Towards Using Prosody to Scaffold Lexical Meaning in Robots

Joe Saunders, Hagen Lehman, Yo Sato and Chrystopher L. Nehaniv

Adaptive Systems Research Group

School of Computer Science

University of Hertfordshire

Hatfield, UK, AL10 9AB

Email: j.1.saunders@herts.ac.uk

Abstract—We present a case-study analysing the prosodic contours and salient word markers of a small corpus of robot-directed speech where the human participants have been asked to talk to the robot as if it were a child. We assess whether such contours and salience characteristics could be used to extract relevant information for the subsequent learning and scaffolding of meaning in robots. The study uses measures of pitch, energy and word duration from the participants speech and exploits Pierrehumbert and Hirschberg's theory of the meaning of intonational contours which may provide information on shared belief between speaker and listener. The results indicate that 1) participants use a high number of contours which provide new information markers to the robot, 2) that prosodic question contours reduce as the interactions proceed and 3) that pitch, energy and duration features can provide strong markers for relevant words and 4) there was little evidence that participants altered their prosodic contours in recognition of shared belief. A description and verification of our software which allows the semi-automatic marking of prosodic phrases is also described.

Fig. 1. Part of a session in the study with the humanoid robot Kaspar [2]. The participant was asked over five 2-minute sessions with the robot to explain the shapes on the box as if the robot were a small child. From session 2 onwards the robot responded to the presented shapes by querying its association between salient words spoken by the participant in the previous session with its sensorimotor history and expressed a chosen word.

I. INTRODUCTION

In this paper we present a case-study where we analyse the prosodic features and salience markers of a small corpus of robot-directed speech to ascertain whether sufficient linguisitic cues are available in it to allow effective scaffolding of language elements and specifically lexical meaning. The work builds on a case study carried out by Saunders et al. [1] where eight human participants were each asked to teach a humanoid robot ('Kaspar' [2]) about shapes during a series of five two-minute interaction sessions (see Fig. 1).

In these experiments each participant was asked to treat the robot as if it were a small child however no other instructions were placed on their use of language. The robot used features derived from research into observations of caregiver speech interactions with human infants [3], [4] to extract salient words from the interaction with the robot and associate these with its sensorimotor feedback. This human-child interaction style is called Child Directed Speech (CDS). What was important was that the human adopted a language style similar to CDS with the robot and that some form of rudimentary shared reference was employed. This being the case the robot learnt the semantics of the shape words, based on its own sensorimotor feedback, with relatively few presentations.

II. ISSUES AND LIMITATIONS OF THE STUDY

The CDS features employed for the heuristics used in the case study were based on the observation [3] that during the early stages in the process of joint understanding salient words in CDS are lengthened with an emphasis on utterance final position. In languages like English this usually produces a grammatical utterance with a salient noun at the end, but it is also observed in languages like Turkish when an ungrammatical (or uncommon) utterance is produced [5]. The study captured the sensorimotor perceptions of the robot during the interactions which included the cartesian position in space that the robot was looking at, robot joint angles, face detection and an integer value representing a unique shape identifier. As the sessions progressed the robot was able to succesfully weight the shape identifier as being more salient than the other sensorimotor attributes and furthermore associate it with the selected words. Thus when subsequently presented with say a 'star' shape the robot would consistently say 'star' regardless of where in space the shape occurred and regardless of the robot's own proprioception. It thus derived lexical meaning based on its own generalised sensorimotor attributes associated to the word form.

A limitation of these experiments was that although the robot could perceive shape cartesian location and its own proprioception it could not perceive attributes such as shape colour or shape size. This is problematical for two reasons, firstly, if words relating to size and colour were used by the participant the robot would have no way to associate them with its own perceptions, and secondly that the salient word heuristic being employed was less likely to mark such words as salient as they normally occur (in English) before the relevant noun and less often at end of utterance. A further issue which was highlighted by the case study was that of the human's perception of shared belief. This was illustrated by the human changing the emphasis and type of utterance used by the fifth session. By this time the robot had effectively associated relevant words with its sensorimotor input and would 'say' the words when presented with the appropriate shape, however the human would then typically praise the robot or start to talk about other attributes of the shape e.g. its size. These changes of emphasis could not be detected by the robot and led to a gradual disassociation of shape noun with its current perceptions and a replacement with words such as 'done' or 'boy' (from 'well done' and 'good boy').

