2010 IEEE-RAS International Conference on Humanoid Robots
Nashville, TN, USA, December 6-8, 2010

Towards Social Integration of Humanoid Robots by Conversational Concept Learning

Florian Kraft, