthermore, action production methods might require proprioceptive cues, such as joint positions. These cues are already used in imitation learning scenarios which include physical interaction with the robot (Calinon and Billard, 2007). In case of reaching and grasping, action cues are needed that help interpreting the scene further by providing information on possible constraints. Obstacles, for example, can thus be considered by using models which integrate information of areas that should be avoided when the system reaches or grasps (Gori et al., 2012).

7.4. 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 to a format that is appropriate for long-term storage. In this format, a higher conceptualization, 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 (Barsalou, 1999). When new acoustic packages are acquired, they are compared to other packages and related to one another. 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. The underlying comparison method might not take place in its original representation, but certain cues could be transformed into a different representation for this process. An example is using a motor representation of a trajectory instead of comparing visual observations.

7.5. Recognizing Repetitions in the Action Representation

The ability to bootstrap action concepts requires clustering of similar parts in the action stream the robot perceives. 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.

7.6. 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 (Tomasello, 2001), 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 Tomasello (2001), 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. If 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.

7.7. Using Linguistic Relationships in Speech for Action Segmentation

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 using 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 sensory 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.

7.8. 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. Initially 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. First steps in this direction were realized by including the prominence module to detect emphasized syllables and using them to provide feedback to the tutor (see Section 6.2). In an interaction with a tutor, this feedback behavior signals either that the robot knows the action or that the demonstrated action consists of new unknown movements. In accordance with this idea, Pitsch et al. (2009) observed that gaze behavior serves as an indicator of a child's knowledge of an action: When a child knows an action he or she gazes at the target (for example, the target cup in the stacking cups task) instead of looking on each single demonstration movement.

In the case that the robot is further advanced in development and capable of performing the demonstrated action, its manipulation actions can itself be seen as a form of feedback. Any kind of imitation is viewed as visual information about the internal representation of action to the tutor (Carpenter and Call, 2007). A related type of feedback is to reach and grasp to initiate physical interaction with objects. This feedback signals that objects taking part in an action can be identified by the system. Furthermore, it communicates the start of a physical interaction to the tutor. Methods based on the passive motion paradigm (Gori et al., 2012) could provide means to flexibly adapt primitives to obstacles and objects which are moved by the tutor during interaction.

7.9. 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 (Pitsch et al., 2009). For example, when children's attention is distracted, parents produce salient movements with the purpose of attracting children's attention to the demonstrated objects and actions. In contrast, when children's attention follows the demonstration, less modified movements can be observed (Pitsch et al., 2009). Thus, it seems that the modifications in movements — called motionese (as summarized in Rohlfing et al., 2006) — are a product of the interaction loop. The development of feedback forms can therefore only be the first step. The tutor's teaching behavior is likely 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.

7.10. Conclusion

Action perception and production based on acoustic packages was analyzed including possible future directions. Furthermore, a roadmap was presented which highlights important capabilities a system for action and language learning in interaction should possess. An important factor for action representations is the capability of keeping connections between different modalities that survive further abstraction processes during development. Acoustic packages are a first step in this direction as they provide an initial action segmentation that links corresponding visual and acoustic events. This link is not only necessary for a system to respond to requests or react to a specific setup but also to share its current level of understanding by providing feedback. For some of the topics addressed in this roadmap section, isolated methods are already available. However, integrating them in one system, which is flexible enough to learn actions over an extended period of time, is still a challenge.

8. Conclusion

This thesis was inspired by the overarching idea that developmental action and language learning in robotics can be realized by learning from interaction with humans. The driving question was "How can we take advantage of speech and action synchrony?". Specifically, synchrony between action and language was assumed to be beneficial for finding relevant parts and extracting first knowledge from action demonstrations. For this purpose a computational model of acoustic packaging was developed which binds visual and acoustic events to acoustic packages based on their temporal overlap. The central contribution of this work comprises the conception, further development, and implementation of a model that has been inspired by the general idea of acoustic packaging as outlined by psychological research. The resulting model of acoustic packaging is able to segment action demonstrations into multimodal units which are called acoustic packages. These units facilitate measuring the level of structuring in action demonstrations. In addition to action segmentation, the acoustic packaging system is able to flexibly integrate additional sensory cues to acquire first knowledge about the content of action demonstrations. Furthermore, the system was designed to process input online, which enables it to provide feedback to users engaging in an interaction with a robot.

For the modeling process a broad area of related work was taken into account. Psychological research on event and action segmentation provided the insight that both adults and children perceive events as variably sized units. Motion features provide one important cue for finding these units. However, more information was required to identify structure in actions that leads to meaningful units. Here, research on modality integration in infant development provided central theories. It is suggested that stimuli which are redundant across multiple modalities help children in finding meaningful units in the stream of multimodal sensory input despite their limited previous knowledge. In this context, acoustic packaging was introduced as a bottom-up cue for language comprehension, since it associates temporally related visual and acoustic events. In contrast to multimodal saliency models, acoustic packaging does not require events that concur exactly; instead, temporally overlapping events are already considered as related. For example, the comments of a human caretaker during an action demonstration are combined with co-occurring visual events and thus reveal structure in the interaction. This view was further extended by the Emergentist Coalition Model which suggests that multiple cues including attentional cues contribute to language development. According to the auditory dominance theory, speech seems to have a certain attentional priority due to its transient nature.

