Ugur, E., Nagai, Y., Oztop, E., and Asada, M. (Eds) Proceedings of Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human-like cognitive abilities? November 29, 2012. Osaka, Japan

Keyword Detection in Human-Robot Tutoring Scenarios

Christian Dondrup[§], Katrin Solveig Lohan[*], Joe Saunders[‡], Hagen Lehmann[‡], Chrystopher Nehaniv[‡] and Britta Wrede[§]
[§]Bielefeld University, Research Institute for Cognition and Robotics, Applied Informatics Group
[*]Instituto Italiano di Tecnologia , Robotics Brain & Cognitive Sciences
[‡]University of Hertfordshire, Adaptive Systems Research Group

***Abstract*—We describe a way of narrowing the search space for descriptive keywords during a human-robot tutoring scenario, where the tutor is explaining names and characteristics of objects to the robot, by employing interaction detection techniques. This system detects attention getting behaviour which is derived from mother-infant interactions and extracts the verbal information during these specific time periods, segmenting it and building up histograms to estimate word frequencies and thus word importance. This method should allow us to create a system that does not rely on a dictionary or normal speech recognition to acquire novel word-object relations but only relies on the pure interaction between the robot and a human tutor.**

I. Introduction

In human-robot interaction speech is an important way of communication. To achieve a natural interaction between a human and a robot we have followed the developmental robotics approach [1] with the intend to create a model for keyword acquisition gained from previous research on adult-child interactions. We have previously studied preverbal infants (6 to 8 months) in an interaction with their parents for clues on how infants learn words (see [2], [3]).

Most speech acquisition approaches typically use a predefined dictionary and a common speech recognition algorithm or manual annotation, see for example [4], [5]. These methods are well suited to their application and more or less accurate in their results, however, we want to build a system that is able to learn important keywords on its own. Such a system should be capable of learning online, so we cannot rely on manual annotation. In addition we want to build a system that profits from the learning behaviour shown by preverbal infants to prevent the need for predefined words which are then recognized by the robot.

In this paper we will show a way of reducing the search space for important words in a human-robot tutoring scenario by emulating the behaviour of preverbal infants, thus trying to achieve a tutoring behaviour in the human tutor which is, as similar as possible, to the behaviour of a mother playing with and teaching her child [6]. We will present a specific scenario where the human tutor is teaching the robot some objects, specifically their names, colours and shapes. We will try to describe a way to encourage the tutor to use more descriptive words (e.g. red, small etc.) than filler words (this, and, here etc.). By this we hope to achieve a search space that allows us to identify important words without knowing their meaning and thereby create a way of learning them with some kind of unsupervised learning algorithm.

Fig. 1. One of the participants explaining shapes and colours to the iCub [9] robot. The shape explained is the blue sun inside of an ARToolKit [10] marker for the object detection.

To recognize the situations where the tutor is more likely to use keywords for the object description we will employ a detector that relies purely on the interaction between the tutor and the robot. Afterwards we will have to segment [7] the recorded speech data and identify similar words [8] to determine a word scale to find the most important ones.

II. Adult-Infant Interaction

The following section describes visual clues that allow us to narrow down the keyword search space. All these clues are derived from the interaction between a mother and her preverbal infant (6 to 8 month) [6]. This is due to the fact that we want to teach our robot novel-words in relation to an object, therefore, we will have a look at how mothers teach their infants such object-word relations. The preverbal infant condition was chosen because we want to learn words that are new for our robot (no internal dictionary), so we need to have a look at infants that need very much assistance with their language acquisition.

A. Maternal Object Naming

As Matatyaho and Gogate describe in their paper [6] there is a connection between the movement of the object and naming it as a part of our natural behaviour in a mother-infant teaching scenario.

Matatyaho and Gogate [6] state that preverbal infants learn an object-word relation better if their mothers use attention getting gestures like forward motions and shaking (or waggling) while synchronously naming the object, compared to infants where the mothers did not use such techniques. So Matatyaho and Gogate [6] have shown that uni-modal (visual) properties occur in combination with inter-modal (synchrony) properties or maternal naming. As described above the mothers used showing gestures like forward and shaking motions more often in synchrony with naming the object than in asynchrony (these findings are consistent with the field studies of Zukow-Goldring [11], [12]). As a result of this Matatyaho and Gogate [6] state that these gestures in synchrony with words are naturally effective tools for conveying novel word-referent relations because they likely elicit greater infant joint attention and thereby facilitate the word mapping.

