Bielefeld University
Faculty of Technology
Applied Computer Science

Statistical analysis of characteristics of Infant-Directed-Interaction with respect to action structure

Master's Thesis
November 2008

Annika Peters
apeters@techfak.Uni-Bielefeld.DE

Thesis Advisors :
Dipl. Inform. Lars Schillingmann
Dr. Ing. Britta Wrede

Acknowledgements

There are many people I would like to thank for their moral and/or scientific support and contribution during this Master's thesis. I would like to thank Horst and Gabi Doelle, John and Dagmar Neate, Angela Manneck-Peters, Christian Peters, Miriam Semin, Angela Grimminger, Anna-Lisa Vollmer, Patrick Holthaus, Peter Nicholson and finally my supervisors Lars Schillingmann and Britta Wrede.

Abstract

Acoustic packaging, intersensory redundancy, motherese and motionese are the main concepts, which describe how acoustic features and body movements of caretakers are adapted intuitively in adult-child-interaction. This thesis will approach the question, whether feature characteristics highlight certain parts of child-directed actions. The main part focuses on the statistical examination of four features, Eye-Gaze, Intonation, Word, Velocity of Hand Motion of child-directed demonstrations. The auditory and visual signals of the recorded toy demonstrations were divided into meaningful segments for this examination. Then the composition of each feature and its values (fc) in each segment is analysed. A conditional probability model $P(Segment_x|fc_j)$ is computed to answer the question: in which segment ($Segment_x$) is, an already appeared feature characteristic fc_j, likely to occur? Patterns of high conditional probability values have been identified for distinct feature characteristics and provide a basis for a future automatically classification.

Contents

Chapter 1

Introduction

1.1 Child-Directed-Communication

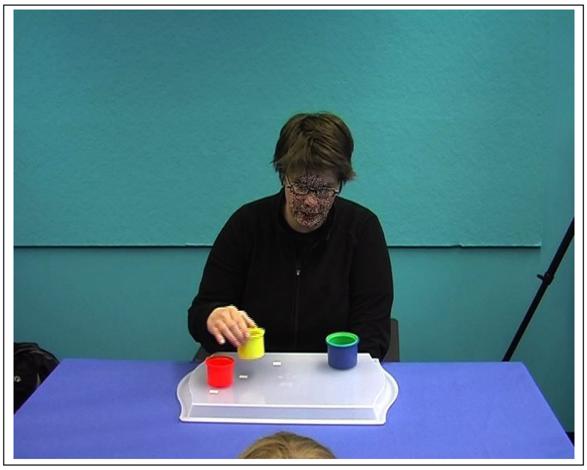

Figure 1.1: *A mother demonstrates to her child, see box below for transcript.*

> *"Look, Mandy, look – cups – We have different sizes* `[silence]`. *See?* `[waves hand, shows object]`. *Put the green one in the blue one,* `[clacking noise]` *– the yellow one in the green one,* `[clacking noise, looks at child and hesitates for a moment]` *and the red,* `[waving cup fast]` *and the red one in the yellow one and they are all gone – they are all gone – did you see it?* `[hands are laid down, observe child, no movements]`."

This is an extract of a transcript made of a tutoring situation. A parent had to explain to her/his 10 month old child, how different sized cups can be stacked together. I have added some descriptions in square brackets to give the reader some notion of how the parent spoke to and acted towards her/his child. These descriptions are not complete, because more changes in prosody (e.g., intonation) and body movements (e.g., velocity of hands) are observable.

Tutoring is a special form of communication. Especially in Parent-Child-Interaction: the parent behaves in a special way when communicating with her/his child compared to adult-adult-interaction. Not only characteristics of parental speech and motion [Snow 77; Bran 02] are adapted to the infant's stage of development, but also *facial expressions* [Chon 03] and gestures are different [Iver 99]. Masataka found evidence for child-directed sign language in deaf mothers [Masa 92]. The above named researchers and many more identified a number of parameters, which are very characteristic for infant-directed (ID) communication. Entire actions towards children have recently been investigated for a complete picture of the interplay of these characteristics. This picture is not complete and research needs to be done in identifying the exact pattern of features over the course of an ID- communicative situation.

1.1.1 Motherese and Motionese

Snow and Ferguson [Snow 77] report that parents simplify the grammatical structure of utterances and repeat sentences when speaking to a child. Prosodic features of speech are more exaggerated. Brand looked upon movements in child-directed demonstrations and identified some parameters, which are modified [Bran 02]. Not only speech, but also motions can be simplified and more repetitive. She found that the rate of hand-movements was slower, and more abrupt and more expansive compared to adult-adult-demonstrations. Additionally researchers found that infants prefer child-directed-speech (*motherese*) and child-directed-motion (*motionese*) over adult-directed-interaction [Kote 06].

Various researchers conclude that an infant directed input directs attention to the relevant information and therefore, language and movements are more easily learned [Bran 02; Bahr 00; Goga 00]. But as Gogate *et al.* points out, infant-directed communication is as seldom unimodal as it is in face-to-face communication between adults. It is a very complex and broad topic.

The entire sensory system perceives information about the dialogue partner. Communication is so much more than speech and simple movements. It encompasses more than simply focusing on visual and audible signals: variations in intonation, pitch, intensity and gestures, facial expression and eye-gaze are recognisable. Cynthia Breazeal pointed out that social robots, as well as, humans need additional information to interpret ambiguous information in a dialogue. Dialogue partners need to establish common ground in order to recognise each others' intentions and theories of mind. To communicate and understand verbal and nonverbal signals is a very challenging task for social robotics according to Cynthia Breazeal [Brea 08] and of course for babies, too.

Infants may, however, have an advantage because child-directed communication is simpler, prosodic and visual features are more exaggerated, which makes certain sensory input more salient and intentions are therefore easier to recognise and patterns in the information stream can be more easily identified.

1.1.2 Interplay of Modalities

There was a clear separation in child-directed research between that of different modalities, mainly auditory (speech) and visual (motion). Only recently has research been oriented towards the interplay of amodal input.

Bahrick and coworkers found another way of catching attention. They theorise that *intersensory redundancy* makes the given information more salient in a constant stream of arbitrary information. She states that young infants prefer temporal synchronous and amodal stimuli over asynchronous stimuli. For example, to hear somebody speak in the same rhythm and to see her/his movements of lips in the same rhythm. Bahrick *et al.* explain that temporal synchronous, redundant and amodal information can also direct attention to meaningful events, which may be beneficial for learning [Bahr 00].

In addition, Gogate and coworkers state that naming an object and touching or moving or rather showing it, is redundant information across the auditory and visual senses and therefore, makes the relation between the name and the object salient. They call child-directed-speech plus co-occurring and redundant information at another modality *multimodal motherese* [Goga 00].

Coming from the perspective of infant-directed-motion research, Brand and Tapscott asked themselves the question, whether utterances (they call it narrations) facilitate the perception of actions in infants. They carried on the idea of *acoustic packaging* from Hirsh-Pasek and Golingkoff (1996) who pointed out that co-occurring sound might emphasise segments within an action [Bran 07b]. Brand *et al.* found out that co-occurring infant-directed-speech and motion can help infants (9.5-11.5 months) to segment the action into smaller units [Bran 07b]. It is assumed that *acoustic packaging* emphasises the importance of visual modality compared to multimodal motherese. The characteristics of child-directed communication and their benefit are summed up in picture 1.2.

The preferential looking paradigm is a well established and often used method within research in child-directed input with very young children (0-13 month and even older [Bran 07b]). In order to focus more on the entire child-directed communication it is necessary to engage the parents in an everyday and communicative situation with their child - a tutoring situation, for example demonstration of toys [Goga 00; Bran 07b].

The next section takes a closer look at the tutoring situation, which provided the data for my analysis of child-directed action. It is hoped to find out more about how and when exactly parents structure their action towards their children. Furthermore it will be defined what an action is and how the term will be understood in regard of the present data.

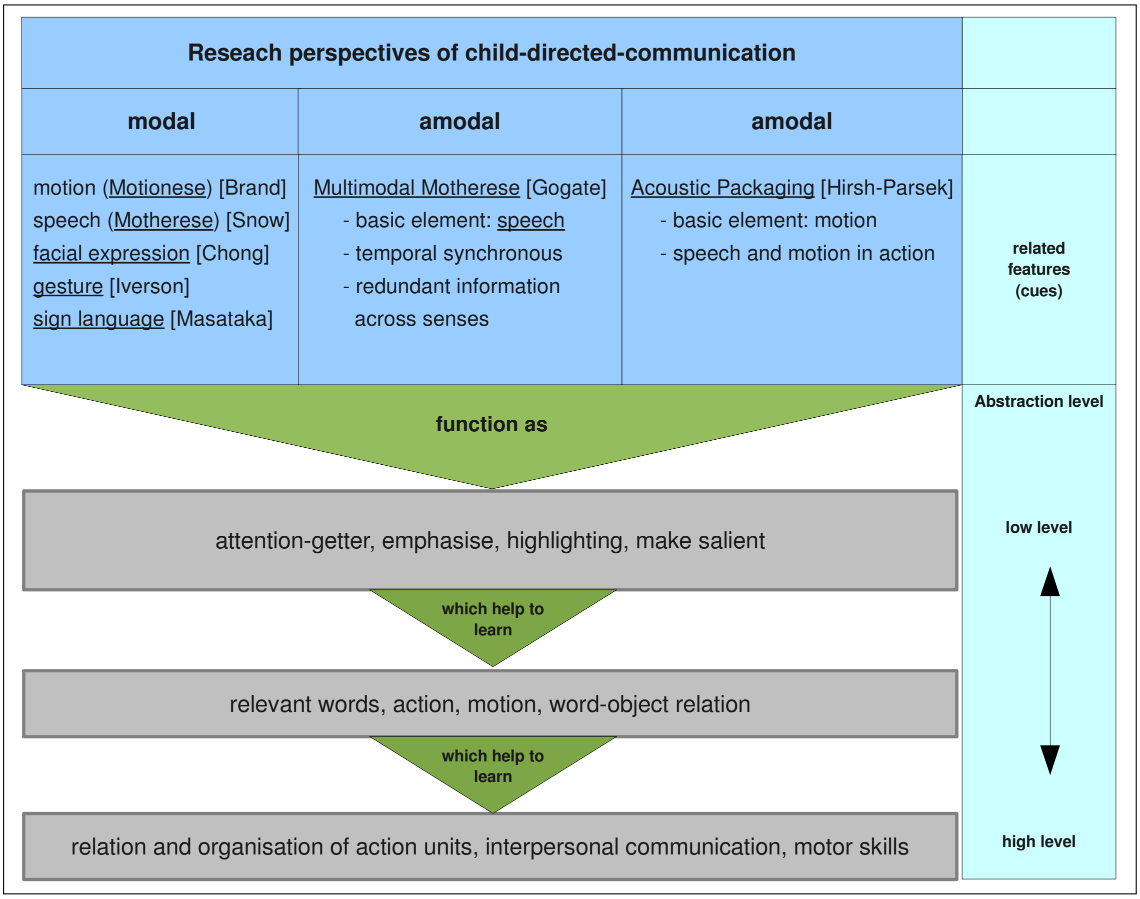
Reseach perspectives of child-directed-communication
modal
amodal
amodal
motion (Motionese) [Brand]
speech (Motherese) [Snow]
facial expression [Chong]
gesture [Iverson]
sign language [Masataka]
Multimodal Motherese [Gogate]
- basic element: speech
- temporal synchronous
- redundant information across senses
Acoustic Packaging [Hirsh-Parsek]
- basic element: motion
- speech and motion in action
related features (cues)
function as
Abstraction level
attention-getter, emphasise, highlighting, make salient
low level
which help to learn
relevant words, action, motion, word-object relation
which help to learn
relation and organisation of action units, interpersonal communication, motor skills
high level

Figure 1.2: *Research perspective of child-directed communication (CDC). Modal approaches concentrate only on one feature (cue). The other two approaches consider input-signals over more than one sense, however, with different features as a basis (speech, motion). Modal, multimodal and amodal cues of CDC function similar on ways to help children learn and cope with the world on various levels.*

Chapter 2

Description of Data

2.1 The Demonstration of Toys

This thesis focuses on the topic of a special tutoring situation, in which parents had to demonstrate to their 8-11 months old children different toy settings.

During the experiments the mother or the father sat at a table opposite their child, who sat in a highchair. Both were videotaped separately but were audio taped together. In this thesis audio and video data of the parent only will be taken into account.

The video- and audio taped experiments were part of a bigger "Motionese" project, which was carried out by the Applied Computer Science Group and Faculty of Psychology at Bielefeld University [Rohl 06], for a detailed overview please see earlier reports [West 08; Pete 08]. I chose these two experimental settings of the Motionese project because the structure of the task is comparable, but the toys were different. Both tasks consist of three toys, which the parents had to manipulate in order to reach the expected goal:

- Three different sized cups had to be inserted into a fourth cup (for picture see Figure 1.1).
- Three different shaped toy blocks had to be put next to each other (for picture see Figure 2.1)

Situations, in which movements play a crucial role (e.g., demonstration and showing)

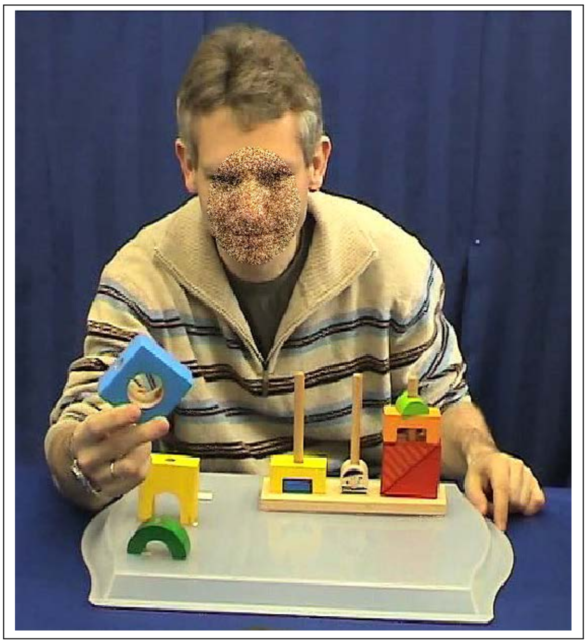

Figure 2.1: *A father demonstrates to his child, how to stack toy building blocks.*

are called *action* in the literature. Unfortunately there is no clear separation between action and motion. A definition for action is needed for this thesis.

2.2 Definition of Action

The definition of action is a very old topic of philosophical discussion. Aristotle defined in his action theory that action is a process of wilful bodily movement [Wiki 08].

At the 2008 Workshop "Intermodal Action Structuring" researchers agreed upon the same definition of action, (Gisa Aschersleben, Annette Baumgärtner and Britta Wrede) [Rohl 08]:

$$\boxed{\textbf{Action} \Rightarrow \textbf{goal directed motion}}$$

The general definition does not seem to be the problem but the interpretation of the beginnings and the end of a goal-directed-motion is ambiguous. The starting point of an action is lifting-up-an-object-motion, according to Brand and coworkers latest experiments, at which mothers had to show toys to their infants [Bran 07a]. The initial movement of the hand towards the toy and grasping the toy are not counted as goal-directed-motions, as was stated by Brand at the workshop [Rohl 08].

Gogate and coworkers use *moving objects, showing object, gestures and action* as synonyms in their paper "A study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures" [Goga 00]. This still conforms to the definition of action, since showing, moving objects and gestures can serve a purpose and are therefore goal-directed. But her definition of initial and end points of an action is different to the definition of Brand *et al.*. For Gogate *et al.* *target words* (names of toys and verbs of the toy's ability: swimming) served as the definition of starting and end of an action as long as they occurred in the same time span [Goga 00].

2.2.1 Hierarchical Action

Brand *et al.* define that action can exist in a hierarchical order. One action (action of demonstration or task) can be split up into smaller actions [Wred 05; Bran 07b]. If this is the case, then the goal is split up as well, for instance the entire goal of the demonstration: "put all cups together" is split up into three minor goals (3 toys), "doing something with each toy". These sub goals can again be divided into even smaller units, Brand and Tapscott call these sub actions, *events*, which include for example, pointing, inserting etc.. Brand *et al.* suggest that other modalities for instance, speech could "highlight at a variety of levels" [Bran 07b]. That means other modalities can be divided to fine-grained levels for analysis (utterances-> words -> intonation).

Seeing the action as defined above as part of a demonstration, it nevertheless provides visual and audible sensory input for the child. A variety of features can be extracted from these signals, which can be hierarchical (word-intonation). The characteristics of a feature describe it precisely (Intonation = falling, rising). A set of features then describes the action.

2.3 My Definition of Action

In my opinion the definition of action depends on the way one looks upon the concept of action.

There is the unimodal perspective, where the action could only be a goal-directed motion. But as mentioned before the interaction between modalities is important in communication. Speech can be goal-directed, too. In my opinion goal-directed speech could be defined as an action as well, especially when utterances are comments on one's own goal-directed movements and therefore very close to the movement itself. For instance, the sentence: “I put this building block on top” describes the movement “put” and the goal “on top”. Buccino and coworkers report that mirror neurons in monkeys are activated not only by viewing or performing an action but also when listening to a sound. Buccino *et al.* follow up this finding and were able to show that processing hand- and foot-action-related sentences involve the motor system as well, because they did not find the same responses when listening or observing a goal-directed motion, Buccino *et al.* assume that listening to an action-related sentence evokes representation of the action at a higher level. Observing and executing a goal-directed motion reveals more details about how exactly the motion is performed, listening to a sentence like “he took the cup” lacks that information but still discharges the same area of the motor system [Bucc 05].

The other way to look at an action is the bimodal or multimodal way where, as in the case of acoustic packaging, goal-directed movements and speech interplay serve one goal: to facilitate the organisation of an action unit. In Brand's and Tapscott's study the question remaines unanswered, whether child-directed speech needs to be goal-directed or whether characteristics of motherese only are enough to highlight the beginning and end of an action, for example intonational contours [Bran 07b]. Gogate *et al.* even pay only attention to (target) words, which occurred in parents' utterances while they were teaching toy names and the toy's purpose. However, both bimodal approaches have in common that they attribute the bimodal features' co-occurrence to play a role in infant learning.

The data for my examination comes from an auditory and visual input. The parent's demonstrations are bimodal, because mostly they commented their actions (object manipulation). In short, in regard of the presented theory, the data could be described as *acoustically packaged action.*

To make my definition complete, I define starting points in regard of goal-directed-motion at the first body movement towards a goal, for example if the goal is: "inserting one cup into another", then the action starts at the first movement of the hand towards the cup (or building block) and stops when the cup is inserted and the hand does not touch the cup anymore and/or the hand engages in another movement or action.

The starting point in speech is the beginning of the first word, when talking about another action: "*[start]* Put the red one into the yellow one *[end]*", please compare the transcript (box) of the experiment in Chapter 1 and for more detailed information about the segmentation of action please read Chapter 5.1.

For an overview of features, which are found in the sensory input of an action please see picture 2.2.

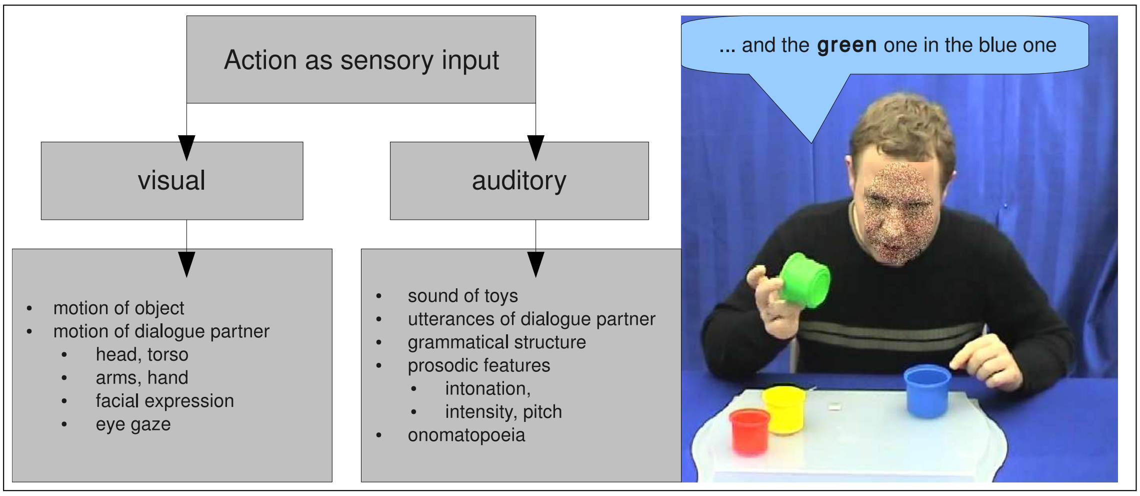

Figure 2.2: *From sensory input to an extraction of features. There are examples for visual and auditory input signals, because they are the basis for the later statistical examination.*

As stated before, research has identified specific features, which are adapted in child-directed communication. Features might reveal the structure of the entire demonstration through features specific changes to infant-directed input. For this examination, it will be necessary to look at selected features separately (in a unimodal way), to better work out the gain of each of them and to be able to combine them freely in future work and my interpretation. In the next chapter features for my analysis will be introduced. Following on from that, the goal and motivation of this thesis will be explained on the basis of the theory in chapter 1 and 2.

Chapter 3

Features and their characteristics

Features, dimensions or parameters are measures, with which a situation can be described. But the type of the situation can only be determined with the characteristics of the measures. For instance if Brand's parameter "proximity to the partner" has the value "very close", then the situation is very likely to be child-directed, especially when the value for adult-directed is the opposite, "very far".

I chose four features; two were extracted from the audio signal and two from the visual signal:

audible utterances of the parent as single Words and ratings of Intonation

visible Eye-Gazes in different directions and Velocity of Hand Movements

All characteristics of features have starting and end points, so the time of their occurrence can be related across the features. Those intervals have different lengths, but their time format is the same. The time is measured in seconds with three decimal places (for millisecond precise output). For example the word "hello" could occur between 20.111 s to 21.342 s in one experiment. All features and their values were extracted offline.

3.1 Word

origin A human transcribed the audio signal. She/he also assigned starting and end points only to meaningful segments (for examples please see box at Chapter 1[1]). The boundaries of the words had to be computed by a standard speech recognition method, forced alignment[2] [Holt 99]. In some cases the forced alignment did not work, therefore start and ends were computed by dividing the wrapping boundary by the number of words.

characteristics each unique word counts for one outcome/characteristic, for e.g. *look, cup, put, the, and ....* Only names of the children (short: ndk=name of child), declensions[3] of adjectives of colour, different spellings of transcribed words[4] were summed up

cause and prognosis Multimodal motherese and especially acoustic packaging emphasise the importance of words within the learning of word-object relation and recognising action boundaries. Distinct utterances or even words, which coincide with action, may help highlight boundaries of action units [Bran 07b]. For example the word "look" could occur at the beginning of an action. It will be interesting to see if characteristics of the feature *Word*, which are not classified into meaningful categories (e.g., categories for colour, object names, attention-getters, verbs of movement), will still reveal certain groups by themselves by occurring at the same time (part of the demonstration) over all subjects and experiment settings, for instance prepositions like "onto, into[5]" and all sorts of conjunctions like "and", "then".