In the design of the acoustic packaging model previous insights and theories were taken into consideration. In contrast to typical systems in robotics, acoustic packaging requires modality specific segmentation methods which do not depend on extensive previous knowledge. Related work on video and action segmentation systems suggests appropriate cues which are consistent with theories from psychological research. A frequent commonality is their aim to detect discontinuities in the visual input. Therefore, the acoustic packaging system uses an approach based on motion history images to segment the visual input into peaks with increased motion. With regard to acoustic segmentation, utterances separated by pauses were identified as sensible units. A temporal association module forms acoustic packages by associating both types of segments. Based on the idea that speech guides the process of finding structure in events, which is also supported by the auditory dominance theory, the model allows to associate multiple motion peaks to one acoustic package. All modules in the acoustic packaging system are designed for processing input online. This capability is important in robotic systems to provide feedback to human interaction partners. Furthermore, the system architecture follows a modular design using a central Active Memory for integrating modules and storing events. This design was chosen to simplify the extension of the system and to facilitate its evaluation by recalling stored events.

Acoustic packaging was evaluated on a corpus of adult-adult and adult-child interactions within a cup stacking scenario. A difference between the structure of child-directed and adult-directed interactions was expected. The evaluation revealed that major differences were found in the number of acoustic packages and in the number of motion peaks per acoustic package. Further analysis of this corpus within age groups ranging from 8–30 months showed that developmental trends are reflected in the statistical properties of acoustic packages. For example, the number of motion peaks per acoustic package increases with children's age suggesting that caregivers adapt the level of complexity of their tutoring based on the infant's improving abilities. In addition to adult-child interaction, a corpus from a similar scenario with a simulated robot was analyzed. The results indicate that adult-robot interaction exhibits a similar structure compared to adult-child interaction.

Additional cues can be easily integrated in the acoustic packaging system. Specifically, integrating acoustic packaging on a robotic platform requires cues which allow for extracting semantic details from action demonstrations, that go beyond structural properties. This semantic information can be used to provide feedback to the tutor. Therefore, acoustic packaging was extended with a color saliency tracking module and a prominence detection module which allow the system to detect moving colored regions and accumulate their trajectories as well as detecting syllables emphasized by the tutor. Tests on the iCub robot showed that semantic information on color terms can be extracted from acoustic packages by connecting visual saliency information with syllables highlighted by the tutor. These results were supported by further analysis of adult-child interactions, which verified that a substantial amount of semantic information can be gathered by exploiting this connection.

Although this work showed that acoustic packaging is able to temporally segment action demonstrations, to assess action structure, and to derive first semantic knowledge, it is only a first step towards developmental learning of actions from interaction. Therefore, future steps in this direction were outlined in a roadmap. One key aspect is the initial interaction loop which requires more complex feedback strategies of the robot. These strategies are necessary to establish a continuous interaction between the tutor and the learner allowing for robotic systems that continuously acquire action and language knowledge. This knowledge includes the link between sensory cues grounded in multiple modalities. In this view, acoustic packages provide the initial representation of action structure in interaction.

A. Additional Evaluation Results on Adult-Adult and Adult-Child Interaction

In this appendix, additional statistics based on acoustic packaging of adult-adult and adult-child interactions are presented.

		ACI	AAI	ACI-AAI	
		M (SD)	M (SD)	Z	p
1	Number of subjects	24	23		
2	Total number of APs	13.25 (7.33)	4.13 (2.56)	5.1	0.000
3	Total length of APs [s]	39.30 (25.57)	15.45 (11.24)	4.3	0.000
4	Average length of APs [s]	2.88 (0.55)	3.73 (1.33)	−2.5	0.011
5	Total number of MPs (in APs)	18.25 (10.94)	8.48 (5.90)	3.9	0.000
6	Total length of MPs (in APs) [s]	22.48 (12.22)	8.25 (5.45)	4.8	0.000
7	Average length of MPs (in APs) [s]	1.27 (0.25)	1.01 (0.21)	3.7	0.000
8	Total number of MPs	26.79 (12.59)	11.48 (5.86)	4.4	0.000
9	Total length of MPs [s]	31.21 (14.12)	10.48 (5.26)	5.2	0.000
10	Average length of MPs [s]	1.20 (0.23)	0.94 (0.18)	3.8	0.000
11	Total number of utterances	15.38 (9.84)	4.61 (2.79)	5.1	0.000
12	Total length of utterances [s]	10.52 (6.33)	6.10 (4.49)	3.1	0.002
13	Average utterance length [s]	0.72 (0.23)	1.38 (0.57)	−4.9	0.000
14	Average utterance length (in APs) [s]	0.75 (0.23)	1.46 (0.55)	−5.1	0.000
15	Total number of pauses in speech	14.38 (9.84)	3.61 (2.79)	5.1	0.000
16	Total length of pauses in speech [s]	21.65 (10.84)	3.08 (2.51)	5.7	0.000
17	Average length of pauses in speech [s]	1.68 (0.68)	1.04 (1.43)	4.1	0.000
18	Average number of MPs per AP	1.37 (0.20)	2.12 (0.61)	−4.4	0.000
19	Ratio of interaction length to speech length	3.63 (1.46)	2.30 (1.43)	4.0	0.000
20	Ratio of AP length to speech length (in APs)	4.08 (1.63)	2.73 (1.11)	3.5	0.000
21	Ratio of AP count to speech length (in APs) 1/[s]	1.34 (0.44)	0.74 (0.26)	4.8	0.000
22	Ratio of all MPs to MPs assigned to APs	1.55 (0.45)	1.65 (1.07)	1.3	0.194
23	Ratio of interaction length to AP length	0.97 (0.31)	0.97 (0.62)	1.5	0.142