In order to cope with these limitations we considered an alternative form of word extraction which could firstly highlight salient words expressed by the human regardless of where they were situated in the utterance and secondly allow the robot to attempt to understand the intentional and attentional features of the human utterances. Both of these aspects were carried out via an analysis of the prosodic features within the utterances expressed by the human. The former step is motivated by the large amount of work carried out on child language acquisition [3], [4] and the latter step motivated from work carried out by Kim et al. [6] applying Pierrehumbert and Hirschberg's thesis on intonational meaning [7] in their study of automatic pitch feature extraction to examine meaning applied to infant directed speech.

III. Prosody, Salience and Language Learning in Children

In order for children to acquire language effectively it is essential that social interaction with others occur [8]. Typically interaction partners include adults, caregivers and other children. Although adults do not specifically attempt to teach language to children there is a need for joint understanding and shared belief with the child and this joint understanding may come about through a learning process specifically biased to draw attention to relevant parts of an utterance or to aspects of the learnable task. This attentional behaviour can take many forms including using a linguistic style of communication such as CDS with certain well documented characteristics [3], [9], as well as using prosody to signal attention and intention [7], shaping and modifying of demonstrations in a learning context [10], the use of shared reference so that a young baby will follow the gaze of its mother [11], and later follow her pointing gestures [12]. There is also evidence for synchrony between object naming and movement [13], [14], [15]. In

this report however we consider only the speech aspects of the caregiver's behaviour.

A. Speech Features of CDS

Typically caregivers of infants during interaction will produce short utterances. For instance, in six child-directed speech corpora in the CHILDES database [16] for children aged 2;6 or younger, the number of words in an utterance averages from 3.37 to 4.01 [17] with many utterances consisting of a single word. Repetition is common, with the use of a number of high frequency words and phrases. Caregivers often repeatedly draw on a repertoire of phrases. As a special case of repetition there is often frequent use of the infant's name. CDS is typically slower than adult directed speech (ADS) and most words are mono- or disyllabic. Salient words are lengthened and prosody is used to give greater emphasis to salient words. Salient words are often placed at the end of the utterance for young infants and occasionally the infant's productions are corrected. Initially, in English CDS, there is typically a preponderance of nouns rather than verbs, the task of learning verbs is usually seen as harder: suggested reasons include the problems of individuating and categorizing actions, their transitory nature, and ambiguity of reference [18, page 355]. The use of CDS declines gradually as the child develops: caregivers evidently have rising expectations about the child's linguistic ability.

B. Prosodic Features of CDS

In natural speech there are noticable variations in the way that some words and specifically the syllables within those words are emphasised. They sometimes appear louder, have a longer duration and seem to have a change in pitch. Pitch, amplitude (or energy) and duration (or rhythm), are the key features of prosody or the 'sound' or 'music' of speech. Different languages exhibit different syllable stress patterns and infants become sensitive to the qualities in their native language environments early in their development [19], [20]. It is also thought that exaggerated stress patterns in CDS allow infants to find word boundaries which they cannot find in ADS [21]. In the work presented in this paper we analyse the speech patterns in our existing corpus of robot directed speech using prosodic features (specifically pitch, energy and duration) to highlight salient words used by a human participant with a robot.

IV. Prosody and Meaning

The work of Pierrehumbert and Hirschberg [7] provides a framework for the analysis of intonational meaning based on pitch levels occuring in prosodic phrases. Pierrehumbert and Hirschberg write:

> "Intonational features such as phrasing, accent placement, pitch range, and tune represent important sources of information about the attentional and the intentional structures of discourse" [7, page 271]