[1] Meaningful segments are parts not divided by square brackets or dashes, see box at Chapter 1

[2] Lars Schillingmann saw to it, thank you

[3] Declension is typical for the German language e.g. gruene, gruenes, gruener, gruenen, ...

[4] Different spellings of the word are due to the dialect/common speech and clipped speech

[5] German: rein und drauf

3.2 Intonation

origin Originally two human evaluators had to decide whether the intonation of the end of a meaningful segment (see Section 3.1) was *rising, falling, continuing* or *unsure*. These two evaluators agree only in 60% of the cases. The rate "rising" does not seem to be the problem, but sorting segments into "falling" and "continuing" seemed to be difficult for evaluators, according to Dominik Westhues in his master's thesis on this topic [West 08]. For this reason I took the ratings of one of the evaluators[6]. Naturally there need to be more evaluators for a statistical significant interpretation.

characteristics Each segment has starting and end points. Originally these ratings were used for a different study. Therefore, each rating contains approximately three words [West 08].

cause and prognosis Brand *et al.* as well as Wrede and coworkers suggest that intonational contours useful for packaging units of action, especially "the falling intonation may mark a boundary of a completed action" (p.323 [Bran 07b]) [Wred 05]. They question the need for words and their meaning when prosodic features like intonation in motherese are so distinct.

3.3 Eye-Gaze

origin Student assistants[7] coded manually from the video stream the gaze direction of the parent into four classes. The videotape displays only the parent, not the infant; therefore the coders had been only able to annotate the direction of the view, not the actual eye contact.

characteristics The gaze-directions are towards the *object, infant, instruction, miscellaneous*. As characteristics before, each outcome has interval starts and ends.

[6] Dominik Westhues calls the evaluator A

[7] Angela Grimminger and Miriam Semin thanks for annotating

cause and prognosis Striano and Stahl found that adult gaze towards an object and the infant in exchange (2-3 seconds) may arise or maintain the infant's attention towards the object, (triadic attention from 3 months on) [Stri 05; Stri 07]. Moreover, *Interactiveness* is one of Brand's dimensions of motionese, eye-gaze bouts are one of its quantifiable measures/characteristics. Brand *et al.* consider eye gaze a global feature, which is there to draw the infant's attention in the joint attention manner [Bran 02; Bran 07c]. An equivalent number of eye-gaze direction changes are expected mainly between object and infant during the demonstration.

3.4 Velocity-of-Hand-Movements

origin Velocity of Hand Motion of the parent was computed from hand-trajectories. Hand-trajectories for each hand were extracted in a semi-automatic way from the video stream. A tracker[8] for each hand computed the x- and y-coordinates per hand and per frame of the video image (every 25 Hz or 40 ms). A student assistant[9] had to control the tracker from time to time. The velocity of each hand movement was computed (velocity value per 40 ms) and smoothed. In order to sum up the movements of both hands the velocities were added as both hands were at no time moved together. The velocity of both hands was divided into five different speed categories with global thresholds. These thresholds were carefully chosen by statistical survey of the velocity data from each person (quartiles, histograms, kmeans).

characteristics The velocity data for each person is annotated with categories of *no, slow, medium, fast* and *very fast* motion of hands and same and successive categories are summed up to bigger intervals. There are no time breaks between the intervals, which mean that the starting point of the current interval is the ending point of the interval before.

[8] Thanks to Philipp Gebert for developing the tracker

[9] Thanks to Miriam Semin

cause and prognosis The computation of the feature ”Velocity of Hand Motion“ differs from Brand’s coding parameter ”rate“. Coders had to valuate the velocity of Hand Movements for the entire experiment duration per subject. Brand and coworkers assume that this is one reason why the difference between infant-directed-action and adult-directed action did not vary significantly. Child-directed motion has merely a tendency toward faster movements [Bran 02]. Consistent with this discovery findings of Rohlfing and coworkers in a similar study [Rohl 06]. The outcome of those studies does not mean that there is not a pattern or regularity hidden in the speed of hand-movements over the time of a demonstration. Brand *et al.* calculated only one mean value (global rating) per subject, the present Velocity data contains on average, 85 intervals per subject per demonstration[10]. If the same speed categories can always be associated with the same part in a demonstration, this could also reveal more about the Motionese dimensions of *smoothness* or *punctuation* of the hand movements [Bran 02; Rohl 06]. These are said to be abrupt and sharp in child-directed-motions.

[10] This means 85 intervals consist of a mixture of the Velocity characteristics, for instance slow, medium, fast, medium, fast and so on

Chapter 4

Goal and Motivation

In regard to the present research from the computational point of view, it is incredible, how well infants cope with this huge amount of information, which is provided by the environment and needs to be processed through various senses. Infants have to learn to understand language and movements from the beginning, similar to robots.

Current research on infants directed input illustrates that adults intuitively help infants to cope with the world by using various means. A lot of features have been detected, which characterise motherese and motionese, for example proximity to the partner, pace, simple and short utterances, variations in pitch, intensity and intonation, rate of hand movements abd eye-gaze-directions [Goga 00; Bahr 00; Bran 02; Wred 05; Rohl 05]. Some have been made quantifiable by counting occurrences within interaction periods [Bran 07c]. Little research, however, has been done on identifying pattern intra and inter feature wide and pairing occurrences to part of an action (e.g. beginning, middle and end of an interaction period).

Brand and coworkers propose that a "fine-grained analysis of characteristics of Motionese" is very fruitful, if characteristics like *interactiveness* are evaluated by "quantifiable sub-features" (sub-features of interactiveness are eye-gaze bouts). All features, therefore, should be computed to avoid dependence on the opinion on human elevators

[Bran 07c].

Quantifiable features are of great interest in robotics, as a distinct pattern can reveal, in which state of an action a distinct feature is situated, for example the word: “look at” could have a pattern: “look at is always heard at the beginning of a tutoring situation”. This pattern or rather prediction can in turn help to direct the information processing unit of a robot to a special signal and its interpretation.

The next step could be predicting the communicative partner’s following action or intention. For example, this person said “look at” and now he is going to show the use of the object, which is somehow important to learn.

The general motivation in regard of the above lookout on robotics is to find dependencies between a feature and the corresponding part of an action. This is so attractive for robotics and child development because dependencies can reveal, which feature enhances, what part of the structure of an action. This piece of information can help to uncover, how exactly a structured action is perceived and processed by the infant [Naga 08], which might help to build or evaluate models of how children acquire social skills (action skills). This work follows up Brand’s example of a fine-grained analysis from a computational point of view and the approach of Wrede to break down actions into subactions or even smaller parts [Wred 05] and a new approach of regarding values of features with their corresponding time of occurrence.

Acoustic packaging, intersensory redundancy, motherese and motionese are the main concepts, which describe how acoustic parameters and body movements of caretakers are adapted intuitively in adult-child-interaction. This thesis will approach the question: *Do feature characteristics highlight certain parts of an action and on which level of the action hierarchy?*

The following chapters will provide an insight into the method I used and the patterns, which occurred during a demonstration of toys. The course of demonstration will be analysed with regard to the four above presented features (see Chapter 3). The chances[1]

[1] Conditional Probability = P(Part x of a demonstration | characteristic j of feature i)

of each part of a demonstration, given that there is a distinct characteristic of one feature, will be computed to provide a first overview of the information features. This will lead to the more precise question of whether these conditional probabilities reveal anything about the structure of an action. The probabilities will be analysed in regard of patterns and discussed.

Before conditional probabilities can be computed the demonstration needs to be divided into meaningful time periods (later: segments) to examine the feature characteristics behaviour in different states of the demonstration. The next chapter explains how the segmentation is done. Then, in the following chapter the mathematical background is described.

Chapter 5

Segmenting the Task

In order to analyse the behaviour of features and their characteristics at certain stages within the course of the demonstration task, this approach needs to determine meaningful segments with prior knowledge. It is essential to know where an *action* segment starts or ends in order to examine features within these segments and to build hypotheses about any pattern like “the intonation contour is always falling towards the *End* of an *action*”.

The difficulty will be balancing the level of partition of the action or rather demonstration task and the resulting information gain. Fortunately, the experiments my analysis is based on, have a natural structure.

5.1 Hierarchy of Demonstration Task

The task (or to be more precise, the recorded data of one experiment) needs to be segmented into smaller parts in order to examine features within an action. There is a certain sequence or structure of events, which is very similar for each parent in the study. Out of 55 subjects, 52 said or did something before starting the demonstration, all parents demonstrated the toys and 48 mothers or fathers did or said something after

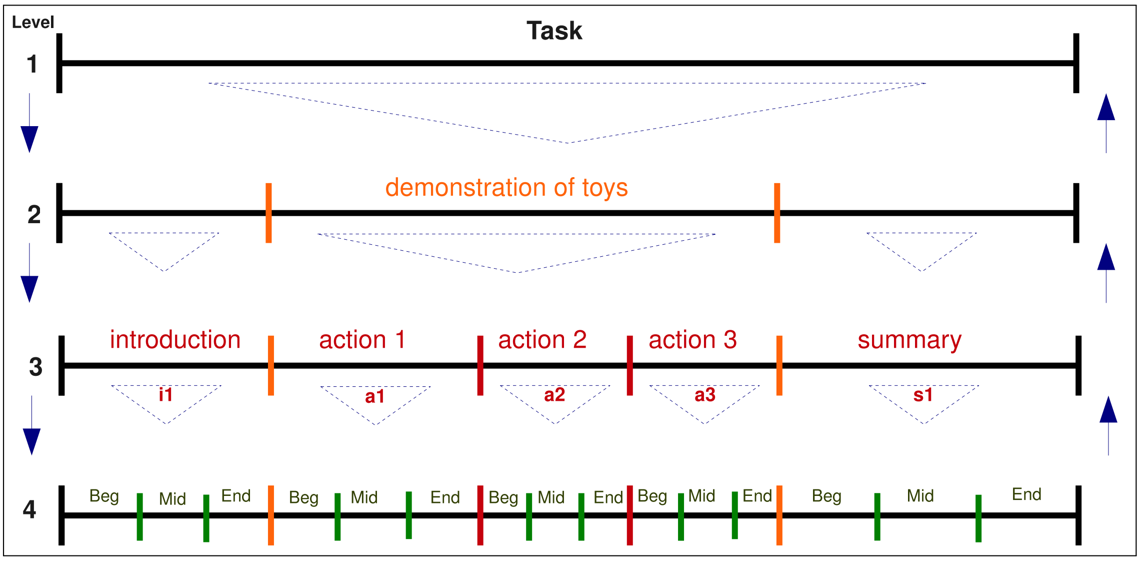

Figure 5.1: *Partition of the "task". The task (experiment) can be divided into a demonstration segment and two bounding segments (level 2). Those segments which occur before and after the demonstration segment, are called "introduction" and summary in the next level of partition (level 3). In the third level, the demonstration is divided into its three action segments. In the highest partition level (and the level of examination), each of the segments are divided into a "Begin", "Middle" and "End" part (segment) (level 4).*

achieving the goal. This was due to the job the parents had to fulfil and the comparable toy settings the parents attended to (please see Chapter 2.1 for description of task). Figure 5.1 level 2-4 represent the partition of the (demonstration) task.

Most of the time the parents gave themselves some time to familiarise themselves with the toys and the task, listening to the experimenter's instruction and either reciting the instruction or addressing the child directly, giving him/her a short overview, for example "do you know these toys, we have them at home" or touching the toys. In short, everything that happens before the demonstration of toys is referred to as *introduction* (short: i1/in), (please consult Figure 5.1 level 2 and 3). Accordingly, everything that happens after the demonstration of toys is called *summary* (short: s1/su). After fulfilling the goal, for example, the parent would summarise verbally what she/he just did, repeat the task or give the toys to their infant. A variety of actions followed, which will be not considered in the computation because they are too variable, (please see Section 5.1.1 for more details).

The demonstration of the objects can be split up into three separate *actions* (short: a1, a2, a3) due to the fact that each experiment setting contained similar toys the parents had to use. This means parents attended to one object in each action and manipulated the toy in order to reach the goal.

I divided the actions as well as the introduction and summary into three equal parts, *Beginning, Middle* and *End* for a more fine-grained analysis, (please see Figure 5.1 level 4).

With this trisection I hope to acquire results, which are comparable to other assumptions and findings made on action structuring. Yukie Nagai *et al.* worked on the cup experiments of the motionese-project, but they analysed the demonstration as one action. They examined the behaviour of visually salient objects *before, during* and *after* the demonstration [Naga 07a]. Since they started the action with "grasping the cup", findings like salient eye-contact before the action should be found in the *Beginning* segment of *action 1* or even before this segment (*i1Beg-i1End*), please refer to Figure 5.1.

Brand *et al.* thinks that features might exhibit distinct patterns after *completion points* or before *initiation points* of an action. Wrede and coworkers suppose that the end of each action might be highlighted through prosodic features because it helps breaking up an entire task and therefore makes the communicated structure easier to comprehend.

The last partition (*Beg, Mid, End*) results in 1-2 seconds per interval. Making the intervals even smaller, would result in a number of empty intervals for the *Intonation* feature, which is only available in intervals of up to 3 seconds (mean duration of Intonation intervals 1-2 s). In addition Word- and Intonation- intervals cannot be divided because they have already the smallest partition level available.

Another approach would be to divide the actions into subactions like grasping, showing, inserting, and not in segments *Beg-End*, but this would be too fine-grained for my analysis and would not serve to analyse the structure on action levels.

I defined an action as either a goal-directed motion or a goal- directed speech. The focus lies on the communicated structure of action. Therefore, the task will be divided on the basis of the audio signal, into the five segments, see Figure 5.1 level 3. To control is, if these audio annotated boundaries are in any way temporally associated with the boundaries of the video signal. A partition is made accordingly on the basis of the video signal.

The following section will explain precisely the rules of segmenting the task separately by both signals. Furthermore, the results, of how well the action annotation of audio and video signal fit together, will be summed up here since it is not the main part of the analysis.

5.1.1 Action Segmenting

The goal was to determine the five time periods as illustrated and described in Section 5.1. Keeping in mind that the audio and video signal of the task are continuous and provide ongoing data, the task was gapless separated, thereby acquiring it a complete overview of feature occurrences. This implies that the beginning of a segment is the ending point of the previous segment, as displayed in Figure 5.2 and Figure 5.3. The examples (no.1-no.4) in Figure 5.2 and 5.3 provide not only an example of the similarity of the length of the action intervals and onset, but also the variation in length of segment *summary* and with that the total length of the experiments.

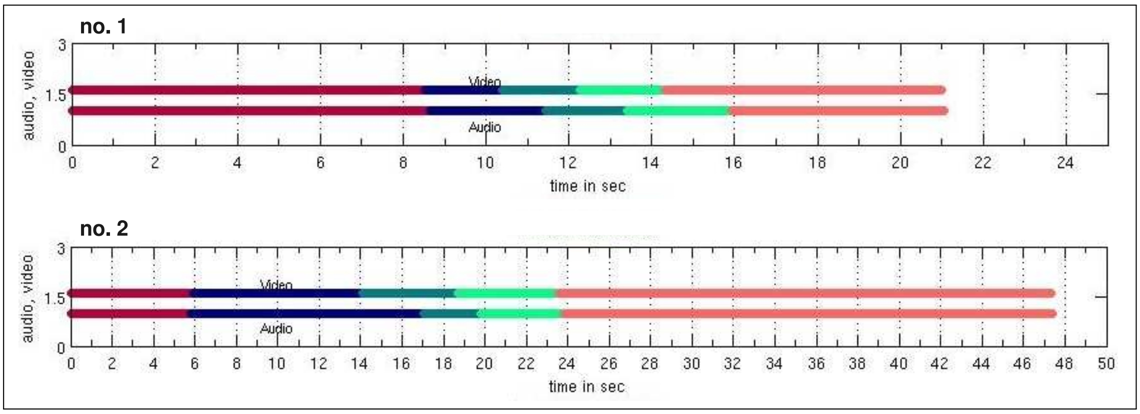

Figure 5.2: *An example of a "Cup" – task segmentation by video and audio signal for two independent subjects (no.1-2). The colours indicate when segments of level 3 start and end. The first bar is the segmentation, which is made on the basis of the video signal. The second bar represents the segmentation of the audio signal. (i1=magenta, a1=dark blue, a2=grey blue, a3=turquoise, s1=rose)*

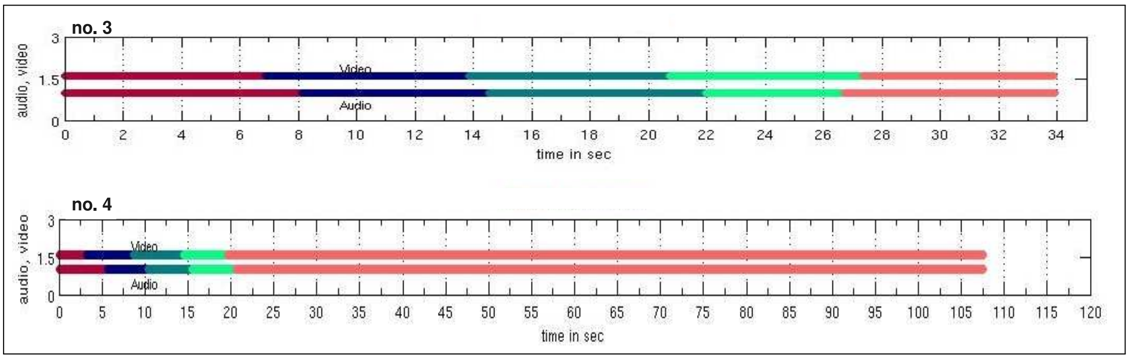

Figure 5.3: *An example of a "Minihausen" – task segmentation by video and audio signal. The diagrams represent segmentations for two subjects with different lengths in the "summary" segment. (i1=magenta, a1=dark blue, a2=grey blue, a3=turquoise, s1=rose)*

The definition of action in Chapter 2.2 and 2.3 forms a basis for the separation of the task up to level 3, (figure 5.1). The exact rules for separation are of course very similar to each other. Note, that start and end are the same and therefore there is only the need for one rule for each signal (see figure 5.4 for rules and examples).

- **audio beginning**
  1. semantics and speech pauses are the key for segmenting. Separation will occur:
     (a) when something new is talked about (the utterance belongs to the next action)
     (b) after a long pause of speech
     (c) with the start of special words
        - "and then" is a key word for something new
        - "look at the green" silence [EndAction BeginAction] "look at the yellow" pause ...
        - "look at the green" silence [EndAction BeginAction] "hey Pete, are you looking – the yellow one goes ..."
- **video beginning**
  1. A separation will be made: when a new goal-directed-movement of any body part starts
     - head turn and eye gaze towards new object
     - moving hand from home position[1]or long pause in motions
     - changing goal and direction of movement in a continuous flow of motion
       - ∗ e.g. ... grasping, lifting, inserting [EndAction BeginAction] grasping...
       - ∗ e.g. ... grasping, lifting, inserting, pausing [EndAction BeginAction] grasping...

Figure 5.4: *Segmenting rules to find start points of actions (segments) for the audio and video signal.*

In general the rule "a new segment starts whenever there is a new goal-directed speech or motion" is easier to apply. A similar rule for endings is harder to apply, because action endings are somewhat blurry. Pauses in motion and speech in a gapless annotation make it harder to define boundaries; a new segment category (pause) would be needed here. For example, consider the following sequence of hand-cup movements "`grasping, lifting, showing, inserting,(E)` ***`putting hands down (E), not moving,`*** `(E)(S) turning torso to new cup, grasping ...`".

If there is a simple rule for action ends, like "end of goal-directed motion" the end-begin (transition) point should be after inserting or somewhere in the middle of the italic text, see (E). But this would make the rule too variable and would not result in a comparable annotation of the data set. The action start will not leave such room for variability, as the above example demonstrates. There is only one point, where a new goal-directed motion starts, see (S).

5.1.2 Task Segmentation Results

The results of task segmentation reveal that all segments except the *summary* segment share a similar mean of duration and have a similar standard deviation. In approximately 80 % of the cases the start of one video annotated segment (e.g., short: *a1 v*) precedes the start of the corresponding audio interval (short: *a1 a*) and so does the ending of the video interval.

Interval Duration

One way ANOVAs[2] illustrate that not only the audio and video segments of the same type (e.g. a1 audio and a1 video) but also intervals from *introduction* to *action 3* (*i1–a3*) have no significant differences in their length; $F(7, 432) = 0.77$, $p = 0.6123$ without *su* segment and with *su* $F(9, 540) = 6.54$, $p < 0.0001$. The multiple comparison procedure (at a 95 % confidence interval) reveals that only *summary* segments audio and video (su a and su v) have significantly different mean durations compared to the others but not to each other. The box plots in Figure 5.5 illustrate the mean duration of all segments, *introduction* till *summary* (*i1-s1*), for audio and video segments.

Moreover, the box plots and Figure 5.6 reveal a broad standard deviation in *summary* segments (short: *s1* or *su*). This is due to the non restricted end of the experiments. The parents determined themselves when to end the task. Figure 5.6 represents the standard deviation for all durations of audio and video annotations of all experiments and separately for each experiment setting “Becher” (cup) and “Minihausen” (building blocks).

[2] ANOVA = here: Fisher’s ANalysis Of VAriance, uses Fisher’s F-distribution as part of the test of statistical significance.

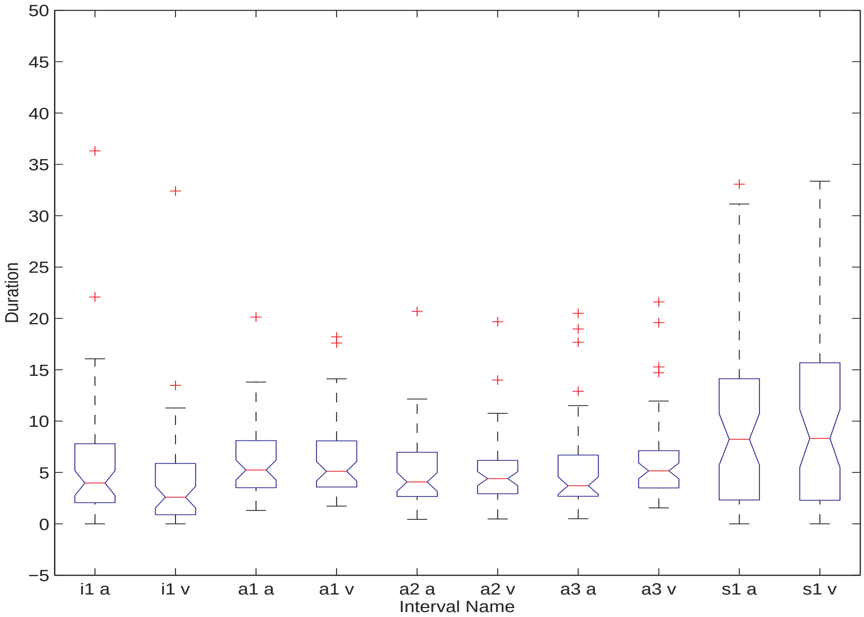

Figure 5.5: *Box Plots represent the length of all segments for audio (a) and video (v) segmentation. All segments, apart from the "summary" (s1) segments, have the same average length and their variance is similar.*

The Figure illustrates the rather low standard deviation for *actions 1-3* compared to *in* and *su* segments

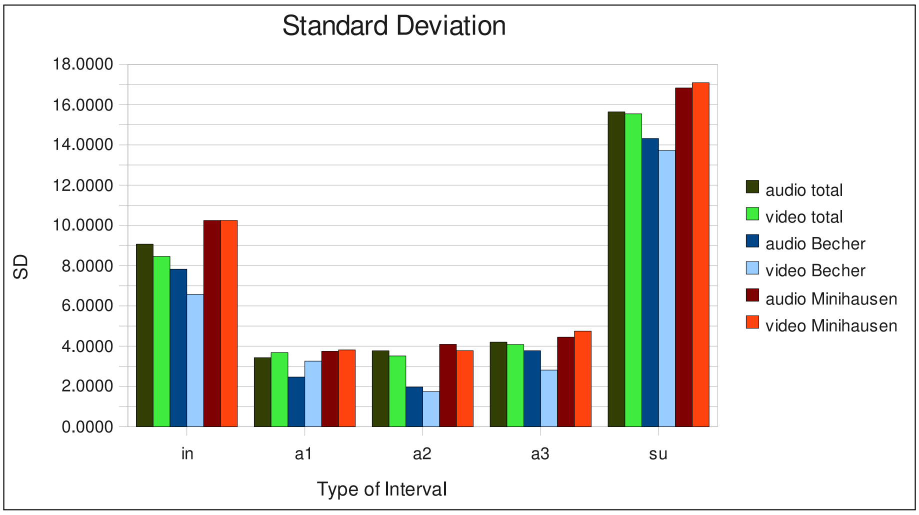

Figure 5.6: *Standard Deviation of the duration per interval for audio and video segmentation. The bars depict SD for all experiments (total) and per experiment setting, Becher = cup and Minihausen = blocks (audio = dark colour, video = light colour)*

On- and Offset

Beginning and end of audio and video segments were compared to each other. In 82.59% of the cases the initial points of segments, which were separated by means of the video signal, precede their corresponding audio segments. The mean onset is 1.125 seconds. In only 12.95% of the cases it is vice versa and in 4.46% of the cases there is no corresponding segment in either audio or video.

These findings are at the first sight not consistent with results of other experiments in this area. Gogate *et al.* and Brand *et al.* report that speech and goal-directed motion start and end temporal synchronous (on- and offset only up to 300 ms) in ID-Input. But considering that Brand *et al.* did not count the grasping or reaching movement towards objects as action, see Chapter 2.2 and [Bran 07c]. Moreover, Gogate *et al.* did not assign start and endpoints to motions. They counted speech and motion as synchronous whenever speech was accompanied by motion, see Chapter 2.2 and [Goga 00]. The mean onset of video (motion) to audio (speech) in this thesis contains the reaching and grasping act towards the object and in 82.59% there is motion when speech starts. For that reason this findings can be considered as consistent with the latest research.

Naturally, the results of the comparison of the end points are very similar as Table 5.1 illustrates. Onsets and offsets are the same; as mentioned before, there are no gaps between the segments. The only difference between onsets and offsets is between *i1-offset*, *a1-onset* and *a3-offset, s1-onset.* 12 parents started with *action 1* and 10 parents had no *summary* segment. According to the reasons above, it will be only spoken of onsets. Only if it is important offsets will be mentioned.

The onsets of segment *a1, a2* and *a3* share a very similar mean length (between 1.2-1.5 seconds, 10% trimmed mean). The onset of segment *su* has a mean duration of only 0.373 seconds. That is the offset of a3, too and the goal of stacking the cups should be fulfilled by then. The completion of the entire task might lead to a tighter alignment of audio and video signals. The standard deviation for *a1* (SD = 2.83) is slightly higher compared to the others (SD 1.3-1.6). This effect is also shown in Figure 5.7. Starting the demonstration might be slightly more variable compared to continuing a demonstration (see onsets of *a2* and *a3* in Figure 5.7).

	video precedes audio	audio precedes video	no partner
onset	82.59%	12.95%	4.46%
offset	81.70%	12.95%	5.36%

Table 5.1: *Onset of audio – video starts and offset of audio – video ends*

As stated before, the segment *summary* represents everything that happens after fulfilling the task. A statistical analysis gives evidence for the difference to the other segments. For this study, the segment *summary* will be excluded from the computation of conditional probabilities. For completeness however, it will be considered in computation up to the point of conditional probabilities and their interpretation.

Having now obtained the initial and completion points of the annotated intervals, the four features and their characteristics need to be associated with their corresponding task segment (*i1-s1*). And then they are again assigned into *beginning-, middle-* and *end* parts.

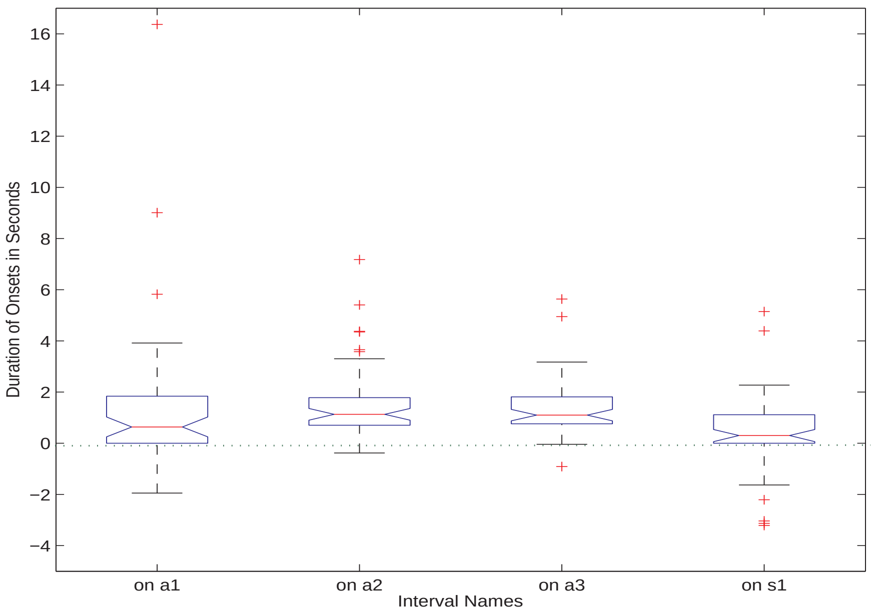

Figure 5.7: *Box Plots of the duration of the Onsets from "a1-s1". The average Onset duration is positive, that means the video segments precede the audio segments. The average duration of the Onset of the "summary" segment (ons1) is closer to zero than the other Onsets (see green doted line)*

Chapter 6

Basis for Analysis

6.1 Assigning Features to Segments

Each feature characteristic is a time interval with a start and an end point. Each interval carries one characteristic value, for example “r” for “rising”, a Word like “put”, a certain Eye-Gaze direction “child” or a certain Velocity of Hand Movements “fast”.

In order to assign feature intervals to their corresponding task segments (first to *i1-s1*, and then to *Beg-End*), the following decisions need to be made.

One has to determine, if start or end points of the intervals are important or in other words, if it is possible to divide the feature characteristic interval into smaller ones. This decision depends mainly on the fact that some feature intervals can not be divided. Words and Intonation are such features. Dividing the Word would result in either duplicating the word or splitting it into syllables, both results are unwanted.

Gogate and co-workers report that for children co-occurring motion and language do not need to be temporally aligned as precisely as it needs to be for adults [Goga 00]. This is the reason, why some blurriness around start and end points of *Beg-End* units is allowed for Word and Intonation intervals.

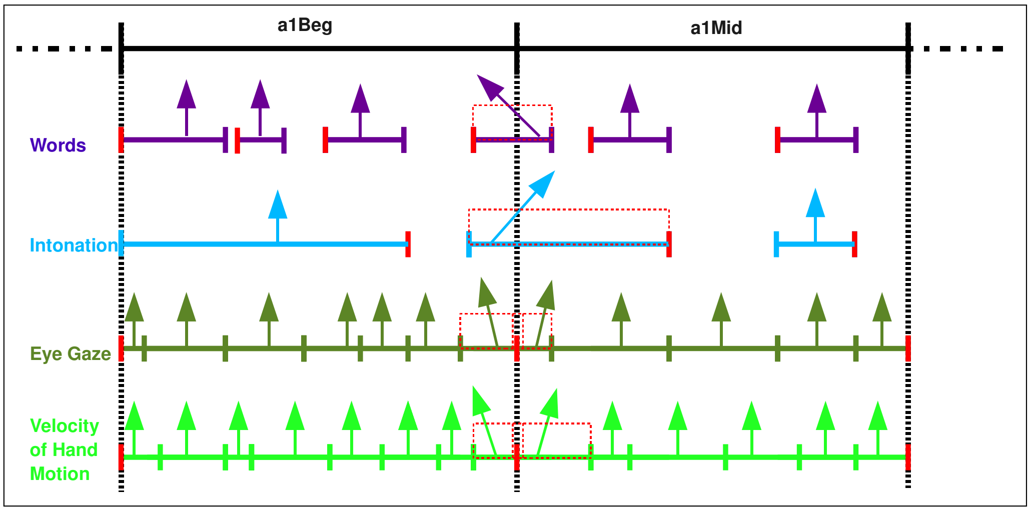

Figure 6.1: *Assigning feature intervals to task segments. Word and Intonation intervals cannot be divided, therefore they are assigned to a segment by rule. The horizontal doted lines represent segment boundaries. Word intervals, which start in a segment, are assigned to this segment. Inclusion is defined by the start point marked in red. Intonation intervals are associated with a segment by the end point marked in red. Eye-Gaze and Velocity intervals are dividable and more frequent.*

Words will be sorted by their beginning time. Intonation intervals, however, will be judged by their end point time because the end of the segment determines the value in this case (please read Chapter 3.2). Eye-Gaze and Hand Velocity intervals will be divided into smaller intervals, if they enclose a segment start or end (*i1-s1* and *Beg-End*) because they are only summaries of smaller units. Here a value was produced every 40 ms, please see Chapter 3.3 and Chapter 3.4 for exact description. Column two and three of Table 6.1 and Figure 6.1 sum up the information given above. Figure 6.1 illustrates through each coloured row and corresponding arrows how each feature interval is assigned and adjusted. Each coloured row in Figure 6.1 represents through arrows, how each feature interval is assigned and adjusted. It also illustrates the proportion between feature intervals and in which frequency intervals occur compared to the other feature intervals.

Before assigning and adjusting feature intervals at level 4 (*i1Beg-s1End* (see Figure 5.1) there are some further problems to be solved. *i1-s1* segments (level 3) need to be divided into smaller units *Beg, Mid, End.* In this case there are two possibilities, how

intervals	sorted by	divisible	Word boundary
Word	start	no	yes
Intonation	end	no	yes
Eye Gaze	start	yes	yes, no
Hand Velocity	start	yes	yes, no

Table 6.1: *Feature intervals can be assigned to the task segments in different ways and task segments can be divided on basis of word boundaries or audio boundaries, column 4.*

those segments can be trisected, either the entire time per each segment *i1-s1* is divided by three or in this case, where language is more in focus, word-boundaries are taken to trisect the segments. This means that the start of the first Word in each segment (*is-s1*) and the end point of last Word interval, make the time length, which will be trisected, (please see at Figure 6.2). This ensures that each unit (*Beg-End*) contains about the same amount of Word intervals.

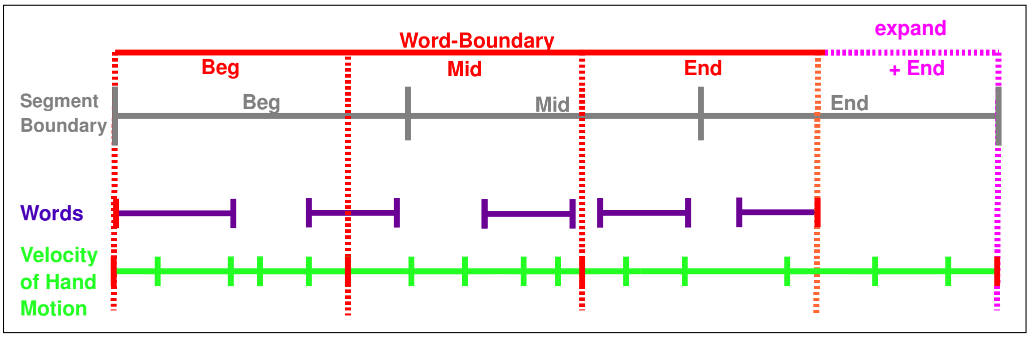

Figure 6.2: *Word Boundaries as boundaries for segment trisection into segments Beg-End. Horizontal grey lines mark the segment trisection (simple trisection). The horizontal red lines represent the trisection boundaries on the basis of the word boundary. The end part needs to be expanded, because Eye-Gaze and Velocity intervals have values in the part, which is cut of by the word boundary trisection as no words are located, there (pink doted lines).*

The problem with this is that features not based on language (Eye-Gaze, Hand Velocity) will have occurrences between the segment's last Word and completion point of the segment. In Figure 6.2 the problematic zone is marked in pink. If the *End* segment is enlarged by about the missing length, then there are on average four for eye gaze intervals and 18 Velocity intervals more in the *End* unit than in every other segment.

The green row in Figure 6.2 shows exemplary a few intervals only. This could result in balancing out the characteristics in the *End* segment. Another possibility would be to ignore word-boundaries with the video based features only. For Eye-Gaze and Velocity of Hand Movement data, both variants will be tried out and results presented and discussed later in Chapter 7.

6.2 Conditional Probability – Bayes' Rule

Without segment *su* there are now 12 segments (*i1Beg-a3End*), which have been assigned feature characteristics. Figure 6.3 illustrates for simplicity only three segments (*a1Beg-a1End*) as boxes (or cylinders) and the corresponding Intonation characteristics as balls with rising (r = green), falling (f = blue), continuing (c = yellow) and unknown (? = red) values.

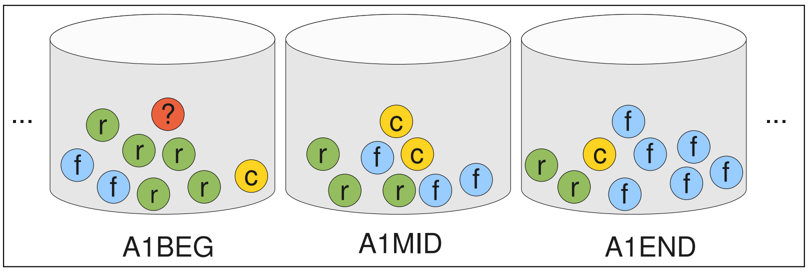

Figure 6.3: *Example for computation of conditional probabilities. Intonation values are represented as balls in three boxes (segments), blue=f=falling, green=r=rising, yellow=c=continuing, red=?=unsure.*

The problem is the calculation of the conditional probability of any segment $Segment_x$ if a distinct feature characteristic fc has already occurred: $P(Segment_x|fc)$.

If one supposes that the feature characteristic "rising" (fc) occurs and one wants to know in which segment ($Segment_x$) it was likely to occur. This problem can be transferred to an easier formulated question. Suppose you have three boxes as depicted in Figure 6.3. "Which box would you guess if the ball drawn is green and what is your chance of guessing right?"

The goal is to have a simple model where as little as possible prior knowledge is needed. The general formula for computing conditional probabilities comes from Bayes. His Rule is used to model the problem, which can be found in (6.1)

$$P(Segment_x|fc) = \frac{P(fc|Segment_x)P(Segment_x)}{\sum\limits_{i=1}^{n} P(fc|Segment_x)P(Segment_x)} \tag{6.1}$$

All components for the computation of the Bayes's Rule can be derived:

- $Segment_x$: the events $Segment_x$ represent n mutually exclusive possible results of the first stage of some procedure. In this case they are the twelve possible segments *i1Beg-a3End*. Which of these events has occurred is unknown. Suppose they contain the results fc

- $P(Segment_x)$: unconditional probability or prior probability. It is assumed that each event $Segment_x$ has the same possibility $\frac{1}{n} = \frac{1}{12}$

- fc: the result of some second stage procedure has been observed, whose chances depend on which of the $Segment_x$'s has occurred. The fc's are the distinct *Feature Characteristics*, for instance as "rising", "falling", "continuing" and "unknown" are the characteristics of feature "Intonation".

- $P(fc|Segment_x)$: likelihoods or conditional probability. In this case it is the ratio of the number of observations (fc) to the total number of observations in each $Segment_x$ (the relative frequency). The example in Figure 6.3 represents three segments as boxes with feature characteristics as balls. The probability for a "rising (green ball)" given the occurrences of $Segment_{a1Beg}$ is
$P(rising|a1Beg) = \frac{\#(rising)}{\#(\Omega = \text{ all fc in a1Beg})} = \frac{5}{9}$

- $\sum\limits_{i=1}^{n} P(fc|Segment_x)P(Segment_x)$ the overall probability $P(fc)$ is the weighted average of the conditional probabilities $P(fc|Segment_x)$ with weights $P(Segment_x)$.

Each feature will be computed separately. For determination of the conditional probabilities for all segments in regard of distinct feature characteristics one needs the number of feature characteristics of the corresponding feature per segments only. A model calculation of the conditional probabilities for all segments, if "rising" has occurred is demonstrated in Table 6.2 on the basis of the example in Figure 6.3.

$P(Segment_x\|fc) =$	$\frac{P(fc\|Segment_x)P(Segment_x)}{\sum\limits_{i=1}^{n} P(fc\|Segment_x)P(Segment_x)} =$ $with\ P(a1Beg){=}P(a1Mid){=}P(a1End){=}\frac{1}{3}$	$\frac{P(Segment_x)P(fc\|Segment_x)}{P(Segment_x)\sum\limits_{i=1}^{n} P(fc\|Segment_x)} =$ $\frac{P(fc\|Segment_x)}{\sum\limits_{i=1}^{n} P(fc\|Segment_x)}$
$P(a1Beg\|rising) =$	$\frac{P(rising\|a1Beg)P(a1Beg)}{\sum\limits_{i=1}^{n} P(rising\|a1Beg)P(a1Beg)} =$ 38.55%	$\frac{\frac{1}{3} * \frac{5}{9}}{\frac{1}{3} * (\frac{5}{9} + \frac{3}{5} + \frac{2}{7})} =$
$P(a1Mid\|rising) =$	41.63%	
$P(a1End\|rising) =$	19.82%	

Table 6.2: *Exemplary calculation of the conditional probabilities for the example in Figure 6.3*

As it is recognisable in Table 6.2 the weights $P(Segment_x)$ can be cancelled from the nominator and denominator, see row one. This is only possible because there is now prior knowledge of $P(Segment_x)$ assumed and the uniform distribution of the segments is taken.

6.2.1 Limitation of Conditional Probability

The conditional probability $P(Segment_x|fc)$ as computed above does say something about which segment is likely to be the segment, which a feature value comes. But it does not reveal anything about the overall rate of distinct values. For example, if there is a single occurrence in one segment but nowhere else, the probability $P(Segment_1|fc)$ is 1 and for all other segments $P(Segment_{2,\ldots,n}|fc)$ it is 0. Following up the example in Figure 6.3, the probability $P(a1Beg|?)$ is 1 for the "unknown" (red) value in segment *a1Beg*. Thus one has to be careful not to judge the quality of the feature distribution only on conditional probability. In the example given, the high probability could induce a wrong interpretation, that there is, for example, a pattern found: *all "unknown" values, which occur, can be associated with the segment "a1Beg".* In other words "unknown" values could be taken for start of an *action.* But single occurrences are not at all statistically significant. The overall rate of a feature characteristic or value gives information on how to interpret the very high probabilities. Another uncertainty with a probability value is that you do not know if the value you are seeing does reveal any information.

For example, suppose you have probability $P(i1Beg|x)$ of 8.3% for some feature x. The value does not reveal anything about the probability of the other segments – the probability could be $\frac{1}{12} = 8.3\%$ for all 12 segments and there would be no information gain or in other words no uncertainty loss about where the feature x comes from. So if the results are equiprobable a measure should then reach a maximum value. The Shannon entropy gives a measure for "uncertainty".

6.3 Shannon Entropy

The entropy by Claude Elwood Shannon is a measure of randomness or "uncertainty". For a discrete set of variables the entropy is measured in *bits.* One bit corresponds to the uncertainty that can be resolved by the answer to a single yes/no question, adapted from [Duda 01]. For the computation of the entropy one needs a set of discrete symbols or variables here: $\{$*i1Beg,...,a3End*$\}$ with associated probabilities, here $P(i) = P(Segment_x|fc)$. The sum of these probabilities is 1: $\sum\limits_{i=1}^{n} P(Segment_x|fc) = 1$, here with $n = \#variables = 12$. The Shannon entropy is computed with the binary logarithm for discrete variables, please compare [Duda 01]:

$$H = -\sum_{i=1}^{12} P(i) \log_2 P(i) \tag{6.2}$$

The equation (6.2) reaches its maximum (H_{max}) if the probabilities are equally likely, thus $P(i) = \frac{1}{n}$. In this case, with $n = 12$, H_{max} is 3.585 bits. This means that on average one needs to ask 3 to 4 yes/no question to be sure into, which segment a seen feature value x belongs, equation (6.3) shows the corresponding calculation.

$$\begin{aligned} H_{max} &= -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n, \quad if \quad \forall P(i) = \frac{1}{n} \\ &= -\sum_{i=1}^{12} \frac{1}{12} \log_2 \frac{1}{12} = \log_2 12 \\ &= 3.585 bits \end{aligned} \tag{6.3}$$

Usually the entropy needs to be normed[1] to make the values comparable with other statistical experiments but since there are always 12 segments in each of this thesis feature calculation there is no need to norm the entropy. All entropy values are between $0 > x \leq 3.585 bits$ (the minimum 0 and maximum 3.585 bits) and therefore comparable with each other.

The entropy value for all probabilities $\{P(Segment_x|fc), ..., P(Segment_x|fc)\}$ is approximately 1.52 bits, for the example, which is depicted in Table 6.2. On average 1.5 questions have to be answered to be certain where the occurred "rising“ value belongs to. The maximum entropy value for this example is 1.58 bits. The result is not far away from the result for the result of the uniform distribution. This conforms with the values shown in Table 6.2. The probabilities 41.63% and 38.55% differ only 3% in that nearly equally probable.

Note, that the computation of the entropy depends on the conditional probabilities, which means a single occurrence will have entropy of 0. Therefore, the overall rate of feature occurrences should always be referred to before any interpretations are made.

[1] e.g. with H_{max} see equation (B.1) in Appendix B

Chapter 7

Results

This chapter presents the conditional probabilities and, when needed, entropy values and overall rates of the four features. Each feature and its values will be examined and interpreted separately. The discussion will be held for all features.

Mainly there are two patterns in the outcome to observe (see Figure 7.1). The conditional probabilities of all segments observing one feature characteristic x, shows mainly either a global or a reoccurring pattern.

A global pattern means, that the probabilities have only high values at successive segments and share most of the probability density, e.g. the first third or half of the task or segments in *action 1* (*a1Beg, a1Mid, a1End*).

