2007 IEEE/RSJ International Conference on Intelligent Robots and Systems,
St Louis, USA, October 2009

Multimodal Word Learning from Infant Directed Speech

Jonas Hörnstein, Lisa Gustavsson, Francisco Lacerda, and José Santos-Victor

***Abstract*— When adults talk to infants they do that in a different way compared to how they communicate with other adults. This kind of Infant Directed Speech (IDS) typically highlights target words using focal stress and utterance final position. Also, speech directed to infants often refers to objects, people and events in the world surrounding the infant. Because of this, the sound sequences the infant hears are very likely to co-occur with actual objects or events in the infant's visual field. In this work we present a model that is able to learn word-like structures from multimodal information sources without any pre-programmed linguistic knowlege, by taking advantage of the characteristics of IDS. The model is implemented on a humanoid robot platform and is able to extract word-like patterns and associating these to objects in the visual surrounding.**

I. INTRODUCTION

Speech plays an important role in human interaction, and can also provide an intuitive way to interact with humanoid robots. When an infant is born, it is faced with the challange of extracting useful patterns from a continuous speech signal and to assign meaning to these. This is a complex task and while it is not fully understood how it is solved by the infant, part of the answer may actually be found in the structure of the speech signal.

Speech directed to infants is highly structured and characterized by what seems like physically motivated tricks to maintain the communicative connection to the infant, actions that at the same time also may enhance linguistically relevant important aspects of the signal. Also, whereas communication between adults usually is about exchanging information, speech directed to infants is of a more referential nature. The adult refers to objects, people and events in the world surrounding the infant [1]. Because of this, the sound sequences the infant hears are very likely to co-occur with actual objects or events in the infant's visual field. The placement of target words, the repetitive structure, and the use of focal stress are likely to play an important role in helping the infant establishing an implicit and plausible word-object link. This kind of structuring might very well be one of the first steps in speech processing, a coarse segmenting of the continuous signal in chunks that stand out because of some recurrent pattern the infant learns to recognize. Infants are very sensitive to the characteristic qualities of this typical IDS style, and a number of studies indicate that infants use this kind of information to find implicit structure in acoustic signals [2] [3] [4] [5] [6].

In this work we design a prototypical system that is able to learn word-like structure from multimodal information sources by integrating acoustic and visual representations. This is not a completely new approach. In the CELL [7], Cross-channel Early Lexical Learning, architectures for processing multisensory data is developed and implemented in a robot called Toco the Toucan. The robot is able to acquire words from untranscribed acoustic and video input and represent them in terms of associations between acoustic and visual sensory experience. Compared to conventional ASR systems that maps speech signal to human specified labels, this is an important step towards creating more ecological models. However, significant shortcuts are still taken, such as the use of a predefined phoneme-model where a set of 40 phonemes are used and the transition probabilities are trained off-line on large scale database. In [8], no external database is used. Instead the the transition probabilities are trained online only taking into account utterances that have been presented to the system at the specific instance in time. While this make the model more plausible from a cognitive perspective, infants may not rely on linguistic concepts as phonemes at all during these early stages of language development.

In contrast to these related works we avoid using any pre-programmed linguistic knowledge. Instead we take a more direct approach and map the auditory impression of the word as a whole to the object. This is in line with the Ecological Theory of Language Acquisition (ETLA) [9], where underlying concepts like phonemes instead are emergent consequences imposed by increasing representation needs [10] [11]. To compensate for the lack of pre-programmed structures we instead take advantage of some of the characteristics of IDS in order to boost the word learning. Taking this approach allows us to investigate how much linguistic structure, available implicitly in the speech signal directed to infants that can be derived without pre-programmed linguistic knowledge. The objective is both to give insights to how infants may be able to derive linguistic structures directly from the speech signal and examplify how a humanoid robot can take advantage of this approach in order to learn word-objects releations.

The rest of the paper is organized as follows. In section II we give further background information on IDS and show results from an infant study on multimodal word learning. In section III we present the computer model used for the experiments in section IV. Conclusions and directions for future work is found in section V.

This work was partially supported by EU Project CONTACT and by the Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS Conhecimento Program that includes FEDER funds.