Here we present a very brief description of the theory. A pitch level is usually computed by extracting the fundamental (f0)

frequency from the speech signal. In describing intonational 'tunes' Pierrehumbert and Hirschberg analysed the shape of the f0 curve in terms of sequences of high (H*) and low (L*) pitch accents. The alignment with a stressed syllables is indicated by the diacritic "*". H* and L* represent simple tones and can be identified by local f0 maxima or minima respectively in the phrase. Pierrehumbert and Hirschberg's
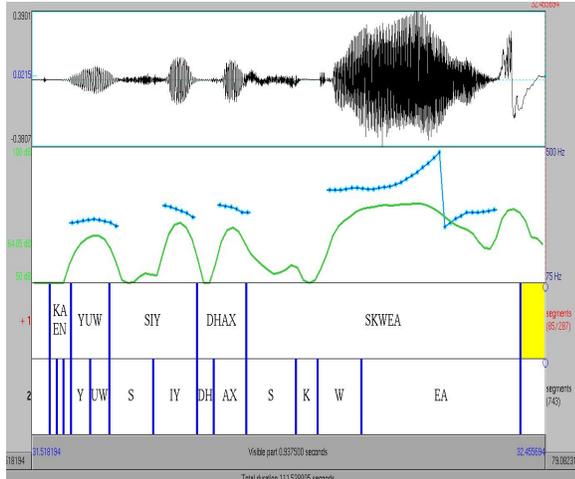

Fig. 2. An example of an utterance "can you see the square?" from one of our interaction sessions shown in Praat. The pitch contour is indicated by the dotted line. The energy is shown as a continuous line. The word and phoneme boundaries are indicated by the two bottom tiers.

theory proposes that when a H* is marked within a phrase the morpheme indicated by the H* marker is *new* in that the speaker believes that it is not in the space of mutual belief that the speaker believes is shared by the listener. An L* indicates the information is *not new*, the speaker believes it to be mutually believed. More complex tones are called two-tones which have either a suffix or prefix and in English comprise L*+H, L+H*, H*+L, and H+L* and show combinations of low and high tones realising or preceeding the stressed high or low syllable. Hobbs [22, page 314] provides a concise explanation of these tones. Thus, L+H* indicates "You might think this information is *not new*, but it really *is new*" whereas H+L* says "You might think this information is *new*, but it really is *not new*". L*+H and H*+L are concerned with incompleteness. The H suffix signalling incompleteness whereas the L suffix fails to signal incompleteness. Following Hobbs [22, page 315] incompleteness means "What I've just conveyed by that morpheme or phrase requires further discussion before it is entered into mutual belief". Further meaning is derived from phrase accents (L or H) and boundary tones (indicated by L% or H%) at the end of a phrase. There are four possible combinations: L-L%, H-L%, L-H% and H-H%. A familiar question contour would be represented with a high rising H-H% boundary tone. An example from the corpus is shown in Fig. 2 and would be transcribed as follows:

```
Can you see the square
L*                  H* H-H%
```

TABLE I
SHARED BELIEF - A SIMPLE SUMMARY OF PIERREHUMBERT AND HIRSCHBERG'S [7] FRAMEWORK ON THE MEANING OF INTONATIONAL CONTOURS

Pitch Accent	Boundary Tone	Meaning
H*	L-L% L-H%	Speaker conveys information to listener and believes listener is unaware of accented item.
H*	H-H% H-L%	Question contours where speaker already knows answer. Speaker conveys information to listener.
L*	L-L% L-H%	Speaker believes that listener shares mutual belief and believes should be aware of accented item.
L*	H-H% H-L%	Question where speaker is in doubt about listener's mutual belief and possible response.
L+H*	various	Speaker believes listerner thinks accented item not new but it is new.
H+L*	varios	Speaker believes listener thinks accented item new but it is not new.
L*+H	various	Signals incompleteness.
H*+L	various	Fails to signal incompletness.

A table summarising the main elements of Pierrehumbert and Hirschberg's theory is shown in table I.

V. RESEARCH QUESTIONS

In carrying out this study we were interested in a number of research questions:

1) Would a human interacting with a robot change their prosody as the human's perception of shared belief increases? More specifically, will there be a change from H* prosodic markings to L* prosodic markings through the five interaction sessions. We thought that this might occur if the human's view of the shared belief space was confirmed by the robots responses. Thus once the robot correctly identifies a shape there would be no need for the human to prosodically mark it as new information.