A reoccurring pattern, as the name says, has reoccurring high probabilities at segments named either *Beg* or *Mid* or *End*. For example, segments with the name *Mid* possesses more than 50% of the conditional probability density (see Figure 7.1).

Figure 7.1 illustrates the global and the reoccurring pattern. The black punctuated line displays the conditional probability per segment and the pink punctuated lines stand for the conditional probability density. Marked is only the density, which exhibits the pattern.

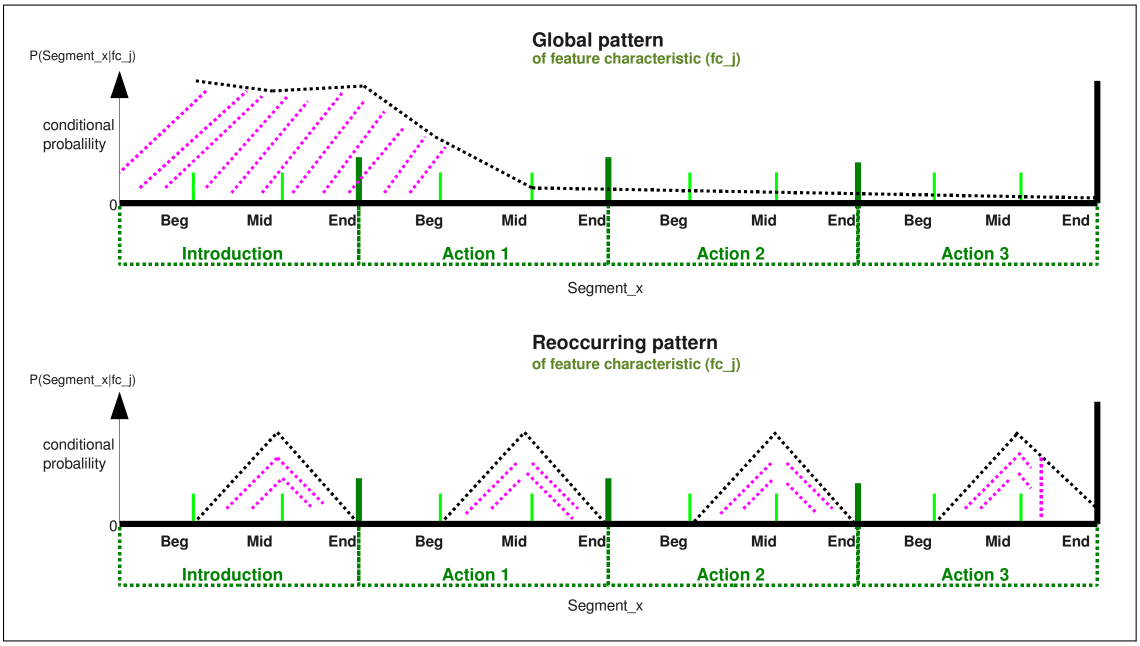

Figure 7.1: *The Figure represents the main patterns for conditional probability (CP) $P(Segment_x|fc_j)$ outcomes (fc_j =feature characteristic). A global pattern has high CP values at successive segments, which share most of the CP density. A reoccurring pattern has high CP values at segments named either "Beg", "Mid", "End".*

The global pattern in Figure 7.1 shows high probabilities only for the segments *i1Beg* to *a1Beg*, thus the feature characteristic x belongs to one of either segments at the beginning of the task. Whereas the reoccurring pattern reveal segments, with the name *Mid*, holding approximately 90% of the conditional probability density.

The results (conditional probabilities) are summed up for interpretation, see above. This is only mathematically correct, if it is assumed that a new statistical experiment is created, when saying: $P(i1Mid|fc_j) + P(a1Mid|fc_j) + P(a2Mid|fc_j) + P(a3Mid|fc_j)$ is equal to $P(Mid|fc_j)$. To keep the equation true, the priori probability $P(Segment_x)$ of segments need to be likewise summed up. In this example, the new segment *Mid*, which consists of all four segments named *Mid*, has the priori probability of the sum of the priori probability of all *Mid* segments:

$$\begin{aligned} P(Mid) &= P(i1Mid) + P(a1Mid) + P(a2Mid) + P(a3Mid) \\ &= \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} \\ &= \frac{1}{3} \end{aligned}$$

Only then are statements, like "all segments named *Mid* share 90% of the conditional probability density (on condition that fc j occurred)" true. If a uniform distribution for the priori probability of summed up segments (*Mid*) and a single segment (*a1End*) is assumed, the ratio of the nominator and denominator in the Bayes' equation will change.

The interpretation of the feature characteristic is not always easy as it is shown in Figure 7.1. Some feature values do not have a clear pattern. They lie between both ways and some do not have a big difference to the uniform distribution. Therefore, only the "best" feature characteristics will be examined in detail. Others will be presented in a more summarised manner and some others will be found in Appendix C. Moreover, each of the following four sections displays a Table with important data about the feature and diagrams to illustrate the probability distribution per feature value.

Note, when referring to probabilities, conditional probabilities are meant except it is clearly expressed otherwise. During the interpretation, the names of feature values are taken to distinguish between probabilities. Sometimes the more correct, but very long form: "the conditional probability of on condition that feature x occurred", is clipped. Other probability will be marked as such.

7.1 Word

The feature Word has 26 values (words), which occurred in more than 0.01% of all recorded words. This threshold is necessary to throw out words, which occurred less than 16 times over all segments and experiments, e.g. “uppsi”, “puh”. 280 distinct words were uttered, 203 occurred only once to three times (a list of all words can be looked up in Appendix C.1).

German speech could be extracted for 49 experiments (49 parents), who spoke an average of 32 words per task. Table 7.1 displays the average amount of words per unit *Beg, Mid* and *End*, which lies between 8-13 words per unit.

The conditional probabilities of 18 feature values are displayed in Figures. For eight words there was no pattern visible and nearly maximum entropy values, which indicate that the conditional probability is uniformly distributed. Table 7.1 shows these words with square brackets. However, these diagrams are displayed in Appendix C for completeness.

# of experiments	49
fc per task	32.37
fc per *Beg* seg.	13.53
fc per *Mid* seg.	8.35
fc per *End* seg.	10.49
# of distinct fc	26
fc values	[auch], [becher], [da], dann, [das], [den], der, die, drauf, du, ein, gelber, gruener, guck, [hier], in, [ist], ja, jetzt, kommt, mal, ndk, rein, so, und, [wir]
overall rate selection	> 1%
selection per segment	all fcs

Table 7.1: *Little profile for feature* ***Word****. Figures for feature characteristics (fc) in square brackets can be refered to in Appendix C.1 (seg. = segments)*

As there are a lot of Word characteristics, they are grouped by similar patterns into subsections.

7.1.1 Pattern: first third and first half of the task

The conditional probabilities for segments with the occurring Words “guck”(look), “mal” (belongs to look) and “ndk” (name of the child) display a global pattern. All three Words occur mostly at segments *i1Beg-a1Beg* – the first third of the task. “Guck” and “mal” are

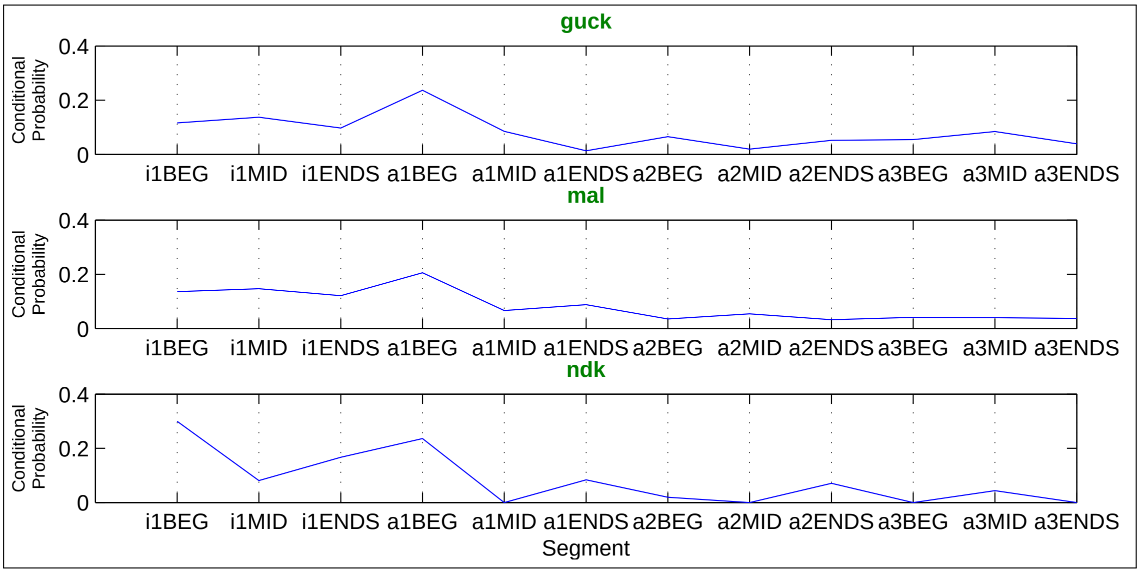

Figure 7.2: *Global pattern for $P(Segment_x|guck)$, $P(Segment_x|mal)$ and $P(Segment_x|ndk)$ in the* ***first third*** *or even half of the task (guck=look, mal= "belongs to guck", ndk=name of the child)*

the Words, which occur most often. 8.57% of all the spoken words are "guck" (look) and 8.26% of all words are "mal". Unfortunately both Words have a high entropy value of 3.3 bits. It is recognisable, the probability graph in Figure 7.2 depicts some low probabilities for the rest of the segments. Since there are four segments with high probabilities, one would need to decide between those four (the max. entropy value for n=4 would be still 2 bits). Even if the entropy value is high, the sum of the probabilities for *i1Beg-a1Beg* show clearly, that a global pattern for each of the three Words "guck", "mal" and "ndk" exists. Table 7.2 displays the sum of the conditional probabilities for the first third and first half of the task which are all above 50%.

With "guck, mal" parents address their child directly. By uttering the child's name the attention is captured even better. Mack *et al.* confirmed that one's own name is a good means to capture attention [Mack 02]. In combination with "guck" (look) parents might direct their child's attention to for example, the toys.

For completeness another feature value should be mentioned, which has a similar pattern to the above. The first half of the task bears 76% percent of the conditional probability

\sum fc \	*i1Beg-a1Beg*= $\frac{1}{3}$ of the task	*i1Beg-a1End*= $\frac{1}{2}$ of the task
guck	59%	68%
mal	60%	76%
ein	59%	76%
ndk	78%	87%

Table 7.2: *Sums of conditional probabilities reveal a **global pattern** for Words **guck, mal, ndk***

density when feature value “ein”(indefinite artical:a) occurs. This can be seen in Figure 7.3. “ein” has a low overall rate of 1.9%. Unlikely “ndk” (name of the child) there are no further interpretations to be made for Word “ein”.

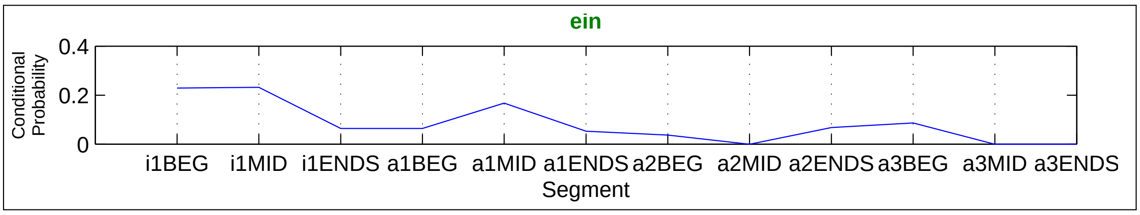

Figure 7.3: *More a global pattern for feature value “ein”, with 76% of the probability density in the first half of the task*

7.1.2 Pattern: mixed *introduction* and *Ends*

Even if a global pattern for feature value “ndk” (name of the child) can be identified, a weak reoccurring pattern can be recognised, too. “ndk”(name of the child) does reappear at the *End* of *action 1* and *2*, the sum of the probabilities for *End* segments on condition that “ndk” has occurred is 32% (see Table 7.3 for an overview).

A similar pattern is shown for the Word “du” (you), (see Figure 7.4 and Table 7.3). The segments summarised by *introduction* have a conditional probability of 53% and the chance that “du” occurred in the time of an *End* segment is 55%. “ndk” and “you” are words, with which parents address their child directly, this might be for preserving or restoring the child’s’ attention after each action and at the start of the task, which might be the reason why “guck” and “mal” occurred there, as well.

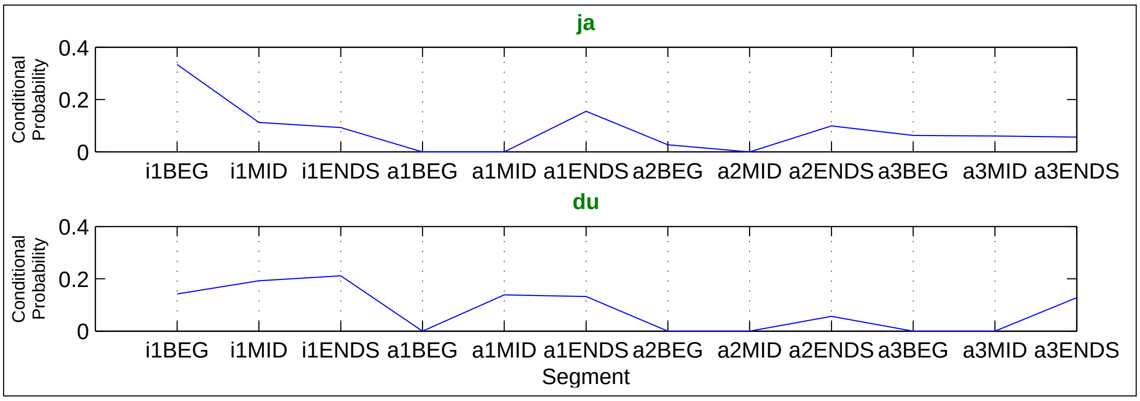

Figure 7.4: *Features reoccurring at action End segments and at the introduction segment*

The Word "ja" (yes) might belong to this group of words, too (*i1*=54%, *End*=40%), see in Figure 7.4 and Table 7.3. "Yes" might be a part of some reassuring questions like, "Did *you* see it? *Yes?*" etc. With the Word "yes" one has to be careful because "yes" was sometimes not addressed towards the child. Especially at the very beginning of the task, speech was sometimes addressed towards the experimenter.

＼sum fc ＼	*i1Beg-i1End*= introduction	*i1-a3End*= all *Ends*
ndk	54%	32%
du	53%	55%
ja	54%	40%

Table 7.3: *Sums of conditional probabilities lean towards a **global pattern** and a reoccurring pattern at for **ndk, du, ja** (ndk=name of the child, du=you, ja=yes)*

7.1.3 Pattern: *action Ends*

The conditional probability distributions for Words "rein" (into) and "drauf" (on top of), which are shown in Figure 7.5, depict a clear reappearing manner at *action Ends*. 59% of the conditional probability density is recorded for "drauf" and even 77% for the "rein". That is one of the cleanest results for the feature Word. Those prepositions do not mean the same, but both describe the position, where the toy is put during the manipulation. The difference in meaning of the prepositions comes from the two

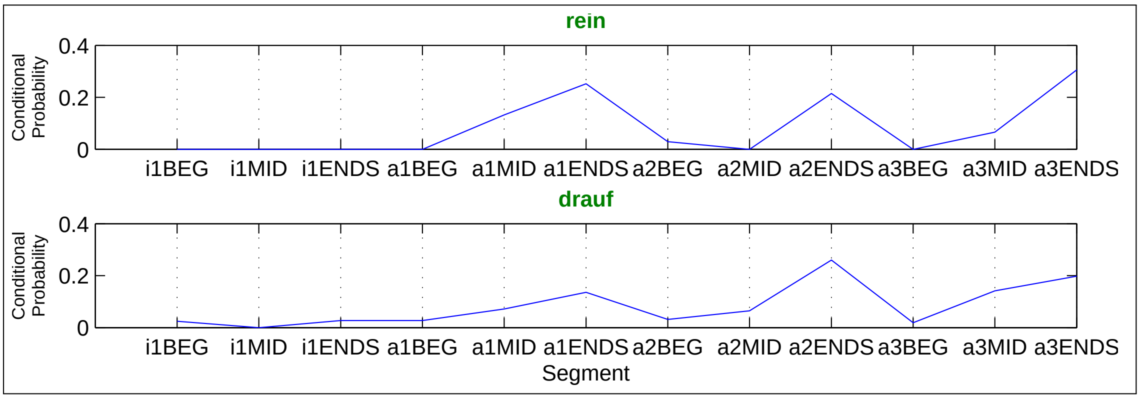

Figure 7.5: *Words, which reoccur especially at actionEnds,* $P(actionEnds|rein) = 77\%$, $P(actionEnds|drauf) = 59\%$ *(rein=into, drauf=on top of)*

different experiment settings: stacking building blocks **on top of** others and inserting cups **into** each other. Despite the different experiment setting a similar structure of occurrences can be identified for those prepositions as assumed in Chapter 3.1. 63% of all "rein" and "drauf" occurrences appear in one of the *action Ends*. This makes 73% $(P(drauf|actionEnds) + P(rein|actionEnds))$ of parents uttering one of the prepositions at the *End* of an *action* segment. For those experiment settings, it can be said that, "rein" and "drauf" stand for *action Ends*.

7.1.4 Pattern: *Mids*

The conditional probabilities for segments on condition, that Words "so" (so) or "die" (feminine article: the) or "kommt" (comes) have occurred, are nearly always present, see Figure 7.6. It is noticable that for "so" and "die" *i1Mid* has the highest conditional probability. The conditional probability $P(i1Mid|so)$ especially stands out with its 31%. That is the third highest conditional probability for only one segment. This peak might be evidence for a communicated start of the real task: "so, what do we have here" or "so, [pause] look at". The third graph in Figure 7.6 depicts the conditional probabilities for "in" (in). The Words "die", "in" and "kommt" show some evidence for a reoccurring pattern at *Mid* segments. The *Mid* segments on condition that "die" occurred hold 51%, "in" 57% and "kommt" 53%. That is not significant because nearly

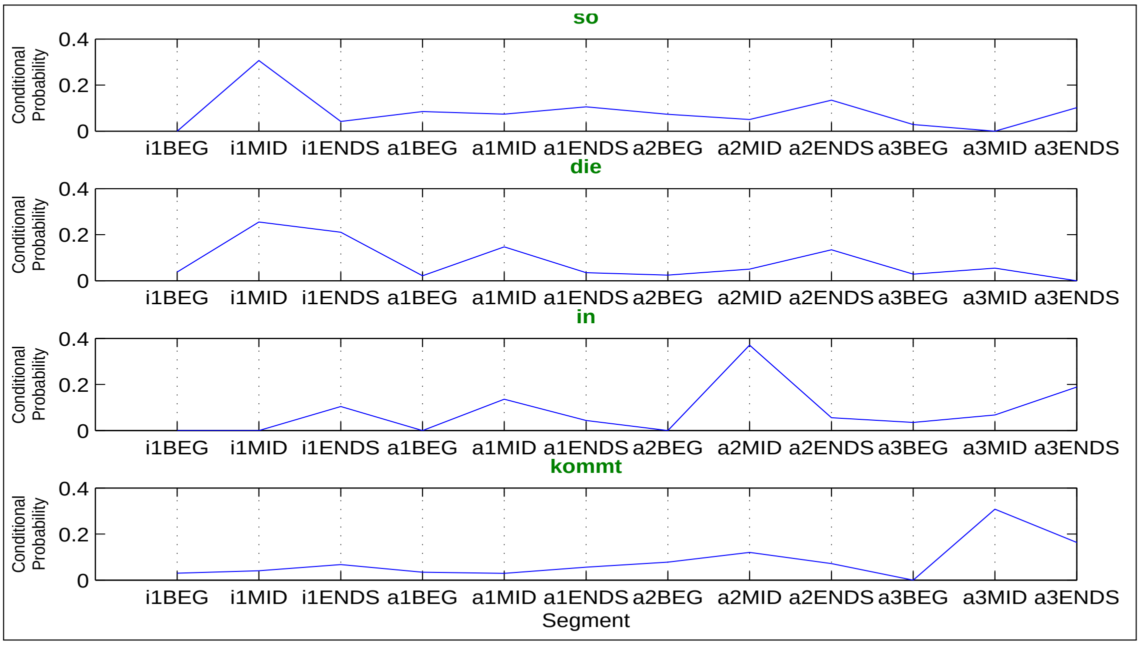

Figure 7.6: *Tendency for **Mid segments** or one peak at a Mid segment, low conditional probabilities at action Beg segments.* $P(i1Mids|so) = 31\%$, $P(Mids|die) = 51\%$, $P(Mids|in) = 57\%$, $P(Mids|kommt) = 53\%$ *(so=so, die=the, in=in, kommt=comes)*

any four segments can be summed up to reach about 50% of the probability density. Noticeable is that "in" and "kommt" have low or zero probabilities at *Begin* segments (3%-6%). Segment *a2Mid* under the condition that "in" occurred has a very high peak (37%) and $P(a3Mid|kommt) = 0.3$. All four Words have a very low overall rate of 2-3%, which makes only 5-7 occurrences for the segment with the highest probability. This means the interpretation should be taken cautiously.

7.1.5 Pattern: *action Begs*

The next four feature values "und" (and), "dann" (then), "der" (masculine article: the) and "jetzt" (now) in Figure 7.7 are grouped together, because the segments *a2Beg* and *a3Beg* in each conditional probability distribution provide the two highest probabilities[1].