Table A.1.: Results from the comparison of child-directed versus adult-directed interaction (group 1: 8–12 months). The right columns show the results of Wilcoxon Mann-Whitney rank sum tests between ACI and AAI.

		ACI	AAI	ACI-AAI	
		M (SD)	M (SD)	Z	p
1	Number of subjects	12	10		
2	Total number of APs	6.58 (4.91)	4.30 (1.89)	1.0	0.318
3	Total length of APs [s]	17.55 (13.09)	15.63 (6.55)	0.0	1.000
4	Average length of APs [s]	2.58 (0.75)	4.14 (2.19)	−2.0	0.048
5	Total number of MPs (in APs)	10.58 (6.92)	8.80 (4.05)	0.5	0.597
6	Total length of MPs (in APs) [s]	11.05 (7.54)	9.24 (4.95)	0.4	0.717
7	Average length of MPs (in APs) [s]	1.04 (0.16)	1.04 (0.16)	0.5	0.644
8	Total number of MPs	15.75 (5.40)	12.30 (4.24)	1.6	0.112
9	Total length of MPs [s]	15.73 (6.05)	11.82 (4.91)	1.6	0.114
10	Average length of MPs [s]	0.99 (0.11)	0.96 (0.16)	0.4	0.692
11	Total number of utterances	7.42 (6.01)	4.80 (1.75)	0.9	0.386
12	Total length of utterances [s]	6.28 (4.92)	6.95 (4.19)	−0.6	0.553
13	Average utterance length [s]	0.88 (0.44)	1.90 (2.15)	−1.6	0.114
14	Average utterance length (in APs) [s]	0.91 (0.44)	1.91 (2.15)	−1.5	0.147
15	Total number of pauses in speech	6.42 (6.01)	3.80 (1.75)	0.9	0.386
16	Total length of pauses in speech [s]	7.17 (5.25)	3.83 (3.55)	1.7	0.086
17	Average length of pauses in speech [s]	1.27 (0.46)	0.86 (0.72)	1.5	0.129
18	Average number of MPs per AP	1.65 (0.45)	2.45 (1.77)	−1.3	0.208
19	Ratio of interaction length to speech length	5.78 (8.80)	2.47 (1.61)	1.7	0.086
20	Ratio of AP length to speech length (in APs)	3.14 (1.00)	3.03 (2.19)	1.5	0.147
21	Ratio of AP count to speech length (in APs) 1/[s]	1.35 (0.96)	0.90 (0.73)	1.8	0.075
22	Ratio of all MPs to MPs assigned to APs	2.30 (2.17)	1.53 (0.59)	0.5	0.644
23	Ratio of interaction length to AP length	1.65 (1.68)	0.87 (0.16)	1.1	0.291

Table A.2.: Results from the comparison of child-directed versus adult-directed interaction (group 2a: 12–18 months). The right columns show the results of Wilcoxon Mann-Whitney rank sum tests between ACI and AAI.

		ACI	AAI	ACI-AAI	
		M (SD)	M (SD)	Z	p
1	Number of subjects	10	13		
2	Total number of APs	11.70 (5.79)	4.23 (1.64)	3.5	0.000
3	Total length of APs [s]	31.50 (15.45)	14.72 (8.25)	2.7	0.008
4	Average length of APs [s]	2.68 (0.46)	3.45 (1.09)	−1.8	0.072
5	Total number of MPs (in APs)	18.30 (9.06)	8.69 (4.61)	2.7	0.007
6	Total length of MPs (in APs) [s]	21.26 (10.21)	8.40 (3.83)	3.4	0.001
7	Average length of MPs (in APs) [s]	1.19 (0.16)	1.01 (0.22)	2.1	0.035
8	Total number of MPs	24.10 (11.70)	10.85 (4.58)	3.0	0.003
9	Total length of MPs [s]	27.33 (13.67)	10.08 (3.71)	3.6	0.000
10	Average length of MPs [s]	1.15 (0.16)	0.96 (0.19)	2.4	0.018
11	Total number of utterances	13.60 (6.40)	4.46 (1.81)	3.9	0.000
12	Total length of utterances [s]	12.56 (6.51)	5.98 (3.91)	3.0	0.002
13	Average utterance length [s]	0.92 (0.20)	1.45 (0.82)	−1.7	0.094
14	Average utterance length (in APs) [s]	1.02 (0.24)	1.48 (0.80)	−1.6	0.107
15	Total number of pauses in speech	12.60 (6.40)	3.46 (1.81)	3.9	0.000
16	Total length of pauses in speech [s]	15.56 (8.21)	3.28 (1.90)	4.0	0.000
17	Average length of pauses in speech [s]	1.26 (0.36)	0.92 (0.53)	2.2	0.026
18	Average number of MPs per AP	1.56 (0.21)	2.17 (0.81)	−1.8	0.077
19	Ratio of interaction length to speech length	2.51 (0.48)	2.06 (0.55)	2.0	0.041
20	Ratio of AP length to speech length (in APs)	2.76 (0.85)	2.70 (1.08)	0.0	1.000
21	Ratio of AP count to speech length (in APs) 1/[s]	0.96 (0.24)	0.81 (0.30)	1.0	0.321
22	Ratio of all MPs to MPs assigned to APs	1.34 (0.17)	1.31 (0.19)	0.6	0.555
23	Ratio of interaction length to AP length	1.04 (0.31)	0.86 (0.30)	1.3	0.193