J. Hörnstein and J. Santos-Victor are with Institute for System and Robotics (ISR), Instituto Superior Técnico, Lisbon, Portugal {jhornstein,jasv}@isr.ist.utl.pt

L. Gustavsson and F. Lacerda are with Department of Linguistics, Stockholm University, Stockholm, Sweden {lisag,frasse}@ling.su.se

II. Background and motivation

In this section we first review important aspects of infant directed speech, and then present a study on how these may influence infants' ability to create word-object associations.

A. Infant directed speech

An important portion of the physical signal in the ambient language of almost every infant is in the form of Infant Directed Speech (IDS), a typical speech style used by adults when communicating with infants. IDS is found in most languages [12] [13] [14] and is characterized by long pauses, repetitions, high fundamental frequency, exaggerated fundamental frequency contours [3] and hyperarticulated vowels [13]. A very similar speech style is found in speech directed to pets [15] [16], and to some degree also in speech directed to humanoid robots [17], and pet robots [18].

The function of IDS seem to change in accordance with the infant's developmental stages, phonetic characteristics in the adult's speech are adjusted to accommodate the communicative functions between the parents and their infants, for example a gradual change in consonant specifications associated with the infants communicative development was found in a study by Sundberg and Lacerda [19]. In longitudinal studies it has been shown that parents is adapting their speech to their infants linguistic and social development the first post-natal year. On the whole they use higher fundamental frequency, greater frequency range, shorter utterance duration, longer syllable duration, and less number of syllables per utterance when speaking to their infants as compared to speaking to adults. Sundberg [20] suggests that these phonetic modifications might be an intuitive strategy adults use automatically that is both attractive and functional for the infant.

In a study more directly related to infants word learning, Fernald and Mazzie [21] found that target words in infant directed speech were typically highlighted using focal stress and utterance-final position. In their study 18 mothers of 14-month-old infants were asked to tell a story from a picture book called Kelly's New Clothes, both to their infants and to an adult listener. Each page of the book introduced a new piece of clothes that was designated as a target word. When telling the story to the infants target words were stressed in 76% of the instances, and placed in utterance-final position in 75% of the instances. For adult speech the same values were 40% and 53% respectively.

Albin [22] found that an even larger portion of the target words (87% - 100% depending of subject) occured in final position when the subjects were asked to present a number of items to an infant.

B. Infant word learning study

The review above shows that IDS is structured in a way that can potentially facilitate the word learning task for infants. To investigate if infants are able to take advantage of utterance-final position and focal stress of target words when learning audio-visual links, a study involving a total of 50 infants has been performed. The study has previously been presented in [23] and is summarized below.

This study used a Visual Preference Procedure that is essentially a modified version of Fernald's Preferential Listening Procedure [24]. In general terms the procedure can be described as inducing the infant's response from its looking time towards alternative pictures displayed simultaneously and where one of the pictures is associated with the expected response.

The speech material used were Swedish sentences recorded by a female speaker. The sentences were recorded in nine different conditions where main stress and target-word were placed in all possible combinations of sentence initial, medial and final positions. The sentences introduced non-words as possible names of objects (e.g. "It is a nice Dappa" where Dappa is the name of one of the puppets). The speech materials were produced in IDS-style, which is characterized by modifications as described earlier, such as frequent prosodic repetitions and expanded intonation contours.

Nine films were created to include all the possible combinations of position of the target word (initial, medial or final position in the utterance) and the utterances focal accent (falling on the utterances initial, medial or final words). The syntactic structure of the utterances was different from film to film but within each film the position of the target word and the part of the utterance receiving focal accent was kept constant. Furthermore, although the utterances within each film were structurally equal, the non-target words were different from utterance to utterance in an attempt to mimic the variation typically observed in natural utterances.

The results are shown in Figure 1, grouped according to the placement of the focal accent and the position of the target word in the utterances. Given the reduced number of subjects each condition and the typical variance observed in this type of experiments, it is perhaps not surprising that no significant main effects for the target word or placement of the focal accent could be observed. Nevertheless in the combinations of target word in final position and focal accent in initial or final positions the gain was all positive.

There were no significant main effects or interactions for word position and placement of the focal accent. However a tendency for longer looking times was observed for the target word in focal position [$F[1,73]=2.957$, $p<0.090$]. If the case of the target word in final position, with a focal accent in the initial position of the sentence is excluded, then a significant effect of the placement of the target word in focus is obtained [$F[1,65]=4.075$, $p<0.048$]. Furthermore, there was a significant difference between the mean looking times for the group of sentences with the target word in focal position plus the sentences with the target word in final position and focal accent in initial position [$F[1,73]=5.579$, $p<0.021$].

III. Word learning model

In this section we describe the computer model that is used to extract words from Infant Directed Speech and associate

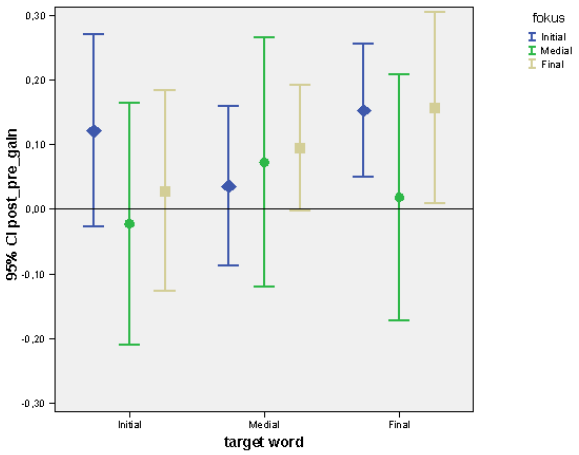

Fig. 1. Average gains in seconds for the infants' responses as a function of the placement of the target word and focal accents. The target word positions are displayed on the x-axis; target word in initial position to the left, target word in medial position in the middle and target word in final position to the right. Blue diamonds represent focal accent on the initial word, green circles focal accent on medial words and yellow squares focal accent on final the word.

those with the objects they refer to. The model is inspired by the CELL-model [7], and consists of acoustic and visual sensors, a short-term memory with recurrence filter, and a long-term memory with a mutual information filter. Raw data from the visual and auditory sensors are stored in the short-term memory where they are processed independently. The auditory stream is searched for recurring patterns which forms potential word candidates. Those word candidates are paired with objects found in the visual input and sent to the long-term memory. Finally the long-term memory is searched for cross-modal regularities in order to form word-object associations.

The main difference between our model and the CELL-model is how data is represented and processed, and that we make further use of the structure of IDS, as will be explained more in detail below. We first describe how we extract word candidates and visual objects form the auditory and visual streams, and then describe in detail how the word-object associations are created.

A. Extracting word candidates

One of the objectives with this work is to investigate how much linguistic structure, available implicitly in the speech signal directed to infants that can be derived without pre-programmed linguistic knowledge. We have therefore deliberately choosen a rather crude computational method for finding recurrent patterns in the speech signal and avoid using any kind of phoneme models.

For simplicity the speech signal is currently recorded one utterance at a time. Our model of the short term memory can store either a fixed number of utterances (typically around 5) or a fixed amount of time (here we have used 10-20 s).

The speech signal of each utterance is windowed using 25 ms windows with 50% overlap between the windows. For each window we then calculate Mel coefficients [25]. Only the first four of the coefficients are used as a representation of the signal when looking for recurrent patterns.

Each utterance within the short term memory at a given time is compared pair-wise with all other utterences in the memory. For each utterance-pair we first make sure that the utterances have the same length by padding the shortest utterance. The utterances are then aligned in time and we calculate the sum of differences between their mel coefficients creating a vector with the acoustic distance between the two utterances at each window. The second utterance is then shifted forward and backward in time and for each step a new distance vector is calculated. These vectors are averaged over 15 windows, i.e. 200 ms, and combined into a distance matrix as illustrated in Figure 2. By averaging over 200 ms we exclude local matches that are too short and can find word candidates by simply looking for minimas in the distance matrix. Starting from a minima we find the start and the end points for the word candidate by moving left and right in the matrix while making sure that the distance metric at each point is always below a certain critical threshold.

In order to take advantage of the structure of infant directed speech and to mimic infants apparent bias towards target words in utterance-final position and with focal stress, we also check for these features. For a word candidate to be considered to have utterance-final position we simply check that the end of the candidate is less than 15 windows from the end of the utterance. To find the focal stress of an utterance we look for the F0-peak. While there are many ways for adults to stress words (e.g. pitch, intensity, length) it has been found that F0-peaks are mainly used in infant directed speech [21]. If the F0-peak of the utterance as a whole is within the boundaries of the word candidate, the word candidate is considered to be stressed. If a word candidates is not stressed and in utterance-final position we reject it with a specified probability.

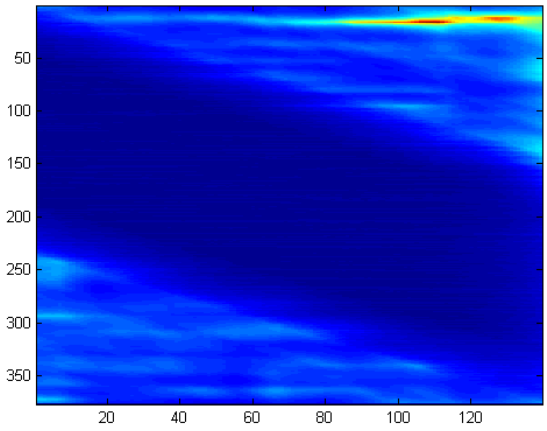

Fig. 2. Finding word candidates in sentences "titta här är den söta dappan" and "se på lilla dappan". The best match is found by shifting the sentences 15 windows.

B. Finding visual objects

When a word candidate is found it needs to be paired with a visual object before it is sent to the long term memory. In this section we describe how the visual objects are extracted and represented.

Starting from a snapshot of the robot's eye view we segment the image and look for the object closest to the center of the image. The segmentation is done by background subtraction followed by morphological dilation.

Using the silhouette of the object we create a representation of its shape by taking the distance between the center of mass and the perimeter of the silhouette. This is done for each degree of rotation creating a vector with 360 columns. The tranformation of an image to the object representation is illustrated in Figure 3.

When comparing two object representations with each other we first normalize the vectors and then perform a pattern maching, much in the same way as for the auditory representations, by shifting the vectors one step at a time. By doing this we get a measurement of the visual similarity between objects that is invariant to both scale and rotation.

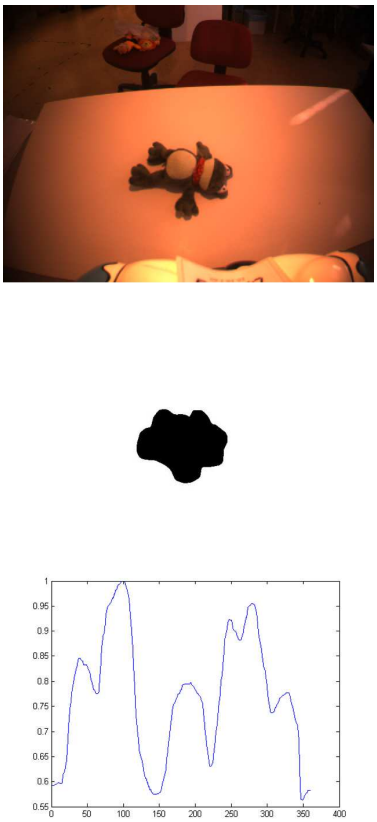

Fig. 3. Original image (top), silhouette image after background substraction and morphologic operations (center), and the silhouette perimeter in polar coordinates (bottom).

C. Creating word-object associations

Each time a recurrent acoustic pattern is found in the short-term memory along with the presence of a visual object, the word candidate and the object representation are given a unique identifier and are sent to the long-term memory. In the long-term memory we first look for acoustic and visual similarity independently. For the visual similarity between objects we directly use the measurement explained above. For the acoustic similarity we use Dynamic Time Warp (DTW) [26] to measure the distance between different word candidates. The reason to use DTW instead of directly applying the pattern matching described earlier is to be less sensitive to how fast the word candidate is pronounced. The distance measurements are used to group similar word candidates and visual objects respectively, using an hierarchical clustering algorithm [27].

The hierarchical clustering results in two tree structures. The first tree structure contains all the word candidates. These are found in the bottom at the leaves. Each branch in the tree represent a cluster containing all word candidates below. As we move upwards in the tree we continue to join more word candidates by allowing bigger acoustic distances between the candidates. In the top we find a single cluster containing all word candidates. In the same way the leaves of the second tree contain all visual objects and by moving upwards in the tree we start to group more and more visual objects by allowing larger distances.

The difficult part here is to decide at what branch we should cut the trees in order to get a good representations of the words and the objects. If we cut too low in the word candidate tree we will not allow sufficient acoustic difference in order to cope with natural variations in the pronounciation of the word. On the other hand, if we cut too high we may end up classifying any sound sequence as a correct observation of the specific word. The same problem holds for the object tree. This is where we make use of our multimodal information.