60% of the probability density share those two segments for "und". In this case segment

[1] to be precise it is the first and third highest probability for "jetzt"

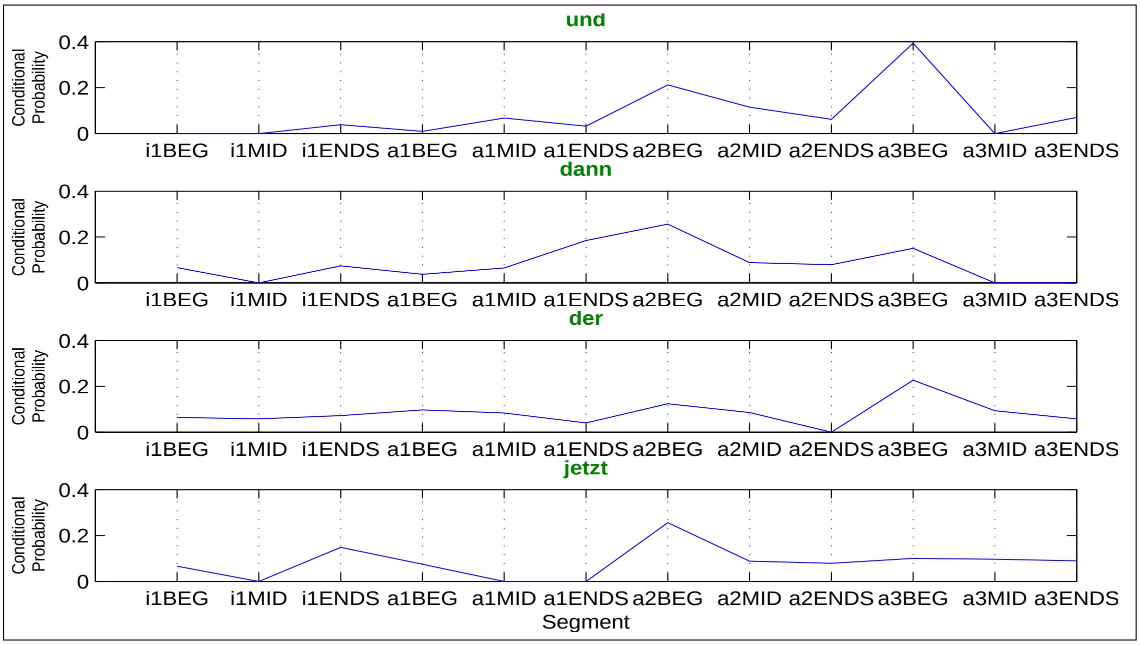

Figure 7.7: *Tendency for a* ***reoccurring pattern*** *at beginnings, especially at segments* ***a2Beg*** *and* ***a3Beg****.* $P(a3Beg|und) = 0.39$ *is the highest conditional probability for one segment. (und=and, dann=then, der=the, jetzt=now)*

a3Beg holds 39% alone, it is the highest probability percentage for a single segment. Naturally, the entropy value is low at 2.55 bits. In addition to this, the Word "und" has an overall rate of 7%. This means about 30 parents have uttered "und" at the *beginning of action 3*, adding the amount from *a2Beg*, 49 parents used "und". A possible interpretation could be that especially the conjunctions "und" and "dann" but "der" and "jetzt[2]",too, might be used to lead from one action to another.

7.1.6 Pattern: *action 2*

The next feature value "gelber" (yellow) has the best entropy value, namely 2.24 bits. The three segments of action 2 share 77% of the conditional probability density, see Figure 7.8. This occurrence of yellow around *action 2* is due to the fact that the second toy of both experiment settings had been yellow. This is why, "yellow" can be taken as a control feature for finding existing patterns and modelling them by conditional

[2] "jetzt" only mentioned for completeness, its entropy value is 3.02 bits

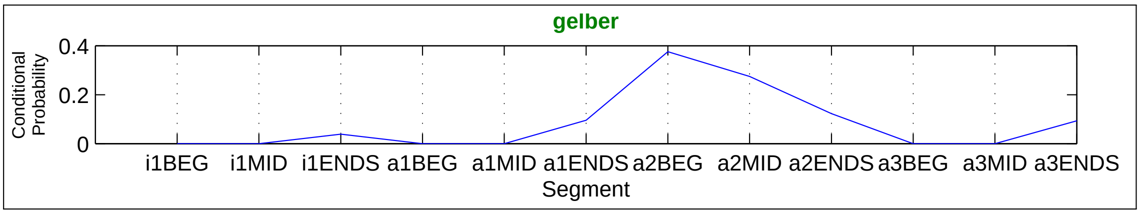

Figure 7.8: *Action 2 on condition that “gelber” (yellow) occurred has 77% of the conditional probabilities density,* $P(action2|gelber) = 0.77$

probabilities.

7.1.7 Pattern: *always present*

This can be said about feature value “gruener” (green) as well. In Figure 7.9 one can see two broad amplitudes at *action 1* and about *action 3*. This is because the green cup was mostly the first toy to be manipulated and the green building block was the last or sometimes second toy to be spoken of. It is not surprising that the entropy value is at 3.05 bits, here.

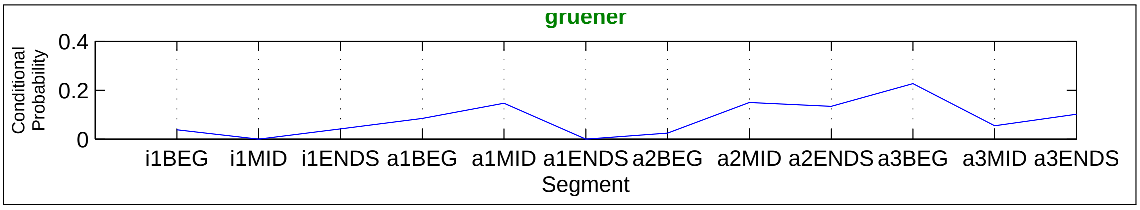

Figure 7.9: *The feature characteristic “gruener” (green) is nearly always present, due to the fact that the first and third toy were green. Entropy value is 3.02 bits.*

7.2 Intonation

The Intonation data comes from 55 experiments, equalling about 10 Intonation values per experiment. On average *Begin* and *Middle* segments contain 2-3 Intonation intervals, which are as long as 3-4 Word intervals. One *End* segment has approximately five Intonation intervals, but, which encompasses only two Word intervals. The exact numbers can be obtained from Table 7.4. The Intonation Section will be much shorter than the Word Section, because there are only four categories to discuss. The category "unsure" (x) will not be presented here, because it has a low overall rate of 3% (17 occurrences over all segments and all 55 experiments = 569 segments, see Appendix C Figure C.4).

# of experiments	55
fc task	10.35
fc per *Beg* seg.	2.76
fc per *Mid* seg.	2.44
fc per *End* seg.	5.15
# of distinct fc	4
fc values	r, f, c, [x]
overall rate selection	$> 1\%$
selection per segment	all fcs

Table 7.4: *Little profile of the feature* ***Intonation****. Figure for feature characteristic [x] will be shown in Appendix C.2 (task without summary, seg.=segment, fc=feature characteristic)*

The categories/values "rising" (r), "falling" (f) and "continuing" (c) have high entropy values of 3.4 bits to 3.53 bits. The conditional probabilities $P(Segment_x|r)$, $P(Segment_x|f)$, $P(Segment_x|c)$ are nearly equally distributed over all segments, consult Figure 7.10. This means, that there are no significant patterns to be recognised, but there are, however, tendencies to present.

The conditional probability $P(Segment_x|continuing)$ shows a tendency towards ***action Mids*** and ***action Ends*** (see green graph in Figure 7.10). The range of the conditional probabilities ($P(Segment_x|continuing)$) is between 3% and 17%.

For the conditional probabilities, $P(Segment_x|rising)$ and $P(Segment_x|falling)$, the range is even more narrow: between 6%-11% for "falling" and 4%-14% for "**rising**". The two highest conditional probabilities, $P(a3Beg|rising) = 0.14$ and $P(a2Beg|rising) = 0.12$ for "rising", occur at *action Begs*, closely followed by *a2Mid* and *a2End*. This hints, that there might not be a reoccurring pattern, but a global one, on ***action 2***.

A slight tendency for $P(\mathbf{a3End}|falling) = 11\%$ as highest probability can be seen for "**falling**". The conditional probability $P(i1Beg|falling)$ is only 0.13% behind. This

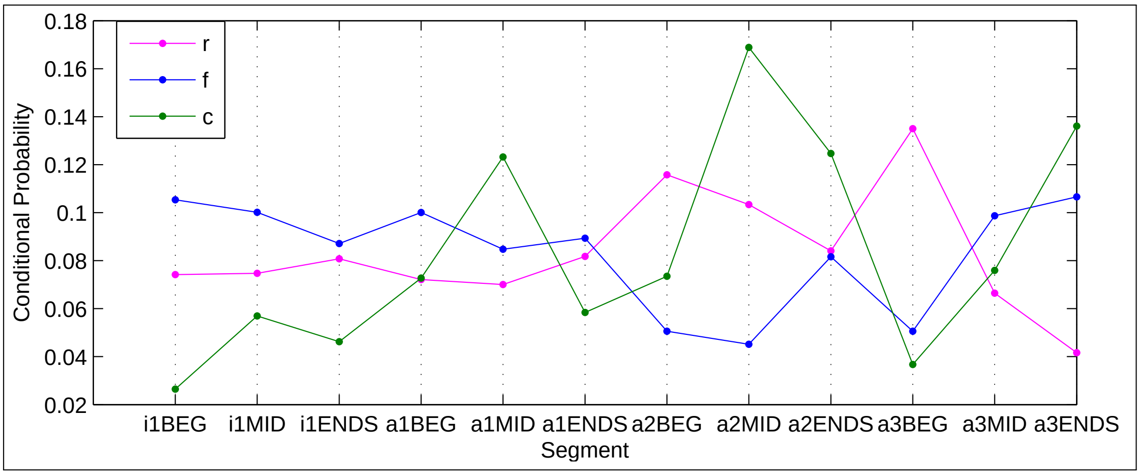

Figure 7.10: *Conditional Probability distribution for $P(Segment_x|r)$ (r=pink=rising), $P(Segment_x|f)$ (f=falling=blue) and $P(Segment_x|c)$ (c=continuing=green). The fine proportion of the y-axis reveals tendencies, even if the conditional probability range is narrow.*

might be due to the fact, that speech in segments *i1Beg-i1End* is directed towards the experimenter in 29% of the cases. Brand and Wrede suggest that a falling intonation contour stands for a completed action [Bran 07b; Wred 05]. This means, that according to the highest probability $[P(a3End|falling)]$ at the end of the task, the intonation contour "falling" works on a higher segmentation level and does not divide the action into three (sub)actions as it is done in this masters' thesis. Those results will not be elaborated any further, because the data does not reveal anything significant. More evaluators and a finer-grained evaluation on word level need to be done to test this hypothesis.

7.3 Eye-Gaze

Across 52 experiments, a task contains approximately 30 changes in Eye-Gaze directions. *End* segments have on average 12 intervals, which is four more than *Beg* and *Mid* segments contain. The four feature characteristics here, Eye-Gaze directions towards the, "Kind" (child), "Objekt" (object), "Instruktion" (instruction), "Sonstiges" (miscellaneous) will be presented in this section. Miscellaneous, means a gaze direction towards something not covered by the directions mentioned before.

First, the section will examine the effects of the "word boundary segmentation" and the "segment boundary trisection".

Eye-Gaze direction is one feature, whose annotation is based on the video signal. It is not, like the feature Intonation, dependent on the parents' utterances. As explained in Chapter 6.1, there are two ways of dividing the segments, *introduction* to *action 3*, into *Begin*, *Middle* and *End* units/segments. The simple approach is to divide the segment into three equal parts by taking the boundaries of the segment (e.g. $\frac{action1End-action1Start}{3}$).

The other possibility is to create new boundaries for trisecting the segments, namely, the start point of the first spoken word and the end point of the last spoken word of the actual segment. This ensures, that there are Words in each *Begin*, *Middle* and *End* unit. The conditional probabilities are computed on basis of both ways. The probabilities do not vary much. The Ansari-Bradley[3] test (alpha = 0.0001) reveals, that the conditional probability outcomes have the same distribution[4]. The average difference of the conditional probabilities[5] for the feature values "Kind" (child) and "Objekt" (object) are 0.3% and 0.47%. However, the mean difference for "Instruktion" (instruction) is 5% and for "Sonstiges" (miscellaneous) 3%. Both have a low rate of occurrences and patterns are somehow better to recognise in the word boundary trisection. Figure C.5 in Appendix C.3 depicts the conditional probabilities on the basis of the segment boundaries. In this section, presented conditional probabilities are based on the word boundary segmentation.

[3] The Ansari-Bradley test tests the hypothesis, that two independent samples come from the same distribution, against the alternative, that they come from distributions, that have the same median and shape but different variances.

[4] I also performed a Kolmogorov-Smirnov test (alpha = 0.0001), which compares both distributions, but assumes, that they are continuous. This test also reveals that the conditional probabilities are from the same distribution

[5] More precisely: the average difference is $P(Segment_{x-segBoundary}|Kind)$ - $P(Segment_{x-wordBoundary}|Kind)$. The short version in the text above should make it more readable.

# of experiments	52
fc per task	29.69 (29.75)
fc per *Beg* seg.	8.79 (9.65)
fc per *Mid* seg.	8.23 (9.40)
fc per *End* seg.	12.67 (10.49)
# of distinct fc	4
fc values	Kind, Objekt, Instruktion, Sonstiges
overall rate selection	$> 0.1\%$
selection per segment	> 2

Table 7.5: *Little profile of the feature* ***Eye-Gaze*** *and word boundary trisection. In round brackets are the numbers of segment boundary trisection of segments (seg.). Translation of feature characteristics (fc): Kind=child, Objekt=object, Instruktion=instruction, Sonstiges=miscellaneous.*

Distinct feature value occurrences had to be more than two per segment and had an overall occurrence rate of more than 0.1%. The overall rate selection is low, to compute the conditional probability, $P(Segment_x|Sonstiges)$, even if the overall rate is only 0.8%. If the gaze direction is neither towards the object nor towards the child nor towards the introduction, then the communication between child and parent could be disturbed. Brand and co-workers found evidence, that there are frequent Eye-Gaze bouts and longer gaze durations of the parent towards the child, when engaged in child-directed interaction (IDI) compared to ADI[6] [Bran 07c]. As a result, there should be no time left for the parent to look at something else. In fact, in 43% (overall rate) of the cases the parents looked at the child and in 54% of the cases parents looked at the toy. This conforms to the findings of Brand *et al.*, who found that children were gazed at 47% of the time in IDI [Bran 07c]. Those eye-gazes towards the child counterbalance the eye-gazes towards the object in each segment, as their overall rate shows.

The conditional probability $P(Segment_x|\textbf{Sonstiges})$ reveals only three segments with probabilities at the start and end of the task: ***i1Beg*** (45.6%), ***i1Mid*** (35.8%) and ***a3End*** (18.6%). This is a tendency, (on basis of Eye-Gaze direction), that parents engaged undisturbed with their child during *i1End-a3Mid* (see Figure 7.11).

The overall rate of feature value "Instruktion" is also low (2%), because it occurred in

[6] Adult-directed Interaction

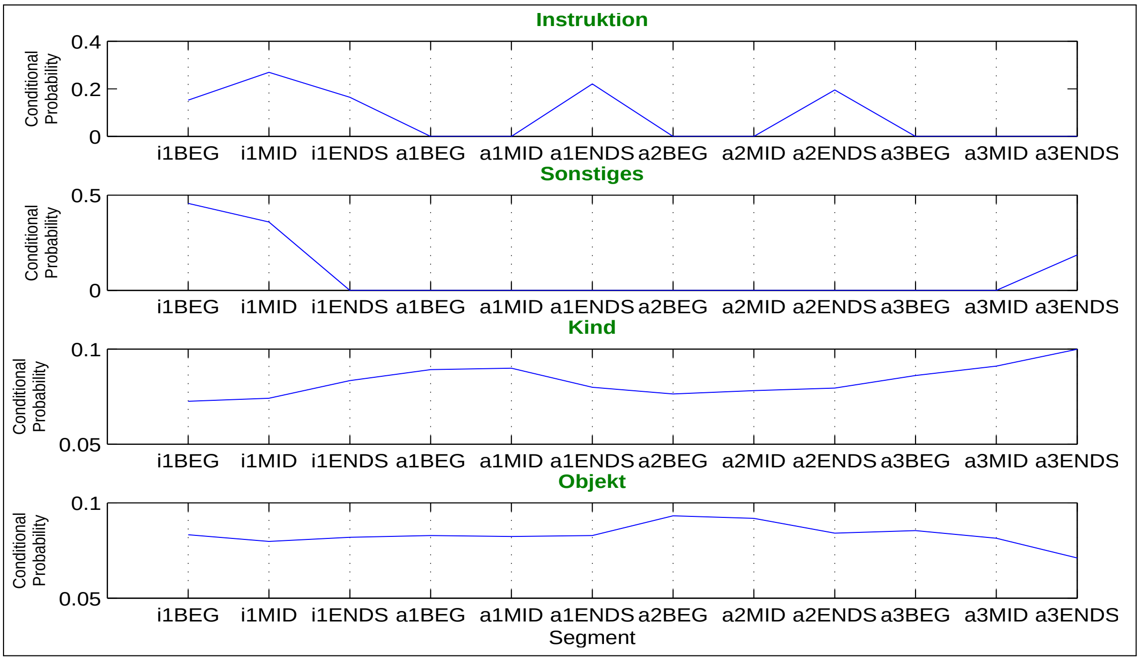

Figure 7.11: *Global and reoccurring pattern for conditional probabilities (CP) $P(Segment_x|fc_j)$, fc_j stands for any of the four Eye-Gaze directions: Instruktion (instruction), Sonstiges (miscellaneous), Kind (child), Objekt (object). CP are created on basis of word boundaries*

Minihausen[7] experiments. However in "Minihausen"- Experiments, there was a little picture (or instruction) next to the toys, showing how to stack the building blocks. The corresponding graph in Figure 7.11 depicts a global and a reoccurring pattern for the conditional probability distribution $P(Segment_x|Instruktion)$. The global pattern contains the *introductory* (*i1*) segments (***i1Beg-i1End***). The summarised conditional probability $P(i1|\mathbf{Instruktion})$ holds 59% of the density. The reoccurring pattern of *action End* segments of *action 1* and *2* share the rest of the conditional probability density,$P(a1a2End|Instruktion) = 0.41$. The gaze towards the instruction might have been interfering with the distribution of Eye-Gaze directions towards the child and object.

The conditional probability distribution $P(Segment_x|Object)$ and $P(Segment_x|Kind)$ are both nearly equally spread (see Figure 7.11). Therefore, the entropy values differ only 0.003-0.006 bits from the maximum entropy value. However, a tendency for a global pattern can be revealed, if the threshold for accounting occurrences per segment is increased from 2 to 50. The results are outlined in Figure 7.12.

A global and a reoccurring pattern for conditional probabilities $P(Segment_x|Objekt)$ and $P(Segment_x|Kind)$ can be recognised in Figure 7.12. The entropy value did not decrease much (from 3.58 bits to 3.51 bits) for the conditional probabilities distribution $P(Segment_x|\mathbf{Objekt})$. Six Segments ***Beg*** and ***Mid*** of *i1, a2, a3* have the highest conditional probability of 11%.

The graph for Eye-Gaze direction "**Kind**" (child) shows the opposite effect for segments. All conditional probabilities are between 16%-19% and appear only at ***End*** segments with $P(End|Kind) = 67\%$ and at the **start of the action** with $P(i1End - a1End|Kind) = 65\%$ (segments *i1End* and *action 1*). The entropy value here is 2.58 bits, because there are still six segments, whose conditional probability are nearly equiprobable.

These tendencies conform to the findings of Nagai and Rohlfing, who applied a saliency

[7] Experiments with building blocks as toys

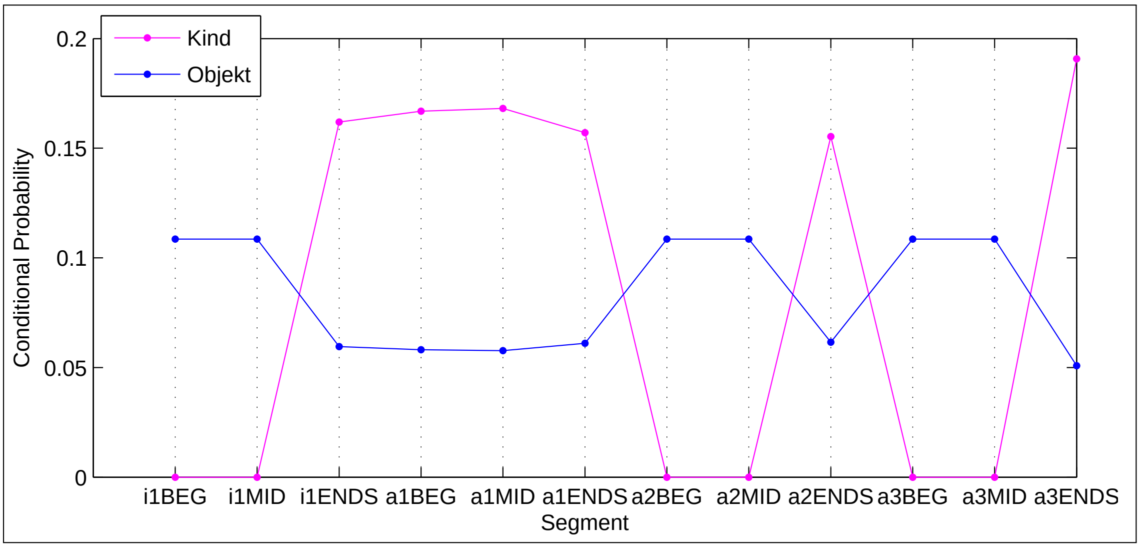

Figure 7.12: *Global and reoccurring pattern for conditional probabilities $P(Segment_x|Object)$ and $P(Segment_x|Kind)$. The occurrences per segment of distinct Eye-Gaze characteristics were filtered. Only occurrences above 50 were taken into account. Segments Mid and Beg of i1, a2, a3 have equally high conditional probabilities (CP) of 11% for "Objekt" (object). The graph for Eye-Gaze direction "Kind" (child) shows CPs from 16%-19% for all segments of the global pattern, i1End + action 1*

based attention model on the cup experiment data [Naga 07a]. Their global model computed, that, before and during the task, the objects attracted more attention. The parents' face attracted more attention after the task. My findings indicate, that this might even be true for a finer grained segmentation of the task.

7.4 Velocity-of-Hand-Movements

The annotation of the feature Velocity of Hand Movements (short:Velocity) was made on the basis of the video signal, like the feature Eye-Gaze (section 7.3). This is why, the time of occurrences of the feature characteristics (fc) are not depending on the utterances of the parent. As explained above and in Chapter 6.1, there are two ways of trisecting the segments (*i1-a3*) into units *Beg, Mid, End*.

First the conditional probabilities (CP) were computed with no limitation of fc selection per segment. Therefore, the results do not reveal any significant pattern, neither if CPs were computed on the basis of the segment boundary nor on the basis of the word boundary trisection (see Figure C.8 in Appendix C.4). Their entropy values are over 3.5 bits (≈ maximum entropy). The high amount of feature characteristic intervals per segments *Beg, Mid* and *End* balances the ratio of distinct feature values across all segments. *Beg* and *Mid* segments contain on average 16 intervals more than the Eye-Gaze segments. *End* segments hold even more intervals: 29 Velocity intervals more than Eye-Gaze fcs. The exact numbers can be seen in Table 7.6.

# of experiments	54
fc task	90.80 (92.91)
fc per *Beg* seg.	25.69 (29.20)
fc per *Mid* seg.	23.46 (30.07)
fc per *End* seg.	41.65 (33.63)
# of distinct fc	5
fc values	nomo, slow, medium, fast, very fast
overall rate selection	>= 0.1%
selection per segment	all vs. only one per segment

Table 7.6: *Little profile of the feature* ***Velocity of Hand Movements****. The numbers of feature characteristics (fc) per task, Beg, Mid, End correspond to the number of intervals without any selection boundaries. With limited selection of one interval (fc) per segment (seg.) per experiment. In round brackets are the numbers of segment boundary trisection of segments .*

There is a reoccurring pattern of high conditional probabilities $P(Segment_x|fc_j)$ for segments *Mid* on the condition that a “medium”, “fast” or “very fast” movement of parent hands occurred. To reveal this pattern, the selection of intervals[8] per segment (*Beg, Mid, End*), which are taken into account for the computation of conditional probabilities, needs to be restricted. Only the very first feature characteristic interval from each *Beg* segments is taken. The middle interval of the *Mid* segment is sampled and the very last element from each *End* segment is counted. This makes, naturally, on average one feature characteristic (fc) per segment and 54 fcs per experiment. As a result, the conditional probability distributions in both trisection ways are very similar[9].