Table A.3.: Results from the comparison of child-directed versus adult-directed interaction (group 2b: 18–24 months). The right columns show the results of Wilcoxon Mann-Whitney rank sum tests between ACI and AAI.

		ACI	AAI	ACI-AAI	
		M (SD)	M (SD)	Z	p
1	Number of subjects	18	20		
2	Total number of APs	8.17 (2.66)	3.90 (1.86)	4.3	0.000
3	Total length of APs [s]	26.24 (9.06)	13.84 (6.31)	3.8	0.000
4	Average length of APs [s]	3.28 (0.82)	3.62 (0.63)	−1.2	0.242
5	Total number of MPs (in APs)	13.33 (4.33)	8.65 (4.11)	2.9	0.003
6	Total length of MPs (in APs) [s]	15.68 (5.49)	8.17 (4.20)	3.7	0.000
7	Average length of MPs (in APs) [s]	1.17 (0.14)	0.93 (0.16)	3.9	0.000
8	Total number of MPs	17.44 (4.85)	12.10 (4.14)	3.2	0.001
9	Total length of MPs [s]	19.34 (5.78)	10.84 (4.08)	3.9	0.000
10	Average length of MPs [s]	1.11 (0.10)	0.90 (0.13)	4.2	0.000
11	Total number of utterances	9.56 (3.29)	4.15 (1.93)	4.5	0.000
12	Total length of utterances [s]	9.36 (4.58)	5.67 (3.33)	2.6	0.010
13	Average utterance length [s]	1.02 (0.54)	1.42 (0.78)	−1.7	0.085
14	Average utterance length (in APs) [s]	1.13 (0.60)	1.50 (0.79)	−1.6	0.114
15	Total number of pauses in speech	8.56 (3.29)	3.15 (1.93)	4.5	0.000
16	Total length of pauses in speech [s]	10.28 (3.62)	3.26 (2.09)	4.6	0.000
17	Average length of pauses in speech [s]	1.24 (0.32)	1.11 (0.50)	1.3	0.198
18	Average number of MPs per AP	1.67 (0.38)	2.35 (0.88)	−2.6	0.010
19	Ratio of interaction length to speech length	2.99 (2.36)	2.57 (1.40)	0.6	0.539
20	Ratio of AP length to speech length (in APs)	3.99 (3.15)	3.24 (2.48)	1.5	0.121
21	Ratio of AP count to speech length (in APs) 1/[s]	1.26 (1.22)	0.92 (0.69)	1.3	0.209
22	Ratio of all MPs to MPs assigned to APs	1.35 (0.27)	1.53 (0.45)	−1.3	0.208
23	Ratio of interaction length to AP length	0.84 (0.20)	0.89 (0.30)	−0.1	0.953

Table A.4.: Results from the comparison of child-directed versus adult-directed interaction (group 3: 25–30 months). The right columns show the results of Wilcoxon Mann-Whitney rank sum tests between ACI and AAI.

Bibliography

J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, Nov. 1983. ISSN 0001-0782. doi: 10.1145/182.358434.

M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive Developmental Robotics: A Survey. *IEEE Transactions on Autonomous Mental Development*, 1(1):12–34, Apr. 2009. doi: 10.1109/TAMD.2009.2021702.

L. E. Bahrick, R. Flom, and R. Lickliter. Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental psychobiology*, 41(4):352–363, Dec. 2002. ISSN 0012-1630. doi: 10.1002/dev.10049.

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*, 13(3):99–102, June 2004. ISSN 1467-8721. doi: 10.1111/j.0963-7214.2004.00283.x.

L. E. Bahrick, R. Lickliter, and R. Flom. Up Versus Down: The Role of Intersensory Redundancy in the Development of Infants' Sensitivity to the Orientation of Moving Objects. *Infancy*, 9(1):73–96, 2006. doi: 10.1207/s15327078in0901_4.