We hope to exploit this teaching technique for our human-robot tutoring scenario by relying on the natural attention getting behaviour of the human tutor. For this purpose we will try to implement a behaviour for the robot that encourages such attention getting gestures and hopefully synchronous object naming, thus narrowing down the search space for meaningful words without knowing what these words actually mean.

B. Looming

As stated by Matatyaho and Gogate [6] mothers use forward and upward/downward[1] and shaking or waggling gestures as attention getting movements. Since we do not care about the position of the tutor in respect to the robot we ignore the upward/downward movement which is combined with the forward movement (Matatyaho and Gogate [6] collapsed forward/downward movements into one because they often co-occurred at the same time) but will just concentrate on the forward movement itself. These forward movements which intend to bring an object into the line of sight of the infant (or robot in our case) are also called *looming*.

If we use this looming behaviour to narrow down our search space, we are more likely to get meaningful information as a result. As a logical consequence we disregard all the other verbal information that is given during the non-looming phases and just process the verbal information given during the looming phases.

C. Robot Behaviour

To induce looming gestures we will have to design a behaviour for our robot that shows some kind of reaction to the looming itself. One way of giving such a feedback would be gazing at the loomed object. This gaze switching to the object and thereby creating a state of joint attention is one of the main features that helps infants learn the relation between the spoken word and the described object [6]. As a result the obvious choice to reward looming would be the joint attention to the object by looking at it. However, to encourage the tutor to use as much looming as possible the robot has to reach a habituation[2] [13] state at some point during a looming gesture and thereby loose interest in the object and show its lack of attention by looking away (at random points for example). This is supposed to trigger as much looming in the tutor as possible, and thus help us gather more meaningful information about the object, by being sensitive to the ostensive stimuli and giving feedback about the capabilities of the robot and thereby creating an environment where the robot is treated infant like [3].

[1]Depending on the position of the tutor in respect to the infant.

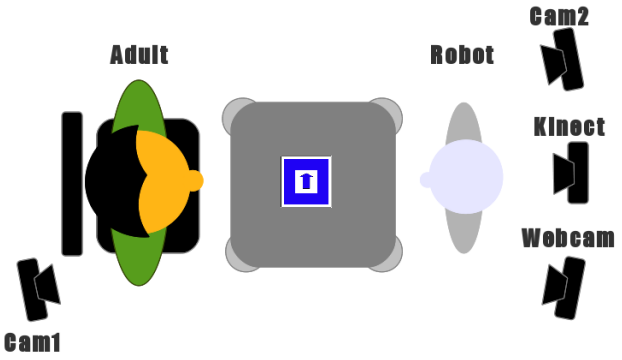


Fig. 2. This figure illustrates the study setting where the human tutor and the robot are placed on opposite sides of a table. The scene is captured by two cameras one facing the human and one facing the robot. In addition we need a Microsoft Kinect for the looming detection (see III-B), a webcam for the object detector and a headset for the voice recording (not shown in this figure). On top of the table is one of the cubes showing the blue arrow shape inside an ARToolKit [10] marker for the object detection.

III. Study

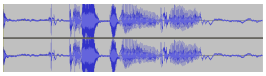
After we have shown how a robot should react and behave to facilitate the acquisition of meaningful data (see II-C) we will now describe a related study which was carried out in the italk project and conducted at the University of Hertfordshire in the beginning of 2012.

A. Parameters

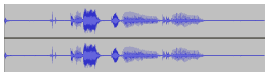
1) Set-Up: We observed 19 participants, which are native English speakers, teaching the iCub robot [9]. The participants were divided into 2 groups which differed in the behaviour the robot showed. The first group was confronted with a random gaze switching, non-responsive[3] robot and the other half taught a robot showing a behaviour according to the Tutor Spotter [14]. The Tutor Spotter tries to create a contingent tutoring environment by showing joint attention according to the gazing

[2]Our definition of habituation differs in the way that not repeated but persistent stimulus triggers the habituation and after the stimulus vanishes the system will immediately recover from said habituation.