[8] Each feature characteristic (fc) has an interval (time period) of occurrence. The term interval is used in order not be confused with segments or distinct fcs.

[9] The Ansari-Bradley and the Kolmogorov-Smirnov test did reveal so, with alpha=0.0001, see Section

The conditional probabilities of both trisection ways differ by only 0.5%-0.8% (for $P(Segment_x|nomo)$ and $P(Segment_x|slow)$). However, the difference of both trisection ways of conditional probabilities $P(Segment_x|medium)$, $P(Segment_x|fast)$ and $P(Segment_x|vfast)$ is larger, namely between 2%-4%. This is recognisable, when comparing the graphs of conditional probabilities of both trisection ways in Fig. 7.13 and Fig. 7.14

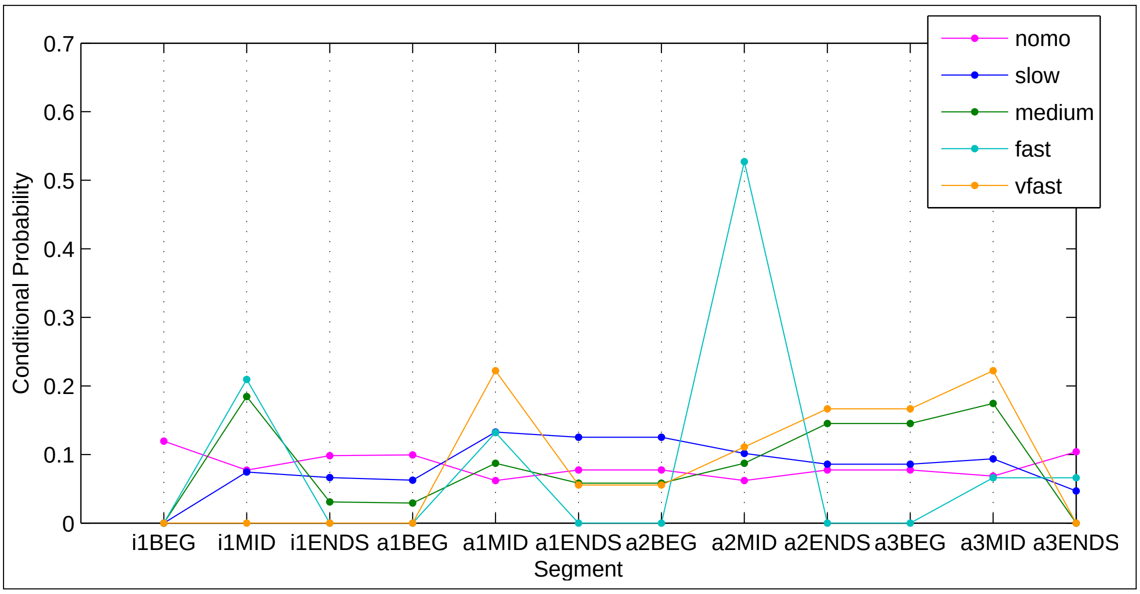

Figure 7.13: *The graphs depict conditional probabilities (CP) $P(Segment_x|fc_j)$ for all feature characteristics (fc_j): nomo (no motion), slow, medium, fast, vfast (very fast) on basis of the word boundary trisection. Selection of one interval (fc) per segment (Beg, Mid, End) and per experiment. Under this condition, the reoccurring high CP of Mid segments for graphs "medium", "fast" and ("vfast") is visible*

Unlike Eye-Gaze patterns, the reoccurring pattern, of high conditional probabilities for *Mid* segments, is better revealed, if the conditional probabilities are computed on the basis of the simple segment trisection (segment boundaries) (see Figure 7.14 for conditional probabilities).

In general all conditional probabilities $P(Mid|fc_j)$ of all ***Mid*** segments with feature characteristic $fc_j = \{$**medium, fast, very fast**$\}$ hold 68%-93% of the conditional probability density. Even $P(actionMid|fc_j)$ with $fc_j = \{medium, fast, veryfast\}$

7.3 for more information.

have 50%-67% of the conditional probability density. Results of Yukie Nagai saliency model indicate, that parents hands (and cups) are salient to the child during the task [Naga 07a]. Salient fast hand motions indicate either parents movements with an object towards inserting / stacking or showing (waving) the object.

The conditional probability $P(Mid|fast) = 93\%$ stands out in both trisection ways. The probability mass is equally distributed among the *Mid* segments ($P(i1Mid|fast)$ ...) if the simple trisection is used. However, if the word boundary trisection is used, the conditional probability $P(a2Mid|fast)$ holds alone 53% (see Figure 7.13: high peak in blue graph). Across all feature characteristics and segments, this is the highest conditional probability peak for one segment. This results of course in a low entropy value of 1.86 bits (else 2.15 bits). The entropy values for conditional probability distribution of “medium” and “very fast” are around 2.7-3.0 bits for both trisecting ways.

The conditional probabilities $P(Segment_x|nomotion)$ and $P(Segment_x|slow)$ are nearly equally distributed across the segments, so the entropy values are still between 3.4-3.5 bits. Noticeable for feature characteristic “no motion” is though, that *Mid* segments have a lower conditional probability than any other conditional probability $P(Segment_x|nomotion)$.

Rohlfing and co-workers stated, that pauses in movements are made between actions or better (sub)actions (like my action 1-3) [Rohl 06; Naga 07b]. These breaks should emphasise the just executed action and indicate start and *End* of an action. My findings conform to this from different perspective. However, *Mid* Segments have a significant high conditional probability $P(Mid|fast)$ on condition, that any “medium” to “very fast” movement has occurred. Around *Mid* segments, (namely in *Begins* and *Ends*), the ratio must be in favour of “no motion” and “slow motion” within other segments than *Mid* segments. Results of conditional probabilities $P(Segment_x|nomotion)$ and $P(Segment_x|slow)$ imply, that “no motion” or at least “slow motion” of hand movements occur everywhere. In fact, the overall rate of “no motion” is 70%, leaving only 3-5% for each hand movement from “medium” to “very fast”. This is not surprising,

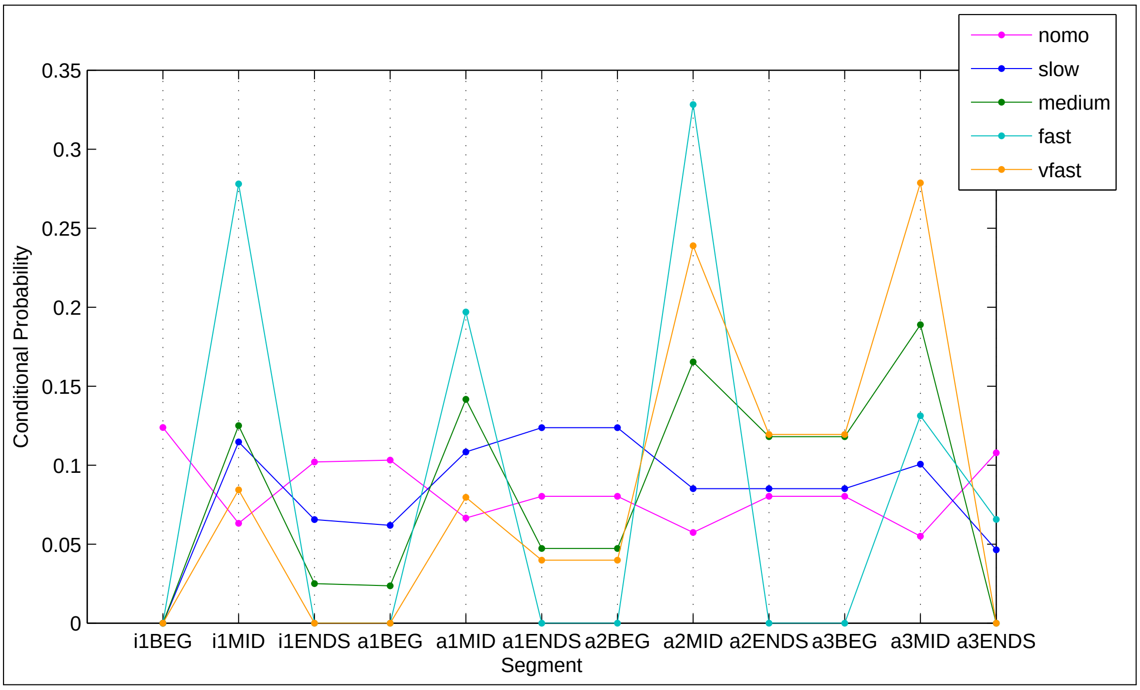

Figure 7.14: *The graphs show conditional probabilities (CP) $P(Segment_x|fc_j)$ for all feature characteristics (fc_j): nomo (no motion), slow, medium, fast, vfast (very fast) on basis of the word boundary trisection. Selection of one interval (fc) per segment (Beg, Mid, End) and per experiment. Under this condition, the reoccurring high CP of Mid segments for graphs "medium", "fast" and "vfast" is even better visible*

because tendencies for abrupt / punctuated movements in child-directed motion have been described by Brand, Rohlfing and co-workers [Bran 02; Rohl 06]. Accordingly, these movements should be short in duration.

The high probability $P(i1Mid|fast)$ cannot result from one of the actions, but from “introducing” (showing) the toys before the actual action starts.

Nagai and co-workers analysed parental demonstrations (stacking cups) only in a global manner with their saliency model. They divided the task into before, during and after the action of stacking all cups [Naga 07a]. The analysing method might be applied to finer grained level of (sub)actions for Velocity feature values and Eye-Gaze characteristics.

Chapter 8

Discussion

Two general patterns have been identified, within the conditional probability examination $P(Segment_x|fc_j)$ (over all segments ($Segment_x$) on condition, that one feature characteristic (fc_j) has occurred).

There are six different reoccurring patterns, which have been found during feature analysis, either *Beg* or *Mid* or *End* segments have high conditional probabilities. Moreover, a distinction can be made between all *Beg, Mid, End* segments or *Beg, Mid, End action* segments with high conditional probabilities .

Four global patterns have been introduced in Chapter 7. High conditional probabilities have been found in the *first third of the task* or a bit finer, high conditional probabilities at segments *introduction (i1Beg-i1End), action 1* and *action 2* have been identified.

A further group of patterns, which cannot be associated with one of the above pattern, is identified. Some conditional probability distributions concentrate mostly on one segment, so that the conditional probability graph depicts one high peak. For example some feature values have been identified to have conditional probabilities $P(Segment_x|fc)$ only at the start and/or end of the task.

Note, that not all feature values, which have been presented, show only one of the above mentioned patterns (for exact numbers refer to Chapter 7). There are mixtures of pattern, mostly a reoccurring pattern at the action (or demonstration) segments and a tendency for a global one at the beginning of the task, see for example “ndk” in Section 7.1.2.

The rest of the feature characteristics (fc) has a conditional probability density, which is evenly spread across the segments, which means that conditional probabilities are always present and not interpretable, if no prior knowledge is available (see for example conditional probability for “gruener”, Section 7.1.7).

Table 8.1 sums up the results of the conditional probability examination for all considered feature values. The feature characteristics (fcs) are associated with their corresponding conditional probability pattern. An evaluation of the significance of the corresponding pattern is coded in the appearance (colour) of feature values.

A fc is coloured in red, if the conditional probability per segment in the corresponding pattern is over 17.5%. For example “rein” (into) is red and is associated with pattern *action Ends*, which then stands for the conditional probability $P(actionEnd|rein) \geq$ 70%. There are always three *End* segments in pattern *action End* $P(actionEnd|fc_j)$, which share on average, a distinct amount of the conditional probability density.

The orange colour represents all feature values, which have an average conditional probability mass per segment of 13.75%-17.5% in their associated pattern, that makes a conditional probability density of 55% for pattern ($P(pattern|fc_j)$), which consists of four single conditional probability $P(Segment_x|fc_j)$.

Black coloured fcs stand for a segment probability $P(Segment_x|fc_j) \geq 10\% < 13.75\%$ within segment in a pattern. Round brackets mean, there is only a tendency towards the corresponding pattern ($P(Segment_x|fc_j) < 10\%$).

pattern		Word	Intonation	Eye-Gaze	Velocity
reoccurring	**Beg**	(no in), (no kommt)		objekt	
	action Beg	und, dann, jetzt, der	(rising)		
	Mid	in, die, kommt		objekt	medium, fast, vfast
	action Mid		continuing		medium, fast, vfast
	End	ja, du, (ndk)		kind	
	action End	rein, drauf	(continuing)	instruktion	
global	**first third**	guck, mal, ein, ndk			
	introduction	du, ja, ndk, guck, mal, ein		instruktion, (kind)	
	action 1			kind	
	action 2	gelber	rising		
peak	**one peak**	jetzt, kommt, so, und, in			fast
	i1Beg		falling	sonstiges, (objekt)	
	a3End		falling	sonstiges, (kind)	
uniform	**always present**	hier, da, den, das, wir, ist, becher, gruener, auch	falling, rising	objekt, kind	nomo, slow

Table 8.1: *Summarised results of the conditional probability (CP) examination of all features. A pattern stands for the conditional probability $P(pattern|fcj)$ per fc. The feature values are coloured, if the CP per segment in the pattern is over 13.75% (orange) and 17.5% (red). Black coloured feature values stand for a CP per segment between 10-13.75%. Round brackets around feature values mean, there is a tendency towards this pattern.*

Looking at Table 8.1 one might imagine, which feature values from the same and different features could occur within the same segment (time period). For example, the intonation contour is extracted/annotated from spoken words, therefore the relation between those feature characteristics is close. It would be interesting, to find out, if these words, for instance "dann" (then) and "und" (and) are the words, from which a "rising" intonation contour has been extracted (referring to Table 8.1 row two). As already explained, per word Intonation extraction should be an important future topic.

In addition, relationships between other features seem reasonable. Due to high conditional probabilities for pattern, $P(End|du)$ and $P(End|kind)$, the feature values "du" (Word:you) and "kind" (Eye-Gaze:child) might be both associated with *End* segments. This is not a great surprise, because whenever one addresses a dialogue partner directly, one should look at him/her, too. However, there are different tendencies noticeable.

$P(introduction|du, ndk, ja)$ and $P(introduction|instruktion)$ have conditional probabilities over 70%. Word values "du" (you) and "ndk" (name of the child) stand for addressing the child directly, but at the same time they still look at the "instruction". The conditional probabilities for Eye-Gaze direction "child" give merely a tendency for segments *introduction*. This reveals that although parents engage with their children verbally, they divide their attention and look at their child in the next time period (segment). $P(action1|kind)$ holds over 41% of the conditional probability density. The conditional probability $P(introduction_{cup}|kind)$ for cup experiments has slightly more probability mass. This means, large disturbances[1] are reflected only slightly by the conditional probability model, it merely models the trend of the majority. Naturally, this should be tested with more experiments and comparable settings.

In addition, comparing feature values of different features can indicate new ways of interpreting the feature values. Considering the Intonation contour "falling" and pattern *i1Beg* and *a3End* in Table 8.1, there is only a tendency towards the pattern. Considering, the Eye-Gaze "sonstiges" (sth. better somewhere else) and this pattern, it is noticeable

[1] Referring to Minihausen experiments, which is half of the data and contained an instruction

that $P(i1Beg|sonstiges)$ and $P(a3End|sonstiges)$ have high probabilities. This could mean, that “looking at something else” hints that the parent is not engaged with his/her child anymore and speech might not be directed to child either. This in turn interferes with a clear pattern for the Intonation contour “falling”, which has been hypothesized and found to occur at the end of a completed action in child-directed speech, [Bran 07c; Wred 05].

Those unexpected shifts of high conditional probabilities to a successive pattern (e.g. introduction, action 1) across distinct features (e.g. Eye-Gaze, Words) might be further evidence, that there is no ground truth and amodal boundary or changing point from one subaction to another. As was found out earlier in this master thesis, there is an onset of boundaries annotated on the basis of the video signal compared to audio boundaries.

Comparing features leads naturally to the question of temporal synchrony, which plays a role in some approaches of child-directed-communication research and therefore, has been the basis for audio-visual attention models [Naga 07a; Rolf 08]. There is evidence, that temporal synchrony across modalities (e.g. hearing and seeing something) triggers attention and therefore helps to highlight for example word-object relation or might emphasise the structure of an action [Goga 00; Bahr 00].

This is the reason for discussing another point, whether the time periods or segments *Beg*, *Mid* and *End*, in which feature occurrences have been examined, are still small enough to speak of synchrony of occurrences of different features, such as “Eye-Gaze direction: kind” and “the uttered Word du” (see Table 8.1). As those feature intervals have different lengths and they can occur anytime in the 2-3 second long time periods, there is no guarantee that they start and end at the same time. This is why, exact temporal synchrony is not modelled with this approach, but a tendency for co-occurrences with highly possible interval overlaps or successions, is modelled. Currently under discussion is the issue of, whether there is a need at all, for exact temporal synchrony of features to highlight structural components of an action or a demonstration, containing several actions. Brand and Tapscott suggest, that it might be enough, if features can be associated with a certain pattern [Bran 07b]. For example, different features identified in one pattern could fulfil this demand by occurring close together.

Considering a higher level, a new pattern could result from the patterns of my conditional probability examination: a reoccurring succession of patterns of feature values. This could be the case on present data: Words “and”, “then”, “now”, which have high conditional probabilities at *action Begs* and “fast” hand movements occur most probably at *action Mids*. This new pattern, might in turn emphasise a global aspect of the demonstration scenario, namely, that something similar happened three times in a row. These are only hypotheses, because this conditional probability model does not compute any dependencies between features. But this is exactly, what this conditional probability model is for, to think of some new hypotheses and the interplay of features.

In the next section the conditional probability model itself will be evaluated, to see if hypotheses can be constructed on the basis of this model.

8.1 Evaluation

Evaluating the conditional probability model means to test, how well the results predict new data or how similar results are of different data sets. The main problem for evaluating the present model is that the data set was too small to divide it into a test and training set. For future evaluation, the left out *summary* segments, in which parents repeated their demonstration, could be used to acquire a training set.

The focus of this masters’ thesis is to examine the child-directed demonstration data statistically. It would be a different approach/goal to build for example a classificator.

One still could have divided the data set in order to test and train a classificator, but with my approach only two experiments were left out to have a tiny evaluation of conditional probabilities. One from each experiment setting (Cup and Minihausen) was not used within the computation of conditional probability model. This means, for each feature, two data sets with annotated characteristics are available and were separated on the basis of the word boundary trisection.

The goal is, to evaluate, how many actual feature values would have been placed by the conditional probability model, into the segments, they originally came from. As the conditional probability results do not favour only one segment (for example $P(a2Beg|guck) = 100\%$), scores were associated to the conditional probability outcomes (for each feature characteristic and their conditional probabilities $P(Segment_x|fc)$). The highest conditional probability per feature value is associated with a score of five. The second highest score is associated with a score of 4 and so on. This makes the interpretation of the probability distribution easier, especially for very narrowly distributed results. All conditional probabilities were rounded to integers. Equal results are associated with the same score accordingly, the successive score is left out. For example there are two results with score 5, the next higher result will have the score 3, (see Appendix C for Figures C.7 and C.9, which show scores instead of conditional probabilities for feature Eye-Gaze and Velocity of Hand Movements).

In 89% of the cases the first three scores correspond with the identified patterns (see Table 8.1). A score of 4 was associated for the conditional probability, if a pattern was identified (for the CP of the fc), which consisted of four conditional probabilities.

The segments with corresponding scores of 4-5 or at least the segments with the three highest scores were marked as "expected" for a distinct feature characteristic. For example, suppose the feature characteristic "fast" occurred in *a2Mid*. The corresponding conditional probability model has the following scores associated:

- $P(a1Mid|fast) = 22\% \Rightarrow score = 5$
- $P(a2Mid|fast) = 20\% \Rightarrow score = 4$
- $P(i1Mid|fast) = 19.9\% \Rightarrow score = 4$
- $P(a3Mid|fast) = 17.5\% \Rightarrow score = 2$

Furthermore pattern *Mid* ($P(Mid|fast)$) is identified and $P(a3Mid|fast) = 17\%$ gets a score of 4. All four segments are marked with "expected". For this example "fast" in *a2Mid* would be counted as match between the actual and the expected outcome. In Table 8.2 the percentages (match rate) stand for the number of all matched occurrences of all fc of one feature divided by the number of all seen (occurred) fc ($\frac{\#\ matched\ fc_j}{\#\ all\ fc}$).

The outcome ranges between 28%-63% for the features. Each of the dataset had similar match rates per feature. Table 8.2 shows the match rate per feature. Eye-Gaze (62.67%) and Intonation (59.09%) feature characteristics seem better modelled by the conditional probabilities / score model than Words (44.23%) or Velocity of Hand Movements (28.64%). Fourteen unseen Words occurred, which had naturally no conditional probabilities available, therefore they are marked as not matched. The low match rate for feature Velocity might be due to the fact, that frequent seen feature values “no motion” and “slow” occurred everywhere and were mostly not matched with the expected place (time, segment). It was noticeable, that about 60% of the “fast” and “vfast” movements occurred at *action Ends* in the training set. This might mean that a trisection on the basis of the segment boundaries is more suitable for the feature Velocity of Hand Movements.

Feature	Match Rate
Eye-Gaze	62.67%
Intonation	59.09%
Word	44.23%
Velocity	28.64%

Table 8.2: *Match rates* $\frac{\#\ matched\ fc_j}{\#\ all\ fc}$ *show the rate between actual occurred fcs in a* $Segment_x$*, which have been matched with expected place of occurrence segment divided by all occurred fcs per feature.*

This simple evaluation approach did not work very well. In order to achieve more correct matches, there need to be rules for not seen events and for computed, but not significant or equally distributed conditional probability results for feature characteristics. The overlap or match rate is higher if only feature characteristics with a significant conditional probability model are considered and the match rate is computed fc wise (match rate:$\frac{\#\ matched}{\#\ all\ fc_j}$). Most of the Word and Eye-Gaze characteristics, which have a significant conditional probability pattern, have a high rate. Four of six Word feature characteristics occur to 100% at the expected segments. Velocity features “medium”, “fast” and “very fast” have low rates with word boundary trisection, but with segment trisection their rate increases to > 57%. Features characteristics with round brackets around, (see Tab. 8.3) have no significant conditional probability pattern. Most of them have low conditional probabilities except fc “continuing”, which has a rate of 60%.

Feature	FC	Match Rate	Pattern
Word	guck	50%	first third
	mal	100%	first third
	drauf	100%	action End
	gelber	100%	action 2
	und	100%	action Beg
	dann	50%	action Beg
Velocity	vfast	25% (62%)	Mid
	fast	8.33% (65.2%)	Mid
	medium	38.46% (57.43%)	Mid
Intonation	(rising)	44.44%	action Beg
	(falling)	37.5%	i1Beg/a3End
	(continuing)	60%	action Mid
Eye-Gaze	(objekt)	21.74%	Mid
	kind	52.63%	End
	instruktion	75%	introduction

Table 8.3: *Match rates ($\frac{\#\ matched}{\#\ all\ fc_j}$) show the rate between actual occurred fc_j in a $Segment_x$, which have been matched with expected place of occurrence segment divided by all occurrences of only fc_j. All unseen fcs and fcs with non significant conditional probabilities were ignored. For convenience the conditional probability pattern are included.*

This evaluation (Tab. 8.3) hints, that the significant conditional probability patterns might be used as a basis to predict, where an occurred feature characteristics comes from (and as basis to compute a complex classificator). Those results can be used to construct new hypotheses about how amodal and modal features interact to help structure actions.