L. E. Bahrick, R. Lickliter, I. Castellanos, and M. Vaillant-Molina. Increasing task difficulty enhances effects of intersensory redundancy: testing a new prediction of the Intersensory Redundancy Hypothesis. *Developmental Science*, 13(5):731–737, Sept. 2010. ISSN 1467-7687. doi: 10.1111/j.1467-7687.2009.00928.x.

D. Baldwin, J. Baird, M. Saylor, and M. Clark. Infants parse dynamic action. *Child development*, 72(3):708–717, 2001. ISSN 0009-3920.

L. W. Barsalou. Perceptual symbol systems. *The Behavioral and brain sciences*, 22(4): 577–609, Aug. 1999. ISSN 0140-525X.

B. I. Bertenthal and J. J. Campos. A systems approach to the organizing effects of self-produced locomotion during infancy. *Advances in Infancy Research*, 6:1–60, 1990.

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*, pages 2401–2404, 2005.

G. Biggs and B. Macdonald. A Survey of Robot Programming Systems. In *in Proceedings of the Australasian Conference on Robotics and Automation, CSIRO*, 2003.

H. G. Birch and A. Lefford. *Intersensory development in children.* Child Development Publications of the Society for Research in Child Development, 1963.

T. W. Boyer and B. I. Bertenthal. Predictive tracking of social and non-social stimuli. In *Biennial International Conference on Infant Studies*, Vancouver, BC, Canada, 2008.

R. J. Brand and W. L. Shallcross. Infants prefer motionese to adult-directed action. *Developmental Science*, 11(6):853–861, Nov. 2008. ISSN 1363755X. doi: 10.1111/j.1467-7687.2008.00734.x.

R. J. Brand and S. Tapscott. Acoustic Packaging of Action Sequences by Infants. *Infancy*, 11(3):321–332, 2007. doi: 10.1080/15250000701310413.

R. J. Brand, D. A. Baldwin, and L. A. Ashburn. Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72–83, Mar. 2002. doi: 10.1111/1467-7687.00211.

C. Breazeal and B. Scassellati. Challenges in Building Robots That Imitate People. In *Imitation in animals and artifacts*, pages 363–390. MIT Press, Cambridge, MA, USA, 2001. ISBN 0-262-04203-7.

C. Breazeal and B. Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481–487, Nov. 2002. ISSN 13646613. doi: 10.1016/S1364-6613(02)02016-8.

C. Breazeal, G. Hoffman, and A. Lockerd. Teaching and Working with Robots as a Collaboration. In *AAMAS '04 Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1030–1037, New York City, New York, USA, 2004. IEEE Computer Society. ISBN 1-58113-864-4. doi: 10.1109/AAMAS.2004.258.

D. Buchsbaum, K. R. Canini, and T. L. Griffiths. Segmenting and Recognizing Human Action Using Low-Level Video Features. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci)*, 2011.

S. Calinon and A. G. Billard. What is the Teacher's Role in Robot Programming by Demonstration? Toward Benchmarks for Improved Learning. *Science*, 8(3):441–464, 2007. ISSN 15720373.

M. Carpenter and J. Call. The question of 'what to imitate': inferring goals and intentions from demonstrations. In C. Nehaniv and K. Dautenhahn, editors, *Imitation and social learning in robots, humans and animals: Behavioural, social and communicative dimensions*, pages 135–151. Cambridge University Press, Cambridge, 2007.

J. W. Davis and A. F. Bobick. The Representation and Recognition of Human Movement Using Temporal Templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928–934. IEEE Computer Society, 1997. ISBN 0-8186-7822-4. doi: 10.1109/CVPR.1997.609439.

R. Dillmann, T. Asfour, M. Do, R. Jäkel, A. Kasper, P. Azad, A. Ude, S. R. Schmidt-Rohr, and M. Lösch. Advances in Robot Programming by Demonstration. *KI - Künstliche Intelligenz*, 24(4):295–303, Aug. 2010. ISSN 0933-1875. doi: 10.1007/s13218-010-0060-0.

S. Ekvall and D. Kragic. Integrating object and grasp recognition for dynamic scene interpretation. In *Advanced Robotics, 2005. ICAR '05. Proceedings., 12th International Conference on*, pages 331–336, Seattle, USA, 2005. doi: 10.1109/ICAR.2005.1507432.

G. A. Fink. Developing HMM-Based Recognizers with ESMERALDA. In V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, pages 229–234. Springer, Berlin, Heidelberg, 1999. ISBN 3-540-66494-7.

K. Fischer, K. Foth, K. J. Rohlfing, and B. Wrede. Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12(1):134–161, 2011. doi: 10.1075/is.12.1.06fis.

P. Fitzpatrick, A. Arsenio, and E. R. Torres-Jara. Reinforcing robot perception of multi-modal events through repetition and redundancy and repetition and redundancy. *Interaction Studies*, 7(2):171–196, 2006. ISSN 1572-0373. doi: 10.1075/is.7.2.05fit.

R. Flom and L. E. Bahrick. The effects of intersensory redundancy on attention and memory: infants' long-term memory for orientation in audiovisual events. *Developmental psychology*, 46(2):428–436, Mar. 2010. ISSN 1939-0599. doi: 10.1037/a0018410.