2) Would prosodic boundary tones change from question and non-question contours during the interaction sessions? We were expecting to see the human using more question contours earlier in the interactions in an attempt to gauge whether the robot understood.

3) Could prosodic salience be used to extract relevant words from the interaction? Specifically, would the words marked as prosodically salient be those which would signal an association with a sensorimotor stream? In the corpus being analysed this would be signalled by the detection of nouns in each interaction session.

In order to analyse the first two questions we ran our prosodic analysis system on a selected set from the existing corpus of interactions with Kaspar. We specifically selected four participants (from the original eight) whose interactions had, in the study [1], provided the robot with the best

sensorimotor associations from their speech patterns and thus allowed the robot to generalise most effectively. From these participants we analysed the distribution of prosodic patterns for each person over their five sessions to ascertain changes in prosodic contours. For the third question we ran our prosodic analysis to extract salient words from the four selected participants and computed the efficiency of extraction of nouns with respect to the size of the sessions corpus, salient words selected and the total number of nouns available.

VI. PROCEDURE FOR MARKING PROSODIC FEATURES AND SALIENT WORDS

In this section we explain the semi-automatic method we employed to derive prosodic phrases, mark them with the appropriate tonal markers and identify salient words. We should note that the tonal markings described in section IV are formalised in the ToBI (Tones and Break Index) annotation system described by Beckman et al. [23] and which we subsequently used to verify our mechanism.

A. Mechanism

In our initial study with the robot [1] we made a number of assumptions. Firstly that words (and specifically the phonetic components of words) would be available to us directly from the speech signal. Current technology is not able to carry this out with sufficient accuracy at present and therefore we imposed some pre-processing steps to achieve this.

B. Pre-Processing

Following each session the participants' speech was manually transcribed and then automatically aligned against the speech signal to yield a set of timed phonemes with word markers. This was achieved using a combination of software components from SysMedia [24] which carried out the initial word alignment and the University College London SFS system [25] which converted the timed aligned words to phonemes and realigned them. Following the alignement process the timed phoneme/word file was processed by Huang, Chen and Harper's Prosodic Feature Extraction Tool (PFET) [26]. This uses Praat [27] as its underlying analytical engine and allows the extraction of various prosodic statistics at the word level[1] via underlying analysis of the phonemic input.

C. Segmenting Prosodic Phrases

We identify individual phrases based on pause duration and word duration. Average pause duration is computed as the sum of each pause between words divided by the total number of words. If the pause between words exceeds the average pause duration then an end of phrase boundary marker is inserted after that word. A further segmentation of each subsequence utterance is then carried out. This is based on studies indicating that (in English) persons using CDS favour introducing new information at ends of utterances [29], [30]. A signal for a

---

[1]Our approach also allow analysis at the syllable level by additionally using a syllabification procedure decribed by Sato et al. [28]

---

new information is extended word duration. We place a phrase boundary marker after a word whose duration is larger than the first standard deviation of all words in that utterance. This has the effect of splitting longer utterances where pause duration is ineffective.

D. Identifying Pitch Accents and Boundary Tones

Using statistics provided by the PFET system above, we extract maximum values for fundemental frequency (f0) at the word level. We then compute the average maximum f0 within each phrase and then mark where the first highest or first lowest f0 value occurs. These are marked with the appropriate H* or L* markers respectively. Two-tone pitch accents are identified by estimating whether there are a consistent set of high or low f0 values preceeding or following the high or low pitch marker. High or low pitch values are characterised as being above or below the f0 average respectively but not the extreme highest or extreme lowest value.

To estimate boundary tones we exploit the PFET's system ability to symbolically characterise the pitch contour. This is based on the f0 value at the phoneme level and results in a string containing combinations of r, f or U denoting *rising*, *falling* or *Unvoiced* respectively. We remove unvoiced symbols and compress similar subsequences of r and f. We then take the final two symbols to compute the boundary tone. Thus the substring *rr* would result in boundary tone of H-H%, similarly *rf* would result in H-L% and *fr* in L-H%. Finally *ff* would give L-L%.