.

Chapter 9

Conclusion

This chapter concludes the Masters' thesis "Statistical analysis of characteristics of Infant-Directed-Interaction with respect to action structure". The next section presents some ideas about different segmenting strategies and different features. The thesis concludes with a summary.

9.1 Suggestions

Prior to the statistical analysis, the audio and video signal were divided into successive time periods, which were called segments (*introduction-summary*). These segments were again trisected into a beginning, a middle and an end part (*Beg, Mid, End*), which were called segments or units. In order to make each of the experiments comparable, each task was divided into those segments. It was expected and found, that features exhibit different values at distinct segments. This analysis was conducted to find patterns in the feature value exhibition (precisely: conditional probabilities), which could in future help to recognise the different parts of an action (here: demonstration) only by means of the features and their characteristics and the conditional probability model.

"Precisely how and when each of these mechanisms for action parsing may be available to infants is unknown ..." (Hollich, Hirsh-Pasek, Golinkoff, 2000). Until today, nobody really knows, where or when exactly which of the feature values might be repeatedly used to reveal something about the structure of an action. There are two ways to approach this problem within the scope of this thesis, either to change the segmenting strategy or to use different features.

9.1.1 Segmenting

It is a fact that features of communication change characteristically when the communication is child-directed and some features appear only in child-directed communication. Mostly those feature values are there to catch the child's attention or intensify attention towards meaningful things, which could be word-object relation, special movements or entire actions [Bahr 00]. If all parents repeatedly use the same means to do so, there should be a recognisable pattern. The difficulty is to establish when those means occur, and how to quantify, summarise and compare them across all experiments. Different segments might reveal different compositions of features and that would result in different or clearer patterns. A good example of a possible dissimilarity of outcomes can be seen at the results of feature Velocity of Hand Movements. These results computed on the basis of word segment - and segment boundaries (the segment boundary results were clearer).

Segmenting the task in different ways and on the basis of different signals might contribute to reveal the great picture. Anyway, segments should not be smaller than 1-2 seconds, because the feature characteristics of Word cannot be divided into meaningful smaller parts and too many segments would be empty.

The present approach was to segment the task gapless, so that ends and starts of the segments are at one point in time. This was difficult, because pauses in speech and motion were difficult to assign to any of the segments. Henceforward, there should be a new segment category *pause* introduced, which could be trisected as the others. Pauses are said to reveal transition points, for example from one (sub)action to another, compare

[Bran 07b; Naga 07a]. Moreover, the concept of redundant information hypothesis can be applied, if in segment *pause*, pauses of speech and motion ("no motion") co-occur. Two amodal inputs would behave equally and therefore the *pause* segment (or parts of it) will be emphasised. In addition, the child's attention is caught or at least the pause could give the parent time to check, if the child's attention is still focused and he/she can proceed with the next part (action).

Another point of discussion should be the finding, that there is an onset of segment starts, if the segmentation is made on the basis of the video signal (mostly on body movements) compared to the audio segmentation (mostly on speech). This means, there are different starting times of, actions in each of the two modalities. Segmentation based on only one modality could result in not finding patterns, which are modality specific.

In the present segmentation only the audio segmentation was considered, that means that for example the movement "reaching towards the cup" was most of the time not considered to be part of the next action as speech did not start before then. However, this movement cannot be associated with the last action. Vice versa words meaningful to the last action could be assigned to the next action, because the video segments suggest otherwise.

For a general examination of multimodal features, the solution to the problem above could be, to assume that there is an area of transition from one part (action) to another and should be considered as a distinct segment in addition to the other segments (*introduction-summary*). Figure 9.1 depicts this segmentation idea.

In the same respect, an examination, on the basis of each modality segmentation, could be conducted separately to reveal, for which feature the current modality segmentation reveals a pattern more clearly.

In addition, another approach of designing segments prior to the analysis could be, creating new segments[1] around starting and ending points (*changing areas*) on and *rest*

[1] Segments on the *introduction-action3* segment level

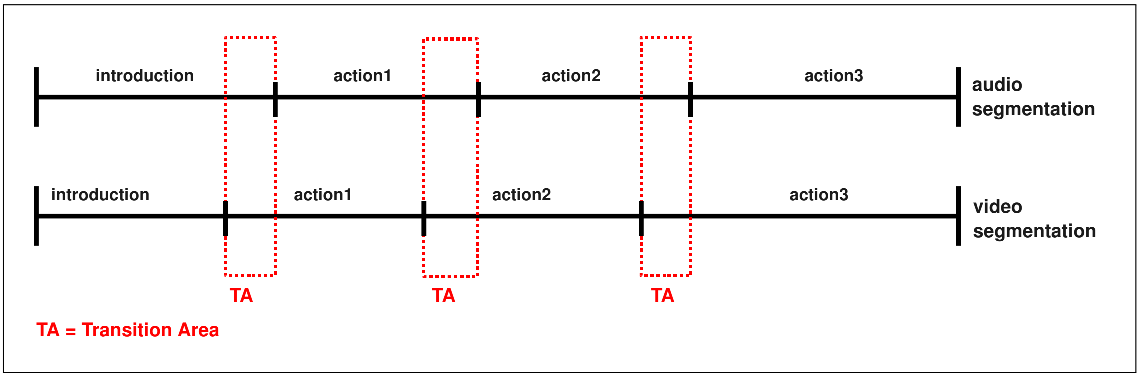

Figure 9.1: *Segmenting strategy: Transition Area is the area between start or end points of the same segment, but different modality.*

segments (see Fig. 9.2). This could also overcome the problem of on- or offset of start and end of segments across different modalities. Those changing area segments might include pauses, which need not be new segments but a new feature or at least a new feature characteristic, which could be beneficial for a statistical examination.

In short *pauses* as new segment category or feature characteristic, could solve the problem of finding an exact start or end point There might not be a fundamental truth change point between last and new (sub)action, there might be an area of change, like pauses in speech and motion suggest.

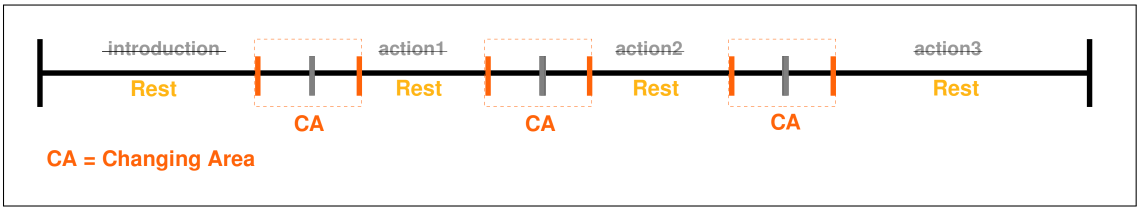

Figure 9.2: *Segmenting strategy: Changing Area segments and rest segments are new segments, created on the basis of the the present segmentation method. Changing areas are around start/end points of the actual segments (introduction-action3)*

9.1.2 Data

Features: As already mentioned in the above Section pauses in motion and speech could play a role in structuring a demonstration. Brand *et al.* states, that pauses and exclamations are salient speech markers [Bran 07b]. Also, Nagai *et al.* report that there are longer pauses between actions in IDI compared to ADI [Naga 07a]. The feature Word could exhibit two new feature characteristics, “a pause longer than normal silence between words” and “all sorts of exclamation and if there is, onomatopoeia”, like “uih”, “ahh”, “ohh” and “peng”, “ding-dong”.

Moreover, words could be summed up into categories, for example, the colour of objects, the adjectives describing the size of objects, because it is not important for analysis of action structures to distinguish between colours and sizes as these vary, of course, with each object. German articles in front of the object could be summed up as well as the names of the toys, especially if one likes to compare different experiment settings with each other. Results of the Words “rein” and “drauf” suggest to sum up prepositions, if they are task related. Naturally, summing up words into categories takes a lot of time and might also be conducted semi-automatically, because based on context, it needs to be decided, if something is task related or not. Another option would be to fine tune the examination or computation of conditional probabilities to antagonise the fact, that the rate of occurrence within a segment will be low if only half or less of the data contains those special words. Features of the same category and those, which exhibit the same pattern, could be summed up after conditional probability computation.

Words, Intonation, Eye-Gaze and Velocity of Hand Movements are the present features. Pitch or sound intensity is also known to have exaggerated values in child-directed speech [Bahr 00]. Like Velocity, Pitch could be divided into categories, for example “no sound”, “quiet sound” ....“very loud sound”. Pitch might reveal exclamations of any kind, as they are often loud.

Data acquiring process: In general, more humans are needed, who code the data (annotate the data), because humans could identify in different ways, during evaluation (for example see Section 3.2). As a result Intonation data should be evaluated by more evaluators and Intonation categories should be assigned word wise. Moreover, humans do make mistakes and therefore cleaning the data takes a lot of time. Coding and acquiring data should be carried out automatically. If possible, when conducting experiments, the setting should be optimal for extracting data automatically with the current annotation and extraction methods.

Naturally, thresholds, which are set manually, are human influenced. Even if they are well considered, the choice could, of course, influence the outcome. The problem is, that continuous values like pitch (decibel) and velocity (pixels per 40 ms) need to be made comparable between all experiments (normalised, smoothed). One approach to make the data comparable is to compute (think of) a threshold, which results in the meaningful feature characteristics (see Velocity of Hand Motion, Section 3.4), but has of course been influenced by human opinion.

9.2 Summary

Four features, namely Word, Intonation, Eye-Gaze, and Velocity of Hand Movements have been examined statistically (two of each available modality (auditory, visual)). Therefore the computed conditional probabilities $P(Segment_x|fc_j)$ answer the question of where, within an action structure (demonstration structure), an occurred feature characteristic most probably comes from? For example, if I hear “guck” (look), what is the chance that this word occurred at the beginning of the task? Prior to conditional probability computation the tasks (experiments) were divided into successive time periods, which were called segments (*introduction-summary*). These segments were again trisected into a beginning, a middle and an end part (*Beg, Mid, End*), which were called segments or units. In order to make each the experiments comparable, each task was divided into those segments. It was expected and found, that features exhibit different values at distinct segments. This analysis was conducted in order to find patterns in the feature value exhibition (precisely: conditional probabilities), which could in future help to recognise the different parts of an action (here: demonstration) only by means of the features and their characteristics and the conditional probability model.

As Chapter 7 (Results) and Chapter 8 (Discussion) state there are some features, which could help to reveal a similarity of structuring the task across the experiments. Mainly two patterns of high conditional probabilities $P(Segment|fc)$ at distinct segments and feature characteristics have been identified: a global pattern and a reoccurring pattern. A global pattern means, that the conditional probabilities have only high values at successive segments and share most of the probability density, e.g. the first third or half of the task or segments in action 1 (*a1Beg, a1Mid, a1End*). A reoccurring pattern has reoccurring high probabilities at segments named either *Beg* or *Mid* or *End.* For example, segments with the name *Mid* possesses more than 50% of the conditional probability density (see Figure 7.1).

The Words “und” (and) and “dann” (then) have high conditional probabilities for segments *action Beg* (all beginnings of actions within a task). The conditional probabilities for segments *action Mid* on condition, that “fast” and “very fast” (Velocity of Hand Movements[2]), have occurred, are significantly high, as well as, the Words “rein” (into) and “drauf” (on top of) would most probably belong to segments *action End.*

The first third of the task (first four segments) have high conditional probabilities $P(firstThird|fc_j)$ for occurring words “ndk” (name of the child) and “guck” (look). The features Intonation and Eye-Gaze have only tendencies to reveal, because the conditional probabilities are nearly distributed uniformly across the segments. A “Rising” Intonation contour tends to occur most probably at segments of *action 2* and “falling” at the very last segment *a3End.* The conditional probability $P(Segment_x|child)$ (kind = Eye-Gaze) are distributed mostly at *End* segments, also $P(action1|child)$ holds a high amount of conditional probability density as well. Eye-Gazes towards the “object” are more likely to found at the *Beg* and *Mid* segments and will be found only with low conditional probability $P(End|object)$ at *End* segments. Moreover, globally seen, parents slightly tended to gaze at the object at the very start of the task ($P(i1Beg|object)$) and stared at the child at end of the task ($P(a3End|child)$).

[2] Example for a “fast” movement: waving object in the child’s focus.

More experiments with different task settings should be conducted to make my findings and their interpretation sounder. However, the presented features are already revealing tendencies about how the identified patterns for feature characteristics help to structure an action in the special case of those demonstrations / tuturing situations.

References

[Bahr 00] L. E. Bahrick and R. Lickliter. “Intersensory Redundancy guides attentional selectivity and perceptual learning in infancy”. *Developmental Psychology*, Vol. 35, No. 1, pp. 190–201, March 2000.

[Bran 02] R. J. Brand, D. Baldwin, and L. Ashburn. “Evidence for motionese: modifications in mothers’ infant-directed action”. *Developmental Science*, Vol. 5, No. 1, pp. 72–83, March 2002.

[Bran 07a] R. J. Brand and W. L. Shallcross. “Six- to 8-month-old infants prefer motionese to adult-directed action”. March 2007. Poster presented at the Society for Research in Child Development, Boston.

[Bran 07b] R. J. Brand and S. Tapscott. “Acoustic packaging of action sequences by infants”. *Infancy*, Vol. 11, No. 3, pp. 321–332, April 2007.

[Bran 07c] R. J. Brand, W. L. Shallcross, K. P. Massie, and M. G. Sabatos. “Fine-Grained Analysis of Motionese: Eye Gaze, Object Exchanges, and Action Units in Infant-Versus Adult-Directed Action”. *Infancy*, Vol. 11, No. 2, pp. 203–214, May 2007. `http://www.informaworld.com/10.1080/15250000709336640`.

[Brea 02] C. Breazeal. *Designing Sociable Robots.* MIT Press, 2002.

[Brea 08] C. Breazeal, A. Takanishi, and T. Kobayashi. “Multimodal Communication”. In: B. Siciliano and O. Khatib, Eds., *Handbook of Robotics*, Chap. Part G:

Social Robots that interact with people, pp. 1349–1369, Springer Berlin Heidelberg, 2008.

[Bucc 05] G. Buccino, L. Riggio, G. Melli, F. Binkofski, V. Gallese, and G. Rizzolatti. “Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study ”. *Cognitive Brain Research*, Vol. 24, No. 3, pp. 355–363, August 2005.

[Chon 03] S. Chong, J. F. Werker, J. A. Russell, and J. M. Carroll. “Three Facial Expressions Mothers Direct to Their Infants”. *Infant and Child Development*, Vol. 12, pp. 211–232, 2003.

[Duda 01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, Inc., 2001.

[Fern 89] A. Fernald. “Intonation and Communicative Intent in Mother’s Speech to Infants: Is the Melody the Message?”. In: *Developmental Psychology 64*, pp. 1497–1510, 1989.

[Fern 98] A. Fernald. “Four-Month-Old Infants Prefer To Listen To Motherese”. In: *Infant Behavior Development 8*, pp. 181–195, 1998.

[Goga 00] L. J. Gogate, L. E. Bahrick, and J. D. Watson. “A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures”. *Child Development*, Vol. 71, No. 4, pp. 878–894, August 2000.

[Holt 99] T. Holter and T. Svendsen. “Maximum likelihood modelling of pronunciation variation”. *Speech Communication*, Vol. 29, pp. 177–191, 1999.

[Inc 07] T. M. Inc. “The Matlab Manual (Part of Matlab)”. 2007.

[Iver 99] J. Iverson, O. Capirci, E. Longobardi, and M. Caselli. “Gesturing in Mother-Child Interactions”. *Cognitive Development*, Vol. 14, pp. 57–75(19), January 1999.

[Kote 06] E. A. Koterba. *Investigating Motionese: The Impact of Infant-Directed Action on Infants' Preference and Learning.* Master's thesis, University of Pittsburgh, July 2006. `http://etd.library.pitt.edu/ETD/available/etd-08082006-162723/`.

[Mack 02] A. Mack, P. Zissis, M. Sivlerman, and R. Gay. "What we see: Inatttention and the capture of attention by meaning". *Consciousness and Cognition*, Vol. 11, pp. 488–506, 2002.

[Masa 92] N. Masataka. "Motherese in signed language". *Infant Behavior and Development*, Vol. 15, pp. 453–460, 1992.

[Naga 07a] Y. Nagai and K. J. Rohlfing. "Can Motionese Tell Infants and Robots. What to imitate?". pp. 299–306, April 2007.

[Naga 07b] Y. Nagai and K. J. Rohlfing. "Parental Signal Indicating Significant State Change in Action Demonstration". March 2007. Poster presented at the 7th International Conference on Epigenetic Robotics, Piscataway, NJ.

[Naga 08] Y. Nagai and K. J. Rohlfing. "Computational Analysis of Motionese: What can infants learn from parental actions?". Vancouver, Canada, March 2008.

[Pete 08] A. Peters. "Gewinnung und Visualisierung von Merkmalen auf dem Motionese Korpus". May 2008. Project Paper.

[Rohl 05] K. J. Rohlfing and T. Jungmann, Eds. *Referenz durch Bewegung: Eine Studie zu Motionese*, Fachgruppe Entwicklungspsychologie (EPSY2005), 17. Tagung in Bochum, September 2005. Poster.

[Rohl 06] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. "How can multimodal cues from child-directed interaction reduce learning complexity in robotos?". *Advanced Robotics*, Vol. 20, No. 10, pp. 1183–1199, 2006.

[Rohl 08] K. Rohlfing. "Intermodal Action Structuring". ZIF Bielefeld, July 2008. http://www.uni-bielefeld.de/(en)/ZIF/AG/2008/07-03-Rohlfing-Programm.html.

[Rolf 08] M. Rolf. *Audiovisual Attention via Synchrony.* Master's thesis, Bielefeld University, June 2008.

[Snow 77] C. Snow and C. Ferguson. *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press, 1977.

[Stri 05] T. Striano and D. Stahl. "Sensitivity to triadic attention in early infancy". *Developmental Science*, Vol. 8, No. 4, pp. 333–343, 2005.

[Stri 07] T. Striano, D. Stahl, A. Cleveland, and S. Hoehl. "Sensitivity to triadic attention between 6 weeks and 3 months of age". *Infant Behavior and Development*, Vol. 30, pp. 529–534, 2007.

[West 08] D. Westhues. *Untersuchung suprasegmentaler Merkmale auf Eltern-Kind-Interaktionsdaten.* Master's thesis, Bielefeld University, June 2008.

[Wiki 08] Wikipedia. 2008. http://en.wikipedia.org/wiki/Main_Page.

[Wred 05] B. Wrede, K. J. Rohlfing, and J. Fritsch. "How can prosody help to learn actions?". 2005.

[Wred 06] B. Wrede, K. J. Rohlfing, and Y. Nagai. "How to make sense of environmental interaction and dynamics. Symposium: Models of infant development: Are we really serious about environmental interaction and dynamics?". Kyoto, Japan, June 2006.

Appendix A

Segmenting

A.1 Segmenting the Task

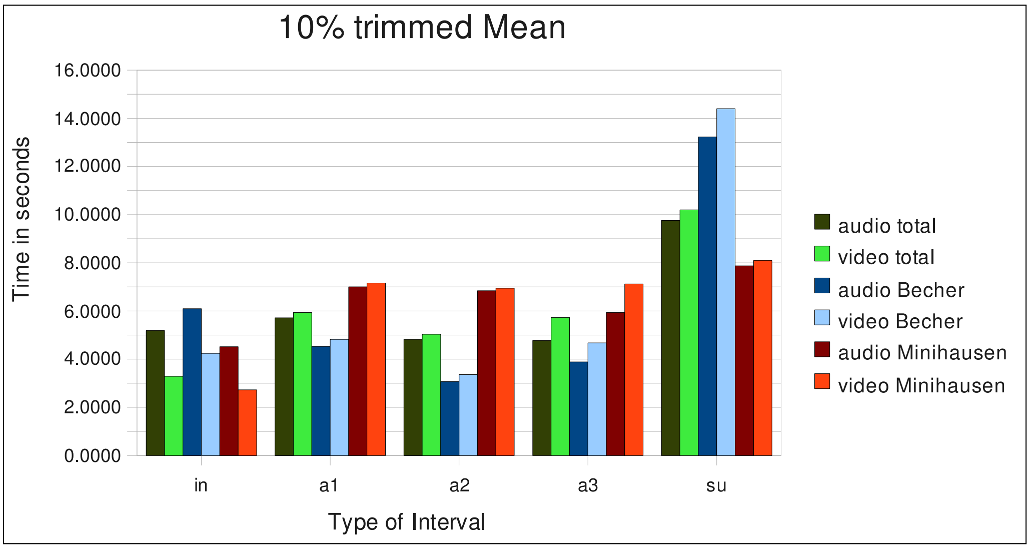

Figure A.1: *Mean of the duration per interval for audio and video annotation in total and per experiment setting (10% trimmed mean, audio=dark color, video=light colour)*

A.2 Onset and Offset

ANOVA of onsets:

- Intervals a1, a2, a3, *s1*: F(3,216)=4.94, p=0.0025
- Intervals a1, a2, a3: F(2,162)=0.13, p=0.8813

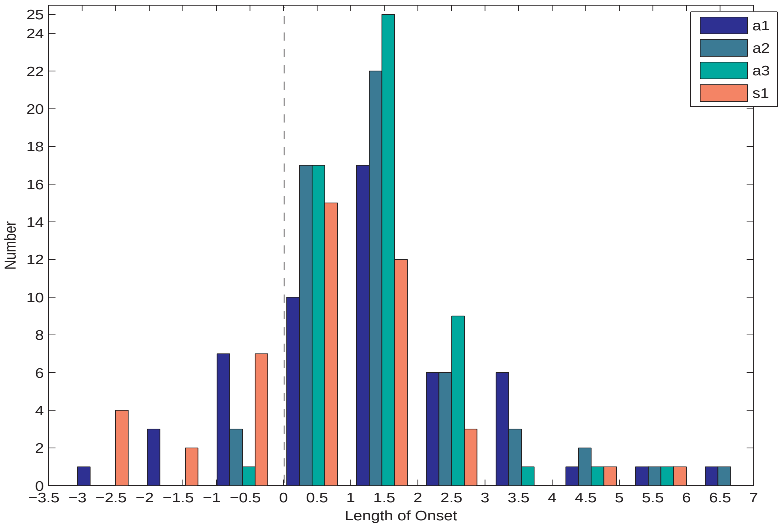

Figure A.2: *The histogram plot represents the duration of Onset of "a1-s1" (result=positive ⇒ video precedes audio)*

Appendix B

Entropy

The maximum entropy can be used to norm the entropy as follows:

$$\begin{aligned}\frac{H}{H_{max}} &= -\sum_{i=1}^{n} P(i) \frac{\log_2 P(i)}{\log_2 n} \\ &= -\sum_{i=1}^{n} P(i) \log_n P(i) \qquad \leq 1\end{aligned} \tag{B.1}$$

With the normed entropy (B.1) (values between 0 and 1) one is able to compare entropies with each other, especially if the number of variables is not the same. The normed entropy is can be used to give a notion about the "uncertainty" of the conditional probabilities for one feature characteristics over all segments but there is no association with the number of yes/no question one has to ask on average. There is one value for all probabilities $\{P(Segment_i|FC), ..., P(Segment_i|FC)\}$. The normed entropy is 0.9584 for the "rising" value (green ball) for example, which is illustrated in Figure 6.3.