[3]Less contingent, does not respond to gazing and looming behaviour of the tutor.



(a) Original signal of three words in one utterance recorded by a headset. Still pretty noisy and distorted. The boarders between the last two words are not very clear.



(b) Signal from Figure (a) after passing through the pre-emphasis and band-pass filter. We can now even make out the borders between all three words.

Fig. 3. Original and filtered Input signal.

behaviour of the tutor [14]. Looming behaviour is rewarded by pointing at the loomed object, thus trying to heighten the joint attention.

The participants had to partake in 3 sessions which had at least one day in between them. In the second and third session the robot spoke back to the participant [15], but this is only mentioned for the sake of completeness and will not be important for our analysis since we will only regard the first session.

2) Task: The task for the participants was to teach the robot about different shapes, sizes and colours. As objects they were given 3 different sized cubes (small, medium and large) with different shapes (sun, heart, cross, circle, arrow and crescent moon) in different colours (red, green and blue) on them.[4] The participants were then advised to explain these 3 different characteristics (sizes, colours, shapes) to the robot in any way they like for about two minutes in each session. In Figure 1 we can see one of the participants explaining the medium sized cube with the blue sun shape facing towards the iCub [9] and Figure 2 shows the general setting during the experiment.

B. Looming Detection

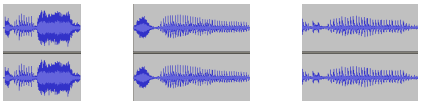
In Section II-B we defined what part of the interaction is meant to help us distinguishing important from less important words. To utilise this we have to detect the looming behaviour of the tutor. We will not talk about the object detection since the object tracker always has to fit to the specific problem[5], but will just focus on the hands to not go beyond the scope of this paper.

We used a Microsoft Kinect camera (as seen in Figure 2) to get a 3D image of the scenery and used the ability to track the position of the hands in 3D space provided by the OpenNI [16] framework. The only value we will observe is the distance (z-coordinate) of the hands to the Kinect camera which is placed behind the robot. These distances are the only important informations for our looming detection since every movement of bringing the object into the line of sight of the robot includes a forward movement [6]. Our approach in the mentioned study (see III) was to use just a fixed distance δ which had to be undershot to trigger the looming detector:

$$L = \begin{cases} \text{true} & \text{if } d < \delta \\ \text{false} & \text{else} \end{cases} \quad (1)$$

[4]On each side of the cubes was only one shape in one colour to make it unambiguous which object is explained.

[5]We used a standard ARToolKit [10] marker tracker.



(a) (b) (c)

Fig. 4. The resulting signal segments after the automatic segmentation into single words.

Where d is the current distance of the hand to the Kinect and δ is the threshold for the looming detection which has to be obtained by manual calibration and testing.

Now that we can detect looming behaviour we have to record the voice during these parts of the experiment and segment it into words [7].

IV. Segmentation into Words

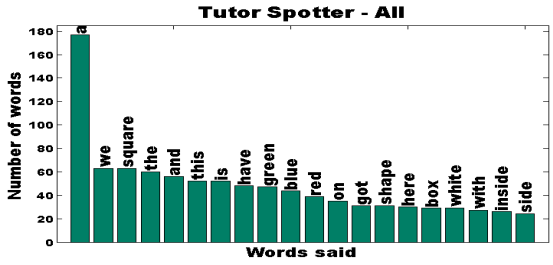
When recording sound we will always get some kind of noise that is distorting the signal we want to process. We can of course try to minimize that by using a headset or unidirectional microphone arrays (in our case we used a headset to enhance the audio track of one of the cameras) but, nevertheless, we will always get some kind of background noise or distortion. To get rid of almost all of the unwanted information in the signal we used two algorithms of noise reduction as suggested in the paper by Waheed et al. [7].