E. Identifying Salient Words

We use two mechanisms for identifying salient words. The first is based on the pitch, energy and duration features of the word following the evidence for prosodic salient outlined in section III-A above. If no words are marked as salient using these features we then use the word marked as salient via the H* markers above. Within each utterance we normalise the values of f0, energy and duration and then multiply the normalised values together to give an overall measure of salience for each word. A word is then marked as salient if the normalised measure is larger than the average normalised measure based on all words in the utterance. This has the effect of highlighting words which have not only high pitch features but also extended duration or are said very loudly. For single word utterances we compute the first standard deviation of pitch, energy and duration respectively for the whole interaction session. If any of the pitch, energy or duration features are larger than the first standard deviation for the complete session for that feature then the word is marked as salient.

F. Verification

In order to verify that the prosodic labelling system was functioning effectively we applied the system on the first 12 examples of the ToBI Labelling Guidelines [23] which had been manually labelled in Praat and the appropriate sound files provided [31], [32]. We measured the efficacy of the marking

by checking how well salient words were marked, whether the boundary tone was identified and whether the first pitch marking was correct. As an example test 4 is shown below:

```
Phrase:       Marianna made the marmalade.
Human
Labeller->   L+H*L-H% L*        H*     L-L%

Salient : Marianna,made,marmalade
 result-> Marianna,made,marmalade 100%
Boundary: L-H%,L-L%
 result-> L-H%,L-L%                100%
Pitch   : L+H*,L*
 result-> L+H*,L+H*               50%
```

The reason for difference in the pitch marking is that the second phrase has the word 'made' signalled by falling then rising pitch. As this is detected before the lowest pitch value on the word 'the' the algorithm assumes a L+H* contour as being more representative.

Complete test results are displayed in table II and showed average performance of 71%, 80% and 71% in the three categories of salient words, boundary tone and pitch marking. Differences occurred firstly due to extremely high peaks in the sound files (background clicks and coughs) which were labelled seperately by the human labeller but interpreted as high pitch markers by our software, secondly due to the fact that we are not processing down-steps (a further sophistication of the labelling system described in the ToBI manual) and finally that the prosodic phrase boundaries were slightly different in some cases. This latter difference in some cases being due to the *subjective* strength of its association with the next word which is employed by human labellers (see [23, page 9]).

TABLE II
VERIFICATION OF LABELLING SYSTEM AGAINST ToBI TRAINING
MANUAL

Test Example	Salient Words Correct	Boundary Tone Correct	First Pitch Highlighted
1	Partial (66%)	All (100%)	Partial (50%)
2	All (100%)	All (100%)	All (100%)
3	All (100%)	All (100%)	All (100%)
4	All (100%)	All (100%)	Partial (50%)
5	All (100%)	All (100%)	All (100%)
6	Partial (66%)	Partial (50%)	Partial (50%)
7	Partial (66%)	Partial (50%)	Partial (50%)
8	Partial (66%)	All (100%)	All (100%)
9	All (100%)	All (100%)	All (100%)
10	All (100%)	All (100%)	All (100%)
11	All (100%))	All (100%)	All (100%)
12	Partial (33%)	Partial (50%))	Partial (50%)

VII. RESULTS

Our first research question asked whether a human interacting with a robot would change their prosody as the humans perception of shared belief increased. This was analysed by comparing the number of contours containing H* in their first pitch markers (e.g. H*, H*+L, L+H*) against those with L*'s (e.g. L*, L*+H, H+L*). The analysis is presented in fig. 3. What is clear is that 2 out 3 utterances used are H* contours throughout the sessions implying that the human is still attempting to convey information to the robot. There is a larger usage of H* contours between the 1st and 2nd sessions and this may be expected as the robot in this study does not provide any feedback in the first session, thus the human has no idea what the robot knows or doesn't know in the first session, but realises the robot knows very little when interacting during the second session. The four participants vary widely on session five (when the robot is using shape words correctly). Three of the participants appear to show a slight drop in H* contours throughout, however the overall change (shown in fig. 4) is not convincing due to the wide variance on the fifth session.
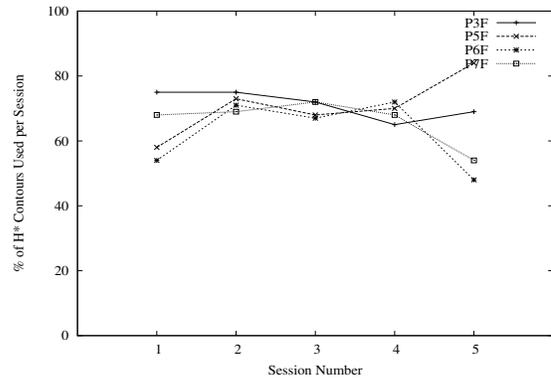