S. Frintrop and M. Kessel. Most salient region tracking. In *2009 IEEE International Conference on Robotics and Automation*, pages 1869–1874. IEEE, May 2009. ISBN 978-1-4244-2788-8. doi: 10.1109/ROBOT.2009.5152298.

J. Fritsch and S. Wrede. An Integration Framework for Developing Interactive Robots. In D. Brugali, editor, *Software Engineering for Experimental Robotics*, pages 291–305. Springer, 2007. doi: 10.1007/978-3-540-68951-5_17.

U. Gargi, R. Kasturi, and S. Antani. Performance Characterization and Comparison of Video Indexing Algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pages 559–565, Santa Barnara, California, 1998.

L. J. Gogate and L. E. Bahrick. Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants. *Journal of Experimental Child Psychology*, 69(2):133–149, May 1998. ISSN 00220965. doi: 10.1006/jecp.1998.2438.

L. J. Gogate and L. E. Bahrick. Intersensory Redundancy and 7-Month-Old Infants' Memory for Arbitrary Syllable-Object Relations. *Infancy*, 2(2):219–231, 2001. doi: 10.1207/S15327078IN0202_7.

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*, 71(4):878–894, 2000. doi: 10.1111/1467-8624.00197.

I. Gori, U. Pattacini, F. Nori, G. Metta, and G. Sandini. DForC: a Real-Time Method for Reaching, Tracking and Obstacle Avoidance in Humanoid Robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems.*, Vilamoura, Algarve, 2012.

B. M. Hard. Reading the language of action. In *Proc. of the Annual Conference of the Cognitive Science Society*, 2006.

B. M. Hard, B. Tversky, and D. S. Lang. Making sense of abstract events: building event schemas. *Memory & cognition*, 34(6):1221–1235, Sept. 2006. ISSN 0090-502X. doi: 10.3758/BF03193267.

J. Hershey and J. Movellan. Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 813–819. MIT Press, 1999.

S. J. Hespos, M. M. Saylor, and S. R. Grossman. Infants' ability to parse continuous actions. *Developmental psychology*, 45(2):575–585, Mar. 2009. ISSN 0012-1649. doi: 10.1037/a0014145.

K. Hirsh-Pasek and R. M. Golinkoff. A Coalition Model of Language Comprehension. In *The Origins of Grammar: Evidence from Early Language Comprehension*. The MIT Press, 1999 edition, 1996. ISBN 0-262-58180-9.

G. Hollich, K. Hirsh-Pasek, R. Golinkoff, R. Brand, E. Brown, H. Chung, E. Hennon, and C. Rocroi. Breaking the language barrier: an emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65(3), 2000a. ISSN 0037-976X.

G. Hollich, K. Hirsh-Pasek, and R. M. Golinkoff. The Emergentist Coalition Model. *Monographs of the Society for Research in Child Development*, 65(3):17–29, June 2000b. ISSN 0037-976X. doi: 10.1111/1540-5834.00092.

J. E. Hunter. *Human action segmentation and recognition with a high dimensional single camera system*. Thesis (phd), Vanderbilt University, 2009.

L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, Mar. 2001. ISSN 1471-003X. doi: 10.1038/35058500.

L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998. ISSN 0162-8828. doi: 10.1109/34.730558.

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 and Applications*, 30(3):273–288, Sept. 2006. doi: 10.1007/s11042-006-0026-2.

A. Jesse and E. K. Johnson. Audiovisual alignment in child-directed speech facilitates word learning. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pages 101–106, Adelaide, Aust, 2008. Causal Productions.

P. Jusczyk. How infants begin to extract words from speech. *Trends in cognitive sciences*, 3(9):323–328, Sept. 1999. ISSN 1879-307X.

S. B. Kang and K. Ikeuchi. Toward automatic robot instruction from perception-recognizing a grasp from observation. *IEEE Transactions on Robotics and Automation*, 9(4):432–443, 1993. doi: 10.1109/70.246054.

K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon. Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL – 3.0, Sept. 1994.

Y. Kuniyoshi, M. Inaba, and H. Inoue. Learning by watching: extracting reusable task knowledge from visual observation of human performance. *IEEE Transactions on Robotics and Automation*, 10(6):799–822, 1994. ISSN 1042296X. doi: 10.1109/70.338535.

C. A. Kurby and J. M. Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, Feb. 2008. ISSN 1364-6613. doi: 10.1016/j.tics.2007.11.004.

F. Lömker, S. Wrede, M. Hanheide, and J. Fritsch. Building Modular Vision Systems with a Graphical Plugin Environment. In *IEEE International Conference on Computer Vision Systems*, pages 2–2, 2006. doi: 10.1109/ICVS.2006.18.

I. Lütkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter, and R. Haschke. The curious robot-structuring interactive robot learning. In *International Conference on Robotics and Automation*, pages 2154–2160, May 2009. ISBN 978-1-4244-2788-8.