1) Pre-emphasis Filter: At first the incoming speech signal is preprocessed using a so called pre-emphasis filter: $y(n) = x(n) - \alpha \cdot x(n-1)$ where n is a discrete time step and $x(n)$ is the corresponding value. The α represents the pre-emphasis factor which usually is 0.95. The pre-emphasis filter in general is used to reduce differences in power of different components of the signal. In speech recognition the pre-emphasis filter is used to *"[...]reduce the effects of the glottal pulses and radiation impedance."* [7] and *"It takes the focus to the spectral properties of the vocal tract."* [7].

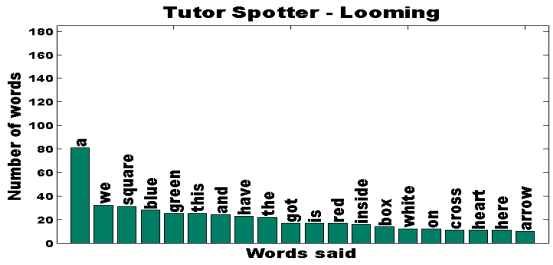
2) Band-Pass Filter: The second algorithm which is designed to reduce low frequency background noise and remove high frequency noise spikes [7] is a band-pass filter. This filter basically consists of a high-pass filter and a low-pass filter. So the band-pass filter passes through frequencies in between an upper and a lower border.

A. Segmentation

To segment the signal we used an algorithm based on an entropic contrast suggested by Waheed et al. [7]. After the signal has been filtered (see IV-1, IV-2) it is divided into windows of 1024 frames (at a signal frequency of $44100Hz$) with a 25% overlap which is then passed into a histogram with 100 bins to determine the probability distribution for that individual frame. The entropy of each of these individual windows is then computed by the standard entropy formula [7]: $H = -\sum_{k=1}^{N} p_k \log_2 p_k$. This gives us a list of entropies which are used to construct the entropy profile $\xi = [H_1\, H_2\, \cdots\, H_m]$ with m total windows of 1024 frames in the signal. From this entropy profile we can now choose a



(a) Words said during whole session.



(b) Words said during looming phases.

Fig. 5. Two histograms of the 20 most said words during the first session of the study. Both histograms are taken from the Tutor Spotter [14] condition.

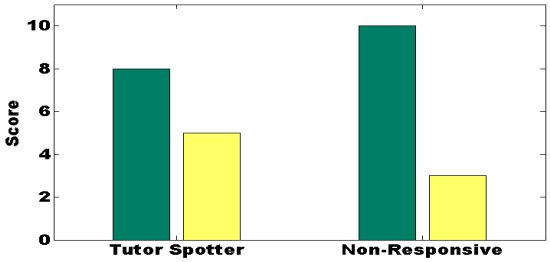


Fig. 6. The scores for the histograms of the two different conditions for the first session of the experiment. The histograms gained a point for every keyword (14 keywords in total) that was listed first. So e.g. if *blue* turned up first in the looming phase histogram then looming gained a point and vice versa. Green: looming, yellow: whole session.

biased threshold to *"[. . .]minimize excessive influence of the background noise."* [7]:

$$\gamma = \frac{\max(\xi) - \min(\xi)}{2} + \mu \cdot \min(\xi) \qquad (2)$$

The bias is defined by $\mu \cdot \min(\xi)$ where $\mu > 0$ and $\min(\xi)$ represents the residual noise floor. After defining the threshold we can consider every window with an entropy above the threshold as speech and every window with an entropy below the threshold as noise [7]. The problem with that assumption is that in many cases non-speech data can be reported as speech data due to artefacts. Also some valid speech data may be ignored because of its physio-vocal characteristics. So Waheed et al. [7] suggest two further criterions in addition to the threshold to determine whether a segment contains speech or not.

The first criterion is the size of the found speech segment $\lambda_i > \kappa$ where κ symbolizes the duration of the shortest phoneme in the target language. Because, *"Humans generally do not produce very short duration sounds."* [7]. The second criterion is the inter-segment distance d_{ij} between the segments i and j. This criterion is required because there can be parts of speech that have been separated into two segments due to its pronunciation [7]. So the criterion is $d_{ij} < \delta$ where δ is the maximum inter-segment distance.