In order to find which branch in the word candidate tree that should be associated with which branch in the object tree we use the mutual information criterion [28]. In the general form this can be written as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p_1(x)p_2(y)} \right)$$

Where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively.

We want to calculate $I(X;Y)$ for all combinations of clusters and objects in order to find the best word representations. For a specific word cluster A and visual cluster V we define the binary variables X and Y as

$$X = \{1 \; if \; observation \in A; \; 0 \; otherwise\}$$

$$Y = \{1 \; if \; observation \in V; \; 0 \; otherwise\}$$

The probability functions are estimated using the relative frequencies of all observations in the long-term memory, i.e.

$p_1(x)$ is estimated by taking the number of observations within the cluster A and dividing with the total number of observations in the long-term memory. In the same way $p_2(y)$ is estimated by taking the number of observations in the cluster V and again dividing with the total number of observations. The joint probability is found by counting how many of the observations in cluster A that are paired with an observation in cluster V and dividing by the total number of observations.

IV. Experimental results

The word acquisition model has been tested both on recordings of real IDS and by implementing the model in a humanoid robot to allow direct interaction. The objective is to test how much linguistic structure that we can derive from our model, and also to test if focusing on words that are stressed and in utterance-final position may improve our word-object associations. The main results presented are obtained when rejecting all candidates that are not stressed and in utterance-final position, and in the end of each experiment we shortly discuss how the result was altered when taking all word candidates into account.

A. Experiment 1: Humanoid platform

The word acquisition model has been implemented in our humanoid robot Chica, Figure 4. In this experiment we show a number of toys for the robot and, at the same time, talk about these objects using infant directed speech style. The objects that were chosen were one ball and two dolls named "Pudde" and "Siffy". The experiment was executed by demonstrating one object at a time by placing it in front of Chica and talk about the object for approximately 20 s. Each utterance contained a reference to a target word and we made sure that the target word was always stressed and in utterance-final position. For the dolls we referred to them both by using their individual names and the swedish word for doll, "docka". The ball were always referred to using the swedish word "bollen".

Fig. 4. Experimental setup for robot test

During the length of one demonstration, sound and images are continuously stored in the short-term memory. The sound is then segmented into utterances by looking for longer periods of silence. Each utterance is then compared to the others as explained in section III. Word candidates are paired with a visual representation of the object and sent to the long-term memory.

After having demonstrated all three objects we repeat the procedure once more, but this time with the objects in a different orientation in front of the robot. This is done in order to verify that the clustering of the visual objects is able to find similarities in the shape despite differences in the orientation of the objects.

When word candidates have been extracted from all six demonstrations, the hierarchical clustering algorithm is used to group word candidates in the long-term memory that are acoustically close. The result from the hierarchical clustering of the word candidates and the visual objects can be seen in Figure 5. The numbers at each leaf shows the unique identifier that allows us to see which of the word candidates that was paired with which of the visual objects.

Looking only at the hierarchical tree for the word candidates it is not obvious where the tree should be cut in order to find good word representations. By listening to the word candidates we notice that the cluster containing candidates (25 26 19 20 2 6 18 14 16 1) represent the word "dockan", the cluster (3 7 4 9 5 8 10 12 15 11 13 17) represent the word "Pudde", the cluster (21 22 23 27 28 29 24 31 30) represent the word "Siffy", and the cluster (32 33 34 36 35) represent the word "bollen". The hierarchical tree for the visual objects may look more simple and it is tempting to select the five clusters in the bottom as our objects. However, it is actually the clusters one level up that represents our visual objects. Of course the robot does not know that at this point.

To find out which branch in the respective tree that should be associated with which branch in the other we calculate the mutual information criterion. Calculating the mutual information criterion for all pair of branches shows that we get the highest score for associating the word candidates (32-35) with the same visual objects (32-35). This is what we could expect since all visual observations of "bollen" were also paired with a correct word candidate. In the case of the objects "Pudde" and "Siffy" part of the observations are not paired with the object name, but instead with the word "docka". Still we get the second and third heighest scores by associating word candidates for the word "Pudde" with object Pudde and the word "Siffy" with object Siffy respectively. We can also find that the branch above the visual representations of Pudde and Siffy receives the heighest score for being associated branch containing word candidates for "dockan".