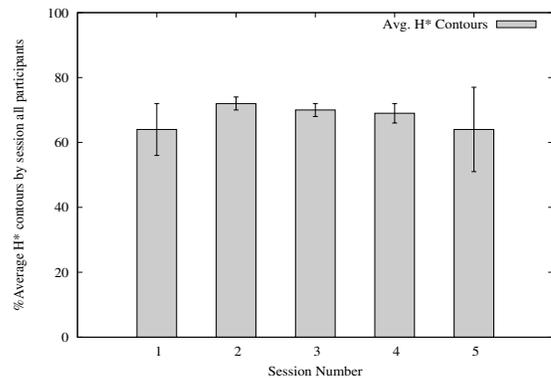Fig. 3. The graph shows the percentage of H* contours for each participant over the five sessions

Fig. 4. The graph shows the average percentage of H* contours by session for all participants

Our second research question asked whether prosodic boundary tones change from 'question to 'non-question' during the interaction sessions. Question contours are signalled by H-H% and H-L% boundary tones. Figs. 5 and 6 indicate that although there is a relatively wide variance there appears to

be a general reduction in question contours over the sessions. This we think would be expected as the robot responds more and more as the sessions progress. Thus the human is more certain that the robot has understood, although it may not have understood correctly. Thus for example instead of saying 'can you see the square?' the human knows that the robot can see something, but may be responding correctly or incorrectly (e.g. it might say 'square' or may be 'moon' or 'star' or something else). And so the response is no longer 'can you...?' but rather 'no, it's a square' or 'yes, that's right'. Our final question
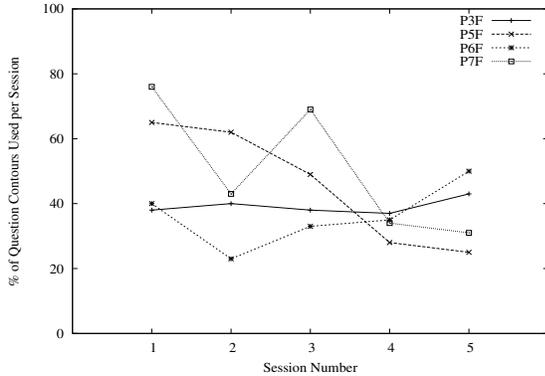

Fig. 5. The graph shows the percentage of prosodic question contours for each participant over the five sessions
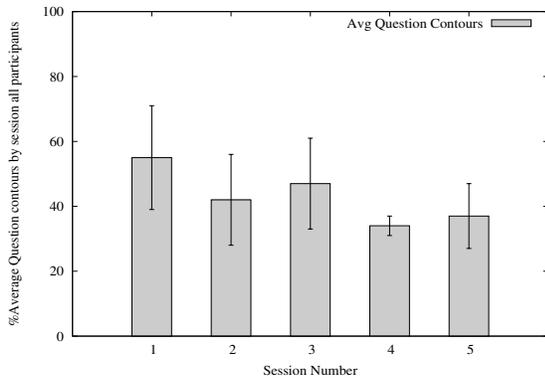

Fig. 6. The graph shows the average percentage of prosodic contours by session for all participants

asked whether prosodic salience can be used to extract relevant words from the interaction. To assess this we counted the number of nouns used overall, the number of words selected as salient and the number of nouns selected in the salient word set. To give a measure of effectiveness we computed how well the extraction worked when compared to a random selection. The results are shown in table III. It seems that the combination of pitch, energy and duration is highly effective for extracting salient words from this corpus. Although we were not counting adjectives it was also noticable that in the later sessions adjectives also started to appear consistently as salient.