As our final distinguishing criterions to determine speech segments we now have our threshold and if λ_i or $\lambda_j > \kappa$ and $d_{ij} < \delta$ the two segments i and j are merged and the space in between will be considered part of the speech, too. On the other hand if $\lambda_i < \kappa$ and $d_{ij} > \delta$, then the segment i will be discarded and thereby considered noise.

The problem with this automatic segmentation algorithm is that the algorithm will just find sentence boundaries or the shortest utterances if the speech is very continuous [7]. Since we are expecting one or two word sentences during the looming phases we hope to achieve good results, nonetheless.

V. Identifying Similar Words

To determine which words are most important for the object description, and by that which words we have to learn, we will need a way to make an assertion about which words are similar to previously heard words. By that we can construct a histogram of words and hope that the most used words are the most descriptive ones. As one possibility to do so we suggest an approach that is similar to the audio fingerprinting algorithm introduced by Yan Ke et al. [8] which is used by the music industry.

This approach was chosen because of its high reliability, the insensibility to noise and the possibility to find single words in longer utterances which compensates for the segmentation where not all of the words can be segmented due to continuous speech. This algorithm Fourier transforms the sound signal and treats it as a 2D image. By that they try to *"[. . .]employ geometric verification in conjunction with an EM-based occlusion model to identify the song that is most consistent with the observed signal."* [8]. These 2D images represent spectrograms of the given signal and could be compared directly by using correlation. This however would be too slow and inaccurate so Yan Ke et al. [8] suggest to use a small set of filters that are robust to small distortions and still give us enough information to distinguish between two different signals. After viewing the spectrogram images Yan Ke et al. [8] suggested that the filters introduced by Viola and Jones [17] are most suitable for their needs. To select a descriptive subset of these filters Yan Ke et al. [8] use a pairwise boosting algorithm that differs from the standard Adaboost [18], [19] in the fact that they only re-weight pairs of filters instead of single filters since their suggested weak classifiers cannot do better than chance on their own. After creating this subset of filters they are used to create a set of descriptors for overlapping windows of the signal.

These descriptors are written to a file and stored in a

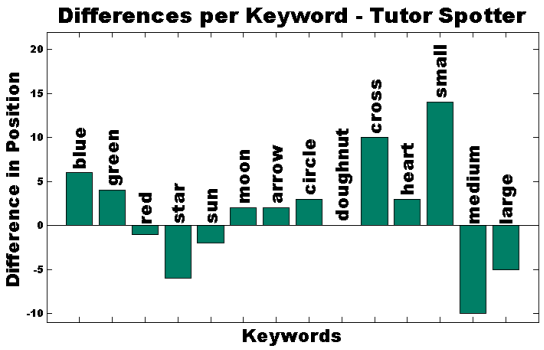


Fig. 7. Tutor Spotter [14] condition: The difference in position for each single keyword. Positive means the keyword moved up this number of ranks in the histogram when comparing looming with the whole session. Negative means the keyword moved down this number of ranks.

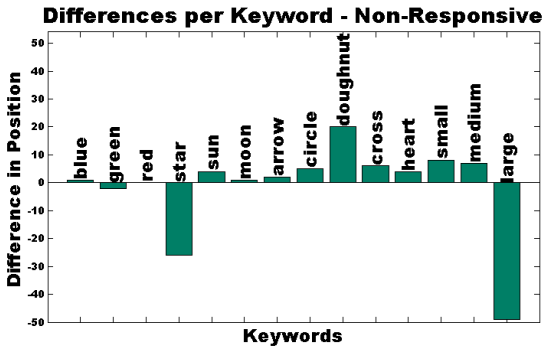


Fig. 8. Non-Responsive condition: The difference in position for each single keyword. Positive and negative values are defined as described in Figure 7.

database and this database can then be queried using other descriptor files. So we are now able to construct a histogram of words (by running cross references) without knowing the actual word itself but just knowing its number of occurrences in the recorded signals.

VI. Results

After we have seen how to construct a system that detects the looming behaviour and builds histograms of the said words we will go back and have a look at our study again (see III).