W. Ma, R. M. Golinkoff, D. M. Houston, and K. Hirsh-Pasek. Word Learning in Infant- and Adult-Directed Speech. *Language Learning and Development*, 7(3):185–201, July 2011. doi: 10.1080/15475441.2011.579839.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. California, USA, University of California Press, 1967.

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

G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano. The iCub humanoid robot: an open-systems platform for research in cognitive development. *Neural networks : the official journal of the International Neural Network Society*, 23 (8-9):1125–1134, 2010. ISSN 1879-2782. doi: 10.1016/j.neunet.2010.08.010.

M. Meyer, P. Decamp, B. Hard, D. Baldwin, and D. Roy. Assessing Behavioral and Computational Approaches to Naturalistic Action Segmentation. In *Proc. of the 32rd Annual Conference of the Cognitive Science Society*, 2010.

M. Meyer, D. A. Baldwin, and K. Sage. Assessing Young Children's Hierarchical Action Segmentation. In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, 2011a.

M. Meyer, B. Hard, R. Brand, M. McGarvey, and D. Baldwin. Acoustic Packaging: Maternal Speech and Action Synchrony. *IEEE Transactions on Autonomous Mental Development*, 2011b. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2103941.

Y. Nagai and K. Rohlfing. Computational Analysis of Motionese Toward Scaffolding Robot Action Learning. *IEEE Transactions on Autonomous Mental Development*, 1 (1):44–54, Apr. 2009. doi: 10.1109/TAMD.2009.2021090.

Y. Nagai, C. Muhl, and K. Rohlfing. Toward designing a robot that learns actions from parental demonstrations. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3545–3550, 2008. doi: 10.1109/ROBOT.2008.4543753.

D. Newtson. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28–38, Oct. 1973.

M. Pardowitz, R. Haschke, J. J. Steil, and H. Ritter. Gestalt-Based Action Segmentation for Robot Task Learning. In *IEEE-RAS 7th International Conference on Humanoid Robots (HUMANOIDS)*, 2008.

U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini. An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1668–1674. IEEE, Oct. 2010. ISBN 978-1-4244-6674-0. doi: 10.1109/IROS.2010.5650851.

P. Peer, J. Kovac, and F. Solina. Human Skin Colour Clustering for Face Detection. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, pages 144–148, 2003.

M. Pereverzeva and D. Y. Teller. Infant color vision: Influence of surround chromaticity on spontaneous looking preferences. *Visual Neuroscience*, 21(3):389–395, Apr. 2004. ISSN 0952-5238. doi: 10.1017/S0952523804213086.

N. J. Pitchford and K. T. Mullen. The role of perception, language, and preference in the developmental acquisition of basic color terms. *Journal of experimental child psychology*, 90(4):275–302, Apr. 2005. ISSN 0022-0965. doi: 10.1016/j.jecp.2004.12.005.

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 *In Gesture and Speech in Interaction*, 2009.

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

J. Ramírez, J. M. Górriz, and J. C. Segura. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In M. Grimm and K. Kroschel, editors, *Robust Speech Recognition and Understanding*, pages 1–22. I-TECH Education and Publishing, June 2007. ISBN 978-3-902613-08-0.

K. Rapantzikos, G. Evangelopoulos, P. Maragos, and Y. Avrithis. An Audio-Visual Saliency Model for Movie Summarization. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 320–323. IEEE, 2007. ISBN 978-1-4244-1273-0. doi: 10.1109/MMSP.2007.4412882.

C. W. Robinson and V. M. Sloutsky. Auditory dominance and its change in the course of development. *Child development*, 75(5):1387–1401, 2004. ISSN 0009-3920. doi: 10.1111/j.1467-8624.2004.00747.x.

C. W. Robinson and V. M. Sloutsky. Development of cross-modal processing. *WIREs Cogni Sci*, 1(1):135–141, 2010. doi: 10.1002/wcs.12.

D. Robinson. Replay Gain - A proposed standard, 2011.

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*, 20 (10):1183–1199, 2006. ISSN 0169-1864. doi: 10.1163/156855306778522532.

M. Rolf, M. Hanheide, and K. Rohlfing. Attention via Synchrony: Making Use of Multimodal Cues in Social Learning. *IEEE Transactions on Autonomous Mental Development*, 1(1):55–67, Apr. 2009. doi: 10.1109/TAMD.2009.2021091.

O. Rosa Salva, T. Farroni, L. Regolin, G. Vallortigara, and M. H. Johnson. The evolution of social orienting: evidence from chicks (Gallus gallus) and human newborns. *PloS one*, 6(4):e18802, Jan. 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0018802.

A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg. Cross-Language Prominence Detection. In *6th International Conference on Speech Prosody*, Shanghai, China, 2012.

D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, and P. Gorniak. The Human Speechome Project Symbol Grounding and Beyond. In P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, editors, *Symbol Grounding and Beyond*, volume 4211 of *Lecture Notes in Computer Science*, chapter 15, pages 192–196. Springer, Berlin / Heidelberg, 2006. ISBN 978-3-540-45769-5. doi: 10.1007/11880172_15.

Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 1111–1118. IEEE Computer Society, Aug. 2000. doi: 10.1109/CVPR.2000.855807.