The experiment was repeated without putting any bias on word candidates that were stressed and in utterance-final position. This resulted in four false word candidates for the object Pudde and one for object Siffy. However, this did not affect the word-object associations as these candidates were found in separate branches in the word candidate tree and only received low scores by the mutual information criterion.

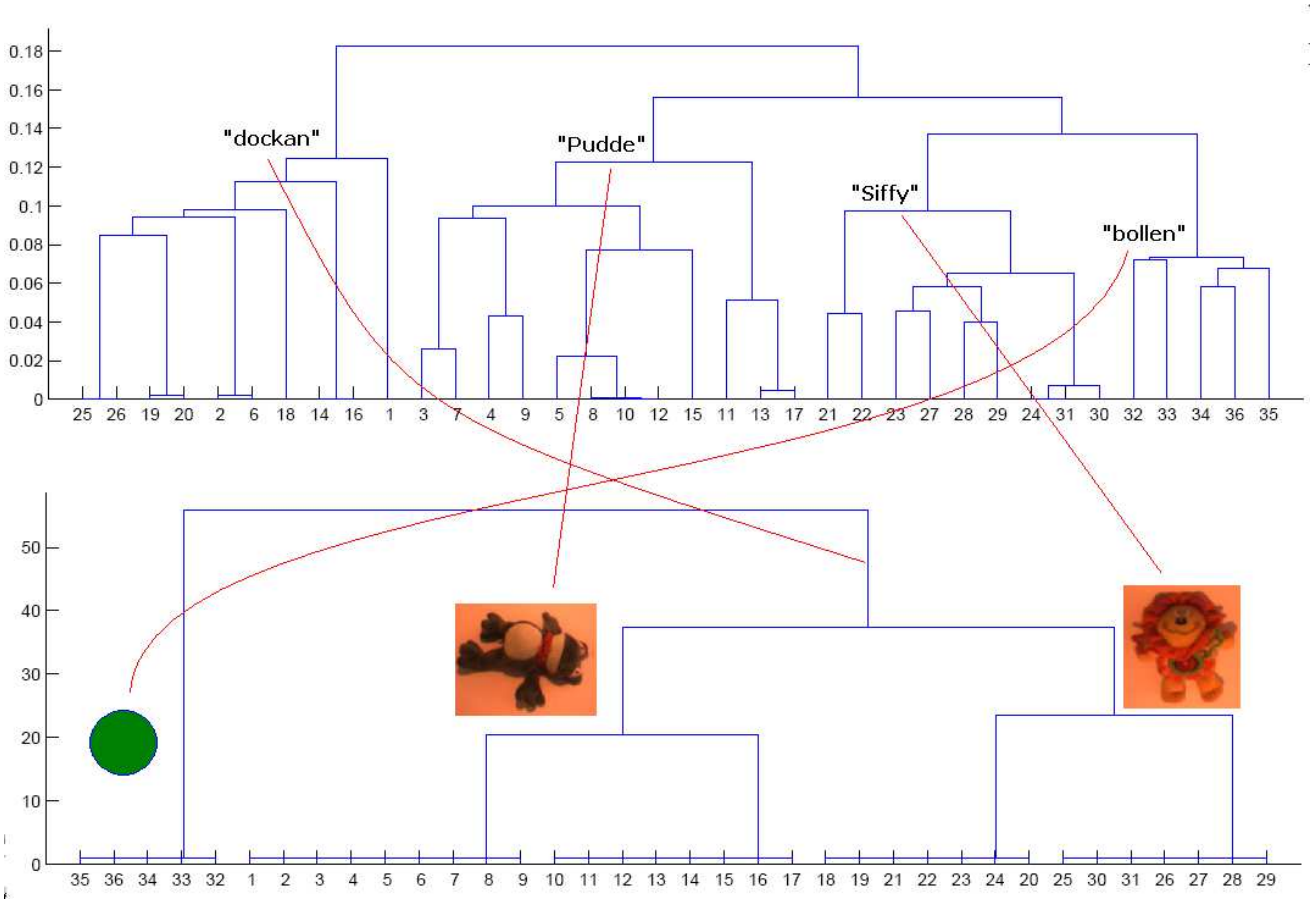

Fig. 5. Above: Clusters of the extracted word candidates during the robot experiment. Word candidates 1-17 are paired with object Pudde, nr 18-29 with object Siffy, and 32-36 with object bollen. Below: Clusters of the extracted visual objects during the robot experiment. Objects 1-17 corresponds to object Pudde, nr 18-29 to object Siffy, and 32-36 to object bollen.