To show that it will be more likely to learn an important keyword[6] during a looming phase than during the whole session we will employ a histogram based analysis as suggested in Section V. For our analysis we just considered the first session to prevent any kind of learning effect from falsifying our results. In Figure 5 we can see two histograms which resulted from the first session of all participants facing the Tutor Spotter [14] condition. Figure 5b still shows that even during the looming phases the most said word is *a* but the amount is considerably smaller than during the whole session as we can see in Figure 5a. We can also see from Figure 5 that the first 3 words are identical in regard to their position in both histograms. The first real change is the 4th word which is *the* for the whole session but *blue* for the looming phases. To highlight this effect of keywords moving up the ranks in the histogram we can have a look at Figure 6 which shows scores depending on the position of the keywords in the histogram. The higher the score the more keywords are mentioned first in the related histogram. So the looming phases generate histograms that contain more keywords on higher ranks than the histogram over the whole session. The difference in ranks which the keywords moved up or down to is pictured in Figure 7 and 8. In Figure 7 the rank difference is 20 which means that we have a gain of 1.43 ranks per keywords on average. In Figure 8 we have a rank gain of -19 which is dominated by one outlier. If we disregard the outlier we achieve a rank gain of 30 which means an average rank gain of 2.31 per keyword.

[6]Keywords used for this analysis are: red, green, blue, star, sun, moon, arrow, circle, doughnut, cross, heart, small, medium and large.

VII. Conclusion

We have seen in the Results Section VI that we have achieved to move the keywords up in the histogram if only regarding the looming phases. This is the essential result if we want to rely on these histograms to create a database of words associated with one particular object. On the other hand we have also seen that we still get a lot of filler words that are meaningless for the object description. These filler words will still have to be filtered out by running a cross reference between the different objects like the Inverse Document Frequency [20] algorithm. The presence of the unwanted information may be explained by the experiment design which was not built to induce looming behaviour in particular but to evaluate the Tutor Spotter [14] (see VIII-2). However, it still narrows down the search space, nonetheless. Figure 6 shows that the Tutor Spotter [14] condition yielded less keywords that moved up in rank than the Non-Responsive condition. This could be due to the reason that the Tutor Spotter [14] itself already creates a state of joint attention triggered by the gazing behaviour of the tutor. Thus, it implies higher cognitive function which results in less use of the synchronous naming behaviour as implied by [6]. But there still is a gain in ranks which we can see in the Results Section (VI). The Non-Responsive condition tends to yield better results because of the general inattention which induces attention getting behaviour (see Figure 6 and 8).

This leads us to the conclusion that considering the looming phases as a clue for meaningful keywords will narrow down the search space and improve the possibility of finding keywords at the top of the histogram if we use an experimental set-up which is either inattentive and/or rewards looming.

In addition to the narrowed search space we gain an unambiguous clue which object these keywords are describing because, as Matatyaho and Gogate [6] state, the attention getting gestures also help to highlight the object-word relation by highlighting the object through movement. So we found a way of combining meaningful words with objects without knowing anything about the object or the word.

VIII. Future Work

1) Looming Detection: The looming detection method we used in the study (see III-B) is very robust concerning the actual detection of looming but also very fragile in regard to calibration and changing environments, e.g. a different position of the tutor that brings him closer to the Kinect could trigger the looming more easily. Due to this problem of exact calibration and adaptability we will follow a new approach of looming detection for future studies which relies on the mean distance of the hands and the variance of that distance. The idea behind this is that the hands (or at least one of the hands) of the tutor will be in front of him while explaining the object of interest. During the explanation phase the tutor will create his own *explanation space* where he moves the object about freely in his normal way of describing it. This so called *explanation space* will be defined by the mean distance over the time and a certain variance to compensate for normal purposeless movement.

Looming will now be triggered if the tutor moves the object in his hand out of the *explanation space* towards the robot (Kinect). To create a more robust detection the hand has to move towards the robot for a minimal distance of twice the radius of the variance sphere. So we end up with the following conditions:

$$L = \begin{cases} \text{true} & \text{if } d < \mu \wedge |d - \mu| > \sigma \cdot 2 \\ \text{false} & \text{else} \end{cases} \quad (3)$$

Where d is the current distance of the hand to the Kinect, μ is the mean distance over time and σ is the standard deviation which equals the square root of the variance $\sqrt{\sigma^2}$.