M. M. Saylor, D. A. Baldwin, J. A. Baird, and J. LaBounty. Infants' On-line Segmentation of Dynamic Human Action. *Journal of Cognition and Development*, 8(1):113–128, 2007. doi: 10.1080/15248370709336996.

S. Schaal. Is Imitation Learning the Route to Humanoid Robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

T. Schack and F. Mechsner. Representation of motor skills in human long-term memory. *Neuroscience letters*, 391(3):77–81, Jan. 2006. ISSN 0304-3940. doi: 10.1016/j.neulet.2005.10.009.

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 and Human-Robot Interaction Workshop*, Toyama, Japan, Oct. 2009a.

L. Schillingmann, B. Wrede, and K. J. Rohlfing. A Computational Model of Acoustic Packaging. *IEEE Transactions on Autonomous Mental Development*, 1(4):226–237, Dec. 2009b. ISSN 1943-0604. doi: 10.1109/TAMD.2009.2039135.

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

C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3, Aug. 2004. doi: 10.1109/ICPR.2004.1334462.

E. Selkirk. On Derived Domains in Sentence Phonology. *Phonology*, 3(1986):371–405, 1986. ISSN 02658062. doi: 10.1017/S0952675700000695.

P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer, 2002. ISBN 3540429883.

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

D. Y. Teller, A. Civan, and K. Bronson-Castain. Infants' spontaneous color preferences are not due to adult-like brightness variations. *Visual neuroscience*, 21(3):397–401, 2004. ISSN 0952-5238. doi: 10.1017/S0952523804213360.

M. Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82, Feb. 2001. ISSN 0936-5907. doi: 10.1515/cogl.2001.012.

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

A.-L. Vollmer, K. Pitsch, K. Lohan, J. Fritsch, K. Rohlfing, and B. Wrede. Developing feedback: How children of different age contribute to a tutoring interaction with adults. In *IEEE 9th International Conference on Development and Learning*, pages 76–81. CoR-Lab., Bielefeld Univ., Bielefeld, Germany, IEEE, Aug. 2010. ISBN 978-1-4244-6900-0. doi: 10.1109/DEVLRN.2010.5578863.

L. Wagner and L. Lakusta. Using Language to Navigate the Infant Mind. *Perspectives on Psychological Science*, 4(2):177–184, Mar. 2009. ISSN 1745-6916. doi: 10.1111/j.1745-6924.2009.01117.x.

D. Wang, L. Lu, and H.-J. Zhang. Speech segmentation without speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*., volume 1, pages 468–471. IEEE, 2003. ISBN 0-7803-7663-3. doi: 10.1109/ICASSP.2003.1198819.

J. F. Werker, L. B. Cohen, V. L. Lloyd, M. Casasola, and C. L. Stager. Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, 34(6):1289–1309, 1998.

W. Wolf. Key frame selection by motion analysis. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 1228–1231 vol. 2, Aug. 2002. doi: 10.1109/ICASSP.1996.543588.

B. Wrede, S. Kopp, K. Rohlfing, M. Lohse, and C. Muhl. Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*, 42(9):2369–2384, Sept. 2010. ISSN 03782166. doi: 10.1016/j.pragma.2010.01.003.

J. Zacks, B. Tversky, and G. Iyer. Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology. General*, 130(1):29–58, Mar. 2001. ISSN 0096-3445.

J. M. Zacks and K. M. Swallow. Event Segmentation. *Current Directions in Psychological Science*, 16(2):80–84, Apr. 2007. ISSN 0963-7214. doi: 10.1111/j.1467-8721.2007.00480.x.

J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127:3–21, 2001.

J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273–293, Mar. 2007. ISSN 0033-2909. doi: 10.1037/0033-2909.133.2.273.

J. M. Zacks, S. Kumar, R. A. Abrams, and R. Mehta. Using movement and intentions to understand human activity. *Cognition*, 112(2):201–216, Aug. 2009. ISSN 00100277. doi: 10.1016/j.cognition.2009.03.007.

D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Multimodal group action clustering in meetings. In *Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks - VSSN '04*, pages 54–62, New York, New York, USA, 2004. ACM Press. ISBN 1581139349. doi: 10.1145/1026799.1026810.

S. Zhang and F. Stentiford. A saliency based object tracking method. In *Sixth International Workshop on Content-Based Multimedia Indexing*, Oct. 2008.

Y. Zhang and J. Weng. Conjunctive visual and auditory development via real-time dialogue. In *in Proc. 3rd International Workshop on Epigenetic Robotics*, pages 974–980, Boston, MA, Aug. 2003.

P. Zukow-Goldring. Sensitive Caregiving Fosters the Comprehension of Speech: When Gestures Speak Louder than Words. *Early Development and Parenting*, 5(4):195–211, 1996. ISSN 1099-0917. doi: 10.1002/(SICI)1099-0917(199612)5:4⟨195::AID-EDP133⟩3.0.CO;2-H.

P. Zukow-Goldring. Assisted imitation: affordances, effectivities, and the mirror system in early language development. In M. Arbib, editor, *From Action to Language*, pages 469–500. Cambridge University Press, 2006.