B. Experiment 2: Infant Directed Speech recordings

A second experiment was performed using recordings of interactions between parents and their infants. The recordings were made under controlled forms at the Department of Linguistics, Stockholm University. A lot of care was taken to create natural interactions. The room was equipped with several toys, among those two dolls called "Kuckan" and "Siffy". The parents were not given any information of the aim of the recordings but were simply introduced to the toys and then left alone with their infants. In this study we have only used a single recording of a mother interacting with her 8 month old infant. The total duration of the recording is around 10 minutes. The audio recording has been segmented by hand to exclude sound coming from the infant. In total the material consists of 132 utterances with time stamps and also object references in those case that an object were present. In 33 of these the doll "Kuckan" was present and in 13 of them the doll "Siffy". In total the word "Kuckan" is mentioned 15 times and "Siffy" is mentioned 6 times.

In this experiment we limit the short-term memory to 10 s. The utterances enter in the short-term memory one at a time and any utterance older than 10 s is erased from the memory. Word candidates that also have an assigned object label are transferred into the long-term memory.

After searching all utterances for word candidates we cluster all the candidates in the long-term memory. The result can be found in Figure 6. Here we don't have any hierarchical tree for the visual objects. Instead we use the labels assigned by hand that can be used for calculating the mutual information criterion. Doing so gives us that the object Kuckan is best represented by word candidates (5 6 3 9 11 10 22 4 7 17 18) and Siffy by (32 33 34). Listening to the word candidates confirms that they represent the names of the dolls, but the segmentation is not as clear as in the robot experiment and there are a few outliers. Among the word candidates associated with Kuckan, nr 22 was unhearable and nr 17 and 18 were non-words but with a prosodic resembly of the word "Kuckan". For the word candidates associated with Siffy all contained parts of preseeding words.

When repeating the experiment without bias on focal stress and utterance-final position, the number of word candidates grew significantly resulting in lots of outliers being associated with both the objects. In the case of Kuckan it even caused the correct word candidates to be excluded from the branch that was associated with the object.

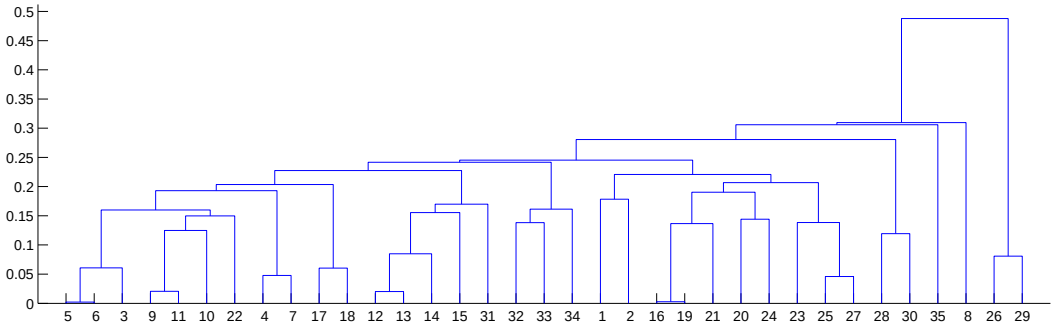

Fig. 6. Cluster formations from word candidates taken from infant directed speech. Word candidates between 1 and 30 are paired with object Kuckan and word candidates between 31 and 35 are paired with Siffy. Using the mutual information criterion, cluster (32 33 34) gets associated with Siffy and cluster (5 6 3 9 11 10 22 4 7 17 18) gets associated with Kucka.

V. Conclusions and Future Work

The objectives with this work has been to investigate how much linguistic structure that can be derived from Infant Directed Speech, without any pre-programmed linguistic knowledge. We have shown that it is possible to extract useful word candidates and to associate those with objects in the visual field, using simple pattern matching techniques. The method make use of the repetitive nature of IDS as well as other phonetic characteristics of IDS such as final placement of target words and the use of focal stress to guide the attention. The initial results are encouraging and future work include using the method in larger scale experiments and extending the method to associate not only objects but also actions and properties.