Participant	P03	P05	P06	P07
Total Nouns Spoken	13%	27%	18%	16%
Nouns Selected As Salient	88%	98%	90%	97%
Selected Nouns vs. Random Selection (=1.00)	1.86	1.97	2.22	2.22

VIII. CONCLUSION

Our original objective in this case study was to gauge whether prosodic features which provide attentional and intentional cues to a listener could be used to scaffold the acquisition of lexical meaning by a robot interacting with a human employing CDS. As part of this objective we have shown that it is possible to label the speech signal in a semi-automatic manner and verified it against human labelled samples. From our analysis of the robot-directed speech corpus it seems apparent that using this method word salience could be used to move beyond simple single word noun association to a two-word stage where adjectives and possibly verbs could be used. Question contours appear to provide a method for partly assessing mutual belief, however it is less clear whether intonational contours could provide the robot with a mechanism to signal whether the speaker is attempting to provide new information. Clearly this issue may be due to inaccuracies with our extraction mechanism or to the limited amount of data in the case study or simply that humans do not react prosodically to robots in a manner which they may do with human chidren.
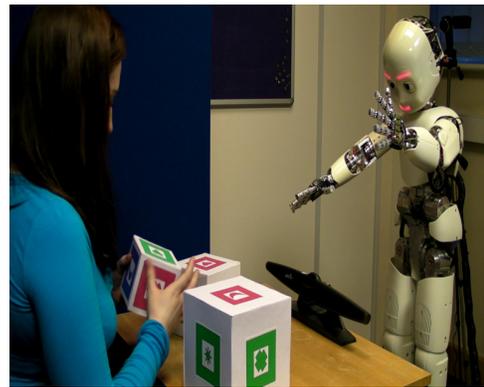

Fig. 7. Part of a session in our ongoing study with the humanoid robot iCub [33]. This study has 20 participants, each participant is asked over five 2-minute sessions with the robot to explain the shapes, colours and size of the boxes. The robot also reacts with various pointing, pushing and grabbing gestures. From session 2 onwards the robot responds to the presented shapes by querying its association between salient words spoken by the participant in the previous session with its sensorimotor history and expressed a chosen word.

It may be that a deeper analysis is necessary and as such our current ongoing research employs the techniques presented above in a much larger study involving 20 participants explaining differently sized, coloured shapes and interacting with the iCub [33] robot. We hope that this will provide a clearer

picture as to whether prosodic contours can be a useful tool for scaffolding meaning via the perception of shared belief.

Acknowledgment

References

[1] J. Saunders, C. L. Nehaniv, and C. Lyon, "Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors," in *Proc. Second International Symposium on New Frontiers in Human-Robot Interaction, AISB Convention, Leicester, UK.* AISB, 2010.

[2] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "Kaspar - a minimally expressive humanoid robot for human-robot interaction research," *Applied Bionics and Biomechanics, Special Issue on 'Humanoid Robots'*, vol. 6, no. 3, pp. 369–397, 2009.

[3] E. V. Clark, *First Language Acquisition*, 2nd ed. Cambridge,UK: Cambridge University Press, 2009.

[4] M. Saxton, *Child Language: Acquisition and Development*. Sage Publications, 2010.

[5] R. N. Aslin, J. Z. Woodward, N. P. LaMendola, and T. G. Bever, "Models of word segmentation in fluent maternal speech to infants," in *Signal to Syntax*, J. Morgan and K. Demuth, Eds. Lawrence Erlbaum, 1996.

[6] E. S. Kim, K. Gold, and B. Scassellati, "What prosody tells infants to believe," in *Proc. 7th International Conference on Development and Learning (ICDL), Monterey, USA.* IEEE, 2008.

[7] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds. MIT Press, 1990, p. 271311.

[8] J. Sachs, B. Bard, and M. L. Johnson, "Language learning with restricted input: Case studies of two hearing children of deaf parents," *Applied Psycholinquistics*, vol. 1, pp. 34–54, 1981.