Appendix C

More results

C.1 Word

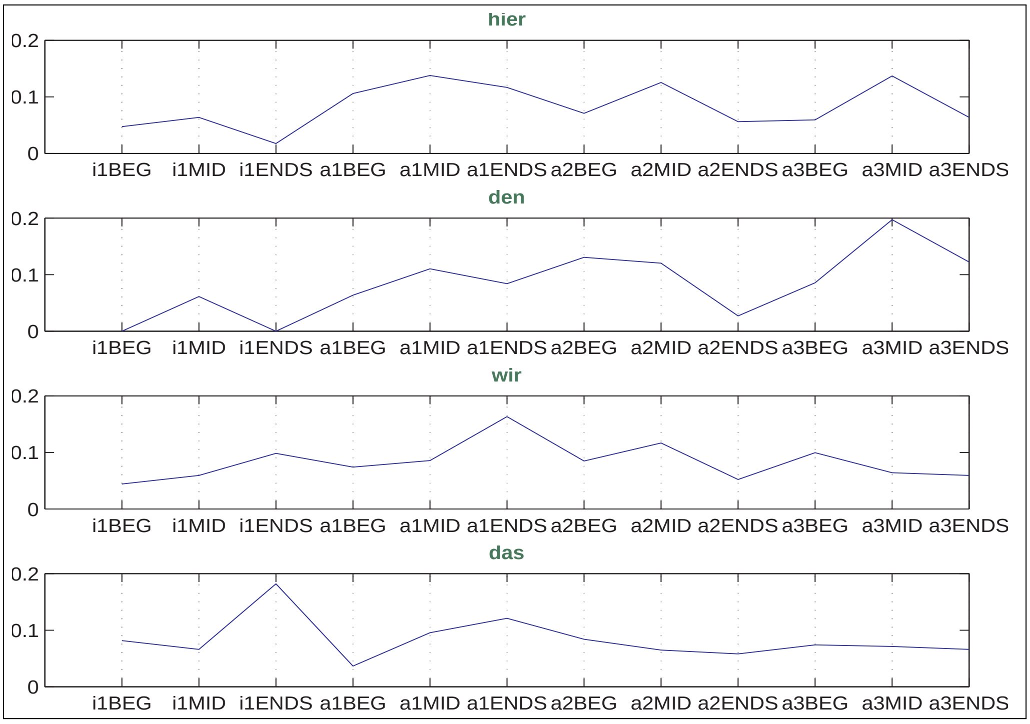

Figure C.1: *Words which are in most segments present and have an entropy value over 3*

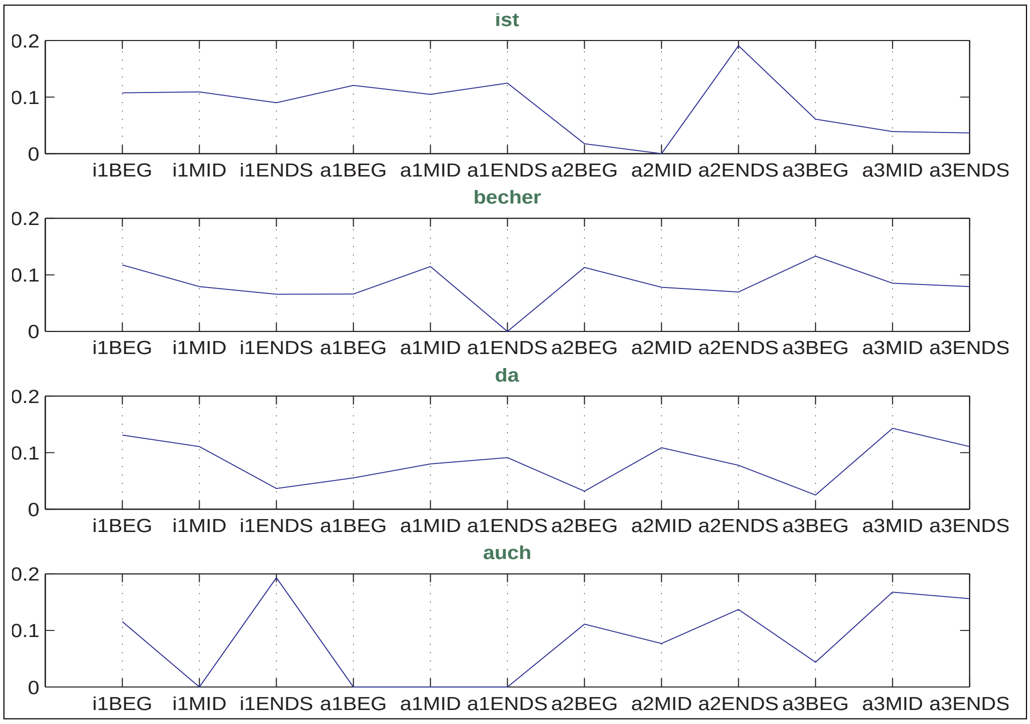

Figure C.2: *Words, which are in most segments present and have an entropy value over 3*

aha
alle
als
alt
am
andere
anderen
anfangen
angefangen
angst
aufpassen
baue
bauen
bestimmt
bild
bist
bleibt
boah
bohrung
bumm
dabei
dach
diesem
dieser
dieses
ding
doch
dran
draufhuepfen
drin
dschii
einer
einfach
einfachheit
ende
erst
erstmal
fangen
farben
fast
foto
ganz
gehen
gib
gibt
glaube
groessen
groesser
groesseren
gross
grossen
guckt
ha
hab
habe
halber
hau
herein
hinein
hinkriege
hinter
hochstecken
huch
huebsch
ihn
ineinader
interesse
jahre
jupp
keine
kenn
klack
klebt
klein
kleine
kloetzchen
koerpernah
komische
kugelstrich
kurz
lachen
laenger
lassen
lustig
macht
mir
mund
nehmen
nimm
nochmal
nummer
obwohl
oder
passen
peng
pfeifen
plumps
plumsen
problem
queh
ran
realisieren
riesig
rot
rum
rund
schoen
schwierig
setzen
sieh
sitzt
sogar
solche
sonst
spatz
spiegel
spielzeug
springen
springt
stellt
suesser
teil
toll
tschueb
tuere
ueberlegen
uih
uihhhh
verschieden
verschiedene
vier
vor
wah
wahrscheinlich
weil
wenn
wertvolle
will
wirds
zeig
zeigen
zip
zusammen
zwar
zweite
ab
aber
an
aus
bisschen
blau
bleiben
bruecke
dein
diesen
durchgucken
fussel
gar
genauso
groesster
grosse
guckst
hast
hast
haus
kannst
kein
koennen
komischen
na
nee
nein
oih
schluss
sie
sieht
sonne
stapeln
willst
wird
wo
zack
zu
zum
baukloetze
denn
einen
fest
im
ineinander
kleinen
kleiner
kleinster
kloetze
maeuschen
muss
oh
pass
selber
sollen
soll
tun
zuhause
zwei
drei
es
geht
machen
mama
mitte
nen
oben
siehst
sind
stellen
turm
was
wie
also
eine
einmal
eins
genau
ok
packen
passt
schon
speedy
dem
hin
kennst
ah
her
klotz
okay
stecken
weg
ach
er
gut
hallo
loch
hm
ich
mit
nicht
auf
ne
blauer
man
kann
noch
roter
mhm
auch
jetzt
becher
dann
ein
in
du
rein
ja
die
wir
so
gruener
kommt
gelber
drauf
ndk
ist
den
der
da
das
hier
und
guck
mal

Figure C.3: *280 words occurred. The list starts with words, which occurred only once (overall rate = 0.06%), the words in red colour indicate that the overall rate has been increased. Please read column wise. The last word "mal" has an overall occurrence of 5.17%*

C.2 Intonation

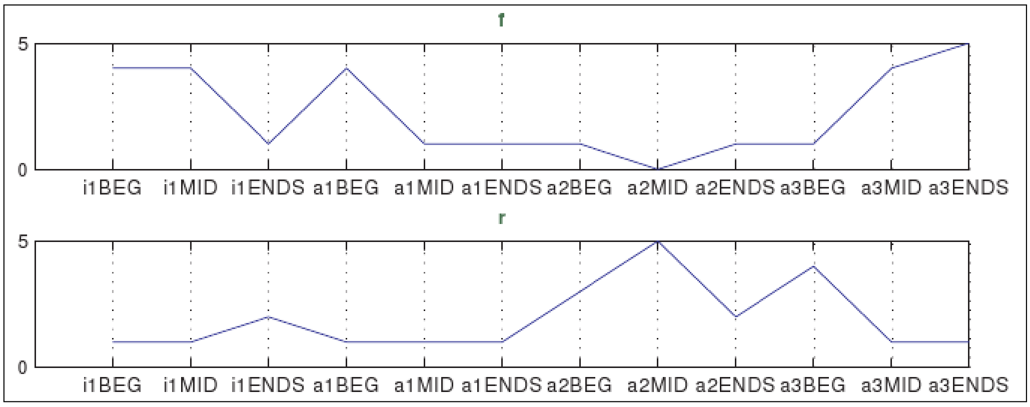

Figure C.4: *Conditional Probabilities for "unsure"(x), low overall rate:* $\frac{17}{569}$

C.3 Eye-Gaze

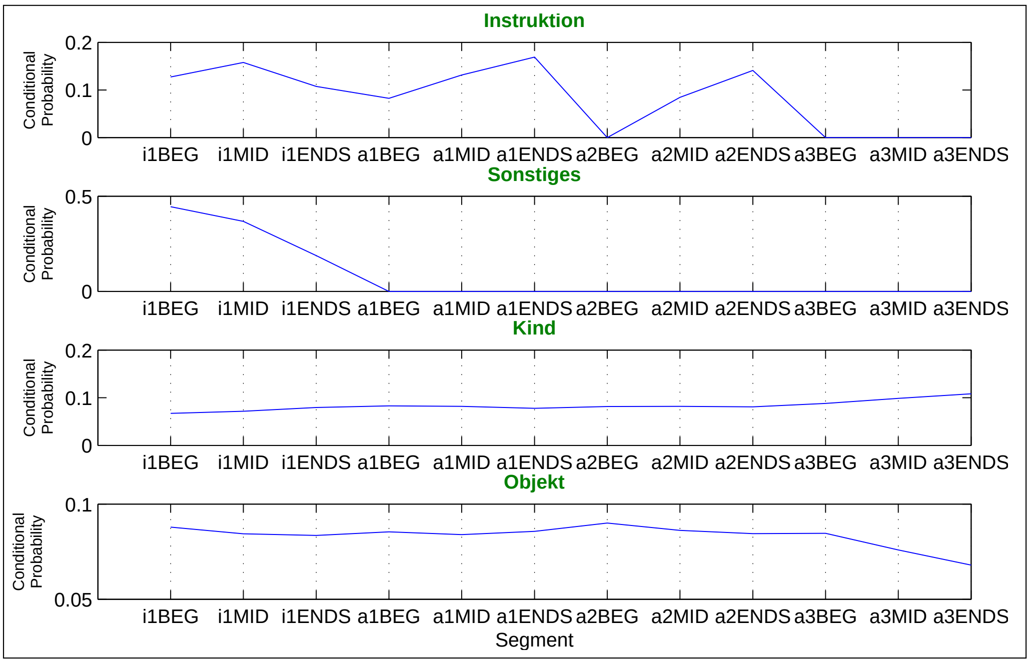

Figure C.5: *Global and reoccurring for conditional probabilities (CP) $P(Segment_x|fc_j)$. fc_j stands for any of the four Eye-Gaze directions: Instruktion (instruction), Sonstiges (miscellaneous), Kind (child), Objekt (object). CP are created on basis of audio segment boundaries*

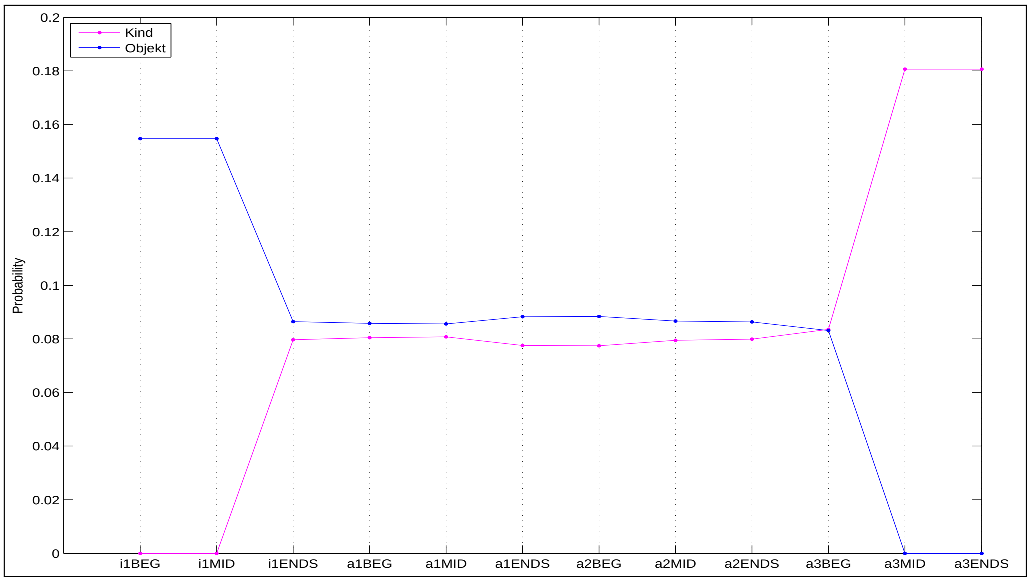

Figure C.6: *Global pattern for Eye-Gaze characteristics "Kind" ($P(Segment_x|child)$) and "Objekt" ($P(Segment_x|Object)$) in one diagram created on basis of audio boundaries, only taking everything over 52 into account.*

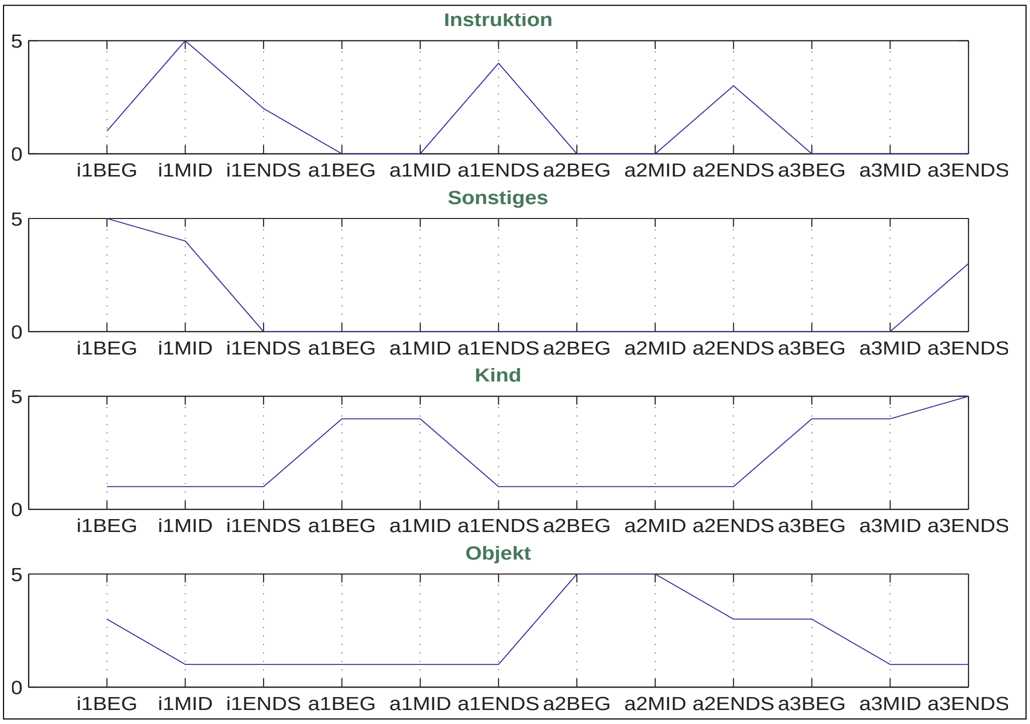

Figure C.7: *Graphs for features have scores as y-value, which is based on the distribution of the conditional probabilities (cp). The highest cp is associated with a score of five. The second highest score is associated with a score of 4 and so on. This makes the interpretation of the probability density easier, especially for very narrow distributed cps.*

C.4 Velocity of Hand Motion

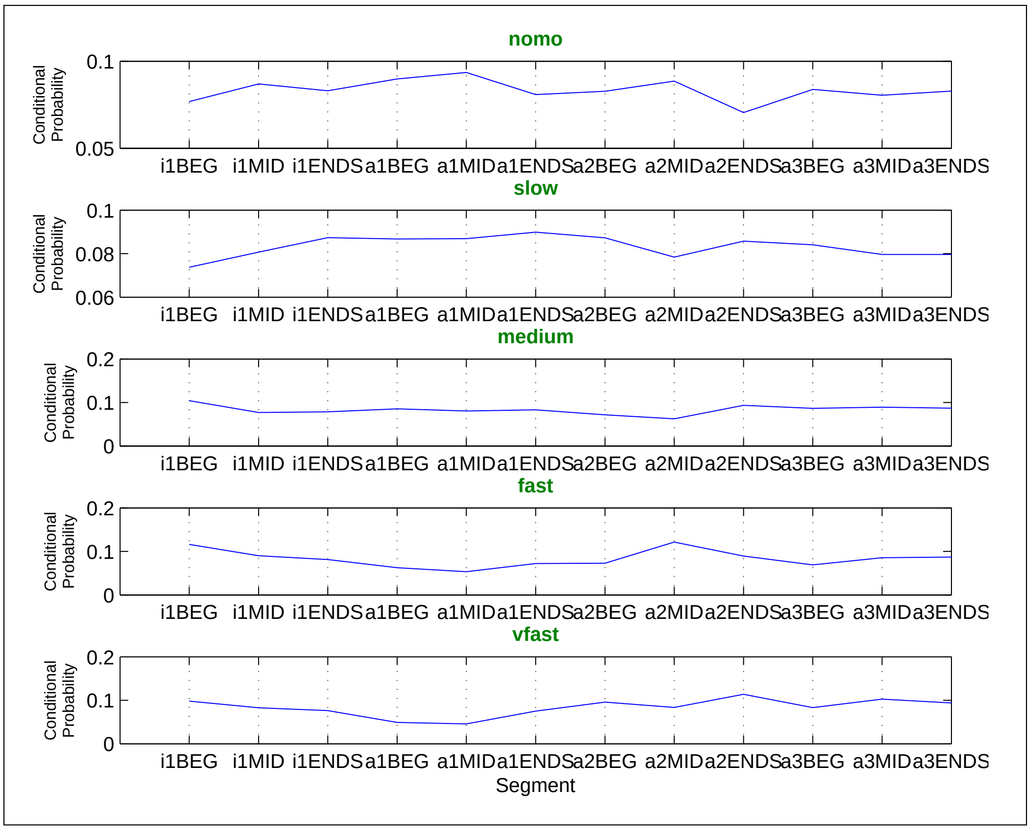

Figure C.8: *Conditional probabilities $P(Segment_x|fcj)$ for all feature characteristics nomo (no motion), slow, medium, fast, vfast (very fast). Trisection of segments on basis of the word boundary. No limitation of interval selection per segment. Entropy value 3.5 bits.*

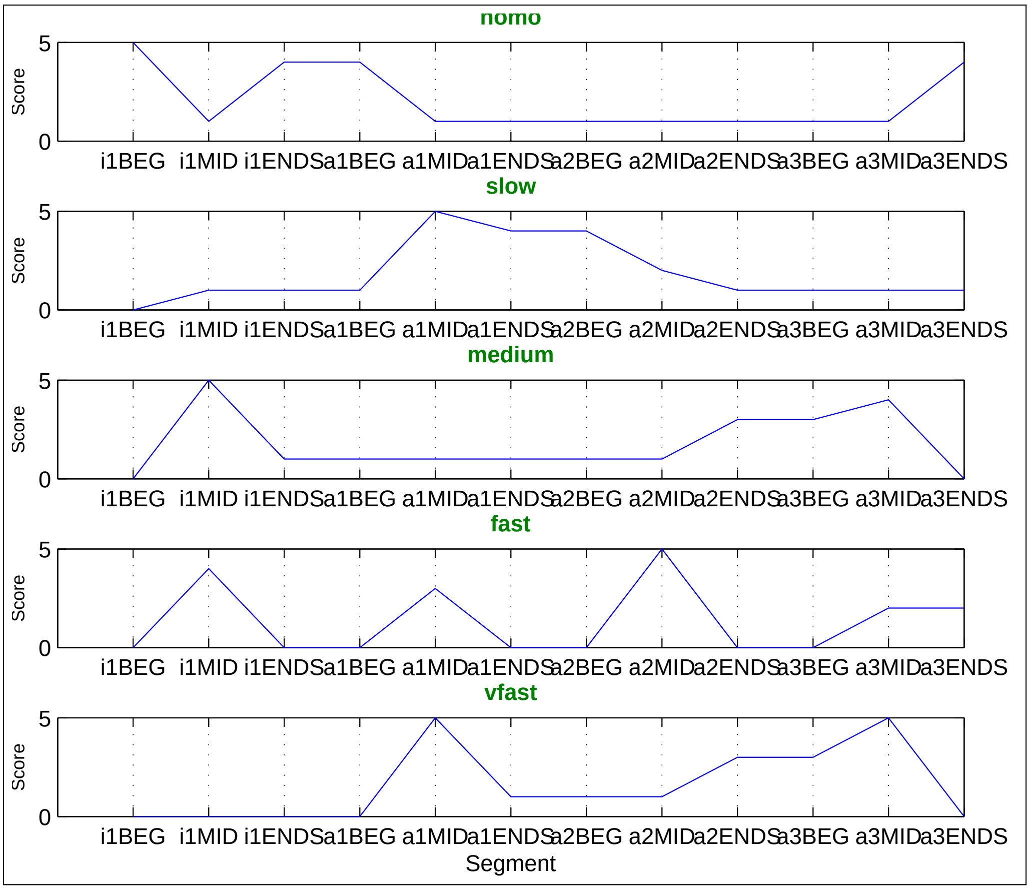

Figure C.9: *Graphs for features have a score as y-value, which is based on the distribution of the conditional probabilities (cp). The highest cp is associated with a score of five. The second highest score is associated with a score of 4 and so on. This makes the interpretation of the probability density easier, especially for very narrow distributed cps.*