References

[1] Lacerda, F., Marklund, E., Lagerkvist, L., Gustavsson, L., Klintfors, E., Sundberg, U., "On the linguistic implications of context-bound adult-infant interactions", In Genova: Epirob 2004, 2004

[2] Jusczyk, P., Kemler Nelson, D. G., Hirsh-Pasek, K., Kennedy, L., Woodward, A., Piwoz, J., "Perception of acoustic correlates of major phrasal units by young infants", Cognitive Psychology, 24, pp 252-293, 1992

[3] Fernald, Al, "The perceptual and affective salience of mothers' speech to infants", In The origins and growth of communication, Norwood, N.J, Ablex., 1984

[4] Kuhl, P. and Miller, J., "Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants", Perception and Psychophysics, 31, 279-292, 1982

[5] Crystal, D. "Non-segmental phonology in language acquisition: A review of the issues", Lingua, 32, 1-45, 1973

[6] Saffran, J. R., Johnson, E. K., Aslin, R. N., Newport, E., "Statistical learning of tone sequences by human infants and adults", Cognition, 70, 27-52, 1999

[7] Roy, D. and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 2002, vol 26, pp 113-146

[8] ten Bosch, L., Van hamme, H., Boves, L., "A computational model of language acquisition: focus on word discovery", In Interspeech 2008, Brisbane, 2008

[9] Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., Sundberg, U., "Ecological Theory of Language Acquisition", In Genova: Epirob 2004, 2004

[10] Nowak, M. A., Plotkin, J. B., Jansen, V. A. A., "The evolution of syntactic communication", Nature, 404, pp 495-498, 2000

[11] Lacerda, F., "Phonology: An emergent consequence of memory constraints and sonsory input", Reading and Writing: An Interdisciplinary Journal, 16, pp 41-59, 2003

[12] Andruski, J. E., Kuhl, O. K., Hayashi, A., "Point vowels in Japanese mothers' speech to infants and adults", The Journal of the Acoustical Society of America, 105, pp 1095-1096, 1999

[13] Kuhl, P., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevnikova, E. V., Ryskina, V. L. et al., "Cross-language analysis of Phonetic units in language addressed to infants", Science, 277, pp. 684-686, 1997

[14] Ferguson, C. A., "Baby talk in six languages", American Anthropologist, 66, pp 103-114, 1964

[15] Hirsh-Pasek, K., "Doggerel: motherese in a new context", Journal of Child Language, 9, pp. 229-237, 1982

[16] Burnham, D., "'What's new pussycat? On talking to babies and animnals", Science, 296, p 1435, 2002

[17] Fitzpatrick, P., Varchavskaia, P., Breazeal, C., "Characterizing and processing robotdirected speech", In Proceedings of the International IEEE/RSJ Conference on Humanoid Robotics, 2001

[18] Batliner, A., Biersack, S., Steidl, S., "'The Prosody of Pet Robot Directed Speech: Evidence from Children", Proc. of Speech Prosody 2006, Dresden, pp 1-4, 2006

[19] Sundberg, U, and Lacerda, F., "Voice onset time in speech to infants and adults", Phonetica, 56, pp 186-199, 1999

[20] Sundberg, U., "Mother tongue Ű Phonetic aspects of infant-directed speech", Department of Linguistics, Stockholm University, 1998

[21] Fernald, A., and Mazzie, C., "Prosody and focus in speech to infants and adults", Developmental Psychology, 27, pp. 209-221, 1991

[22] Albin, D. D., and Echols, C. H., "Stressed and word-final syllables in infant-directed speech", Infant Behavior and Development, 19, pp 401-418, 1996

[23] Gustavsson, L., Sundberg, U., Klintfors, E., Marklund, E., Lagerkvist, L., Lacerda, F., "Integration of audio-visual information in 8-months-old infants", in Proceedings of the Fourth Internation Workshop on Epigenetic Robotics Lund University Cognitive Studies, 117, pp 143-144, 2004

[24] Fernald, A., "Four-month-old infants prefer to listen to Motherese", Infant Behavior and Development, 8, pp 181-195, 1985

[25] Davis, S. B., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, speech, and signal processing, Vol. ASSP-28, no. 4, August 1980

[26] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43- 49, 1978, ISSN: 0096-3518

[27] Hastie, T., "The elements of statistical learning data mining inference and prediction", Springer, 2001

[28] Cover, T. M., Thomas, J. A., "Elements of information theory", Wiley, July 2006