[9] M. Saxton, "The inevitability of child directed speech," in *Advances in Language Acquisition*, S. Foster-Cohen, Ed. Palgrave MacMillan, 2009, pp. 62–83.

[10] K. Pitsch, A.L.Vollmer, J. Fritsch, B. Wrede, K. Rohlfing, and G. Sagerer, "On the loop of action modification and the recipient's gaze in adult-child interaction," in *GESPIN Gesture and Speech in Interaction, Poznan, Poland*, 2009.

[11] C. Moore and D. Povinelli, "Differences in how 12- and 24-month-olds interpret the gaze of adults," *Infancy*, vol. 11, pp. 215–231, 2007.

[12] K. Rohlfing, M.R.Longo, and B.I.Bertenthal, "Pointing. does gesture trigger shifts of visual attention in human infants?" in *Poster presented at 14th Biennial International Conference on Infant Studies, Chicago,USA*, 2004.

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

[14] D. J. Matatyaho and L. J. Gogate, "Type of maternal object motion during synchronous naming predicts preverbal infants' learning of wordobject relations," *Infancy*, vol. 13, pp. 172–184, 2008.

[15] M. Rolf, M. Hanheide, and K. J. Rohlfing, "Attention via synchrony: Making use of multimodal cues in social learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 55–67, 2009.

[16] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk.* Erlbaum, 1995.

[17] P. Monaghan and M. Christiansen, "Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation," *Journal of Child Language*, 2010, in press.

[18] A. Brandone, R. M. Golinkoff, R. Pulverman, M. J. Maguire, K. Hirsch-Pasek, and S. M. Pruden, "Speaking for the wordless: Methods for studying the foundations of cognitive linguistics in infants," in *Methods in Cognitive Linguistics: Human Cognitive Processing*, M. G. Marquez et al., Ed. Benjamins, 2007, pp. 345–366.

[19] C. Moon, R. P. Cooper, and W. Fifer, "Two-day olds prefer their native language," *Infant Behaviour and Development*, vol. 16, pp. 495–500, 1993.

[20] P. W. Jusczyk, D. Houston, and M. Newsome, "The beginnings of word segmentation in english-learning infants," *Cognitive Psychology*, vol. 30, no. 3-4, pp. 159–207, 1999.

[21] E. D. Thiessen, E. A. Hill, and J. R. Saffran, "Infant-directed speech facilitates word segmentation," *Infancy*, vol. 7, no. 1, pp. 53–73, 2005.

[22] J. R. Hobbs, "The pierrehumbert-hirschberg theory of intonational meaning."

[23] M. E. Beckman and G. A. Elam, "Guidelines for ToBI Labelling," *Ohio State University Research Foundation*, no. 3, 1997.

[24] SysMedia, "Sysmedia word and phoneme alignment software," [Last visited 31 July 2009], 2009, http://www.sysmedia.com.

[25] M. Huckvale, University College London *et al.*, "Speech filing system," [Last visited 06 April 2011], 2011, http://www.phon.ucl.ac.uk/resource/sfs/.

[26] Z. Huang, L. Chen, and M. Harper, "An open source prosodic feature extraction tool," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2006.

[27] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2010, retrieved 29 March 2010 from http://www.praat.org/.

[28] Y. Sato, F. Brox, and C. L. Nehaniv, "Computational simulation of syllabification and syllable based word-discovery - submitted for publication."

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

[30] E. V. Clark, "Adult offer, word-class, and child uptake in early lexical acquisition," *First Language*, vol. 30, no. 3-4, pp. 250–269, 2010.

[31] ToBI, http://www.ling.ohio-state.edu/ tobi/, 1999, Tones and Break Indices [last visited 04 April 2011].

[32] A. Gravano, *ToBI on Praat.* Spoken Language Processing Group, University of Columbia New York , USA, 2008, http://www1.cs.columbia.edu/ agus/tobi/tobi-praat/manual.php[last visited 04 April 2011].

[33] iCub, "RobotCub – An Open Framework for Research in Embodied Cognition," http://www.robotcub.org/, 2004.