With this method of detection we hope to achieve a more natural looming detection and an easier experimental set-up and are hopefully able to construct a system that induces looming in a more robust fashion.

2) Future Studies: We believe that the presented study (III) was not optimal to test the real abilities of our system since it was designed to show the benefits of the Tutor Spotter [14], and therefore hope to conduct a new study where we can test an experimental set-up that is tailored to our needs with a system that obeys the rules of creating joint attention when looming is detected like stated in [6] (see Section II-C). We hope to achieve better results and show a more significant difference in rank gain for the desired keywords by doing so.

Also, our system was not implemented and running at the actual study. So we hope to show that with a running system during an experiment we can actually learn at least some of these found keywords and associate them with the given object.

Acknowledgement

We would like to thank for the support of Katharina Rohlfing. This research has been supported by the EU project RobotDoC (235065) from the FP7 Marie Curie Actions ITN and by the EU FP7 Project ITALK (ICT-214668) within the Cognitive Systems and Robotics unit, as well as from the Citec and the IIT.

References

[1] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori *et al.*, "Integration of action and language knowledge: A roadmap for developmental robotics," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, no. 3, pp. 167–195, 2010.

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

[3] K. Lohan, A. Vollmer, J. Fritsch, K. Rohlfing, and B. Wrede, "Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations?" in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[4] T. Plötz and G. Fink, "Robust time-synchronous environmental adaptation for continuous speech recognition systems," in *Proc. ICSLP*, vol. 2. Citeseer, 2002, pp. 1409–1412.

[5] G. Fink, "Developing hmm-based recognizers with esmeralda," *Lecture notes in computer science*, pp. 229–234, 1999.

[6] D. J. Matatyaho and L. J. Gogate, "Type of maternal object motion during synchronous naming predicts psreverbal infants' learning of word-object rselations," *Infancy*, no. 13:2, pp. 172–184, 2008.

[7] K. Waheed, K. Weaver, and F. Salam, "A robust algorithm for detecting speech segments using an entropic contrast," in *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, vol. 3. IEEE, 2002, pp. III–328.

[8] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *CVPR (1)*, 2005, pp. 597–604.

[9] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM, 2008, pp. 50–56.

[10] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *Proceedings of the 2nd International Workshop on Augmented Reality (IWAR 99)*, San Francisco, USA, Oct. 1999.

[11] P. Zukow-Goldring, "A social ecological realist approach to the emergence of the lexicon: Educating attention to the amodal invariants in gesture and speech." *Evolving explanations of development: Ecological approaches to organism-environment systems*, pp. 199–252, 1997.

[12] P. Zukow-Goldring and K. R. Ferko, "An ecological approach to the emergence of the lexicon: Socializing attention," *Sociocultural approaches to language and literacy: An interactionist perspective*, pp. 170–190, 1994.

[13] C. Rankin, T. Abrams, R. Barry, S. Bhatnagar, D. Clayton, J. Colombo, G. Coppola, M. Geyer, D. Glanzman, S. Marsland *et al.*, "Habituation revisited: an updated and revised description of the behavioral characteristics of habituation," *Neurobiology of learning and memory*, vol. 92, no. 2, p. 135, 2009.

[14] K. Lohan, K. Rohlfing, K. Pitsch, J. Saunders, H. Lehmann, C. Nehaniv, K. Fischer, and B. Wrede, "Tutor spotter: Proposing a feature set and evaluating system," *International Journal of Social Robotics*, 2012.

[15] J. Saunders, H. Lehmann, F. Foerster, and C. Nehaniv, "Robot acquisition of lexical meaning - moving towards the two-word stage," 2012.

[16] "Openni - the openni organization is an industry-led, not-for-profit organization formed to certify and promote the compatibility and interoperability of natural interaction (ni) devices, applications and middleware." [Online]. Available: http://openni.org

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[18] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Machine Learning-International Workshop then Conference*. Morgan Kaufmann Publishers, Inc., 1996, pp. 148–156.

[19] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, vol. 37, no. 3, pp. 297–336, 1999.

[20] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from poisson," in *Proceedings of the third workshop on very large corpora*, 1995, pp. 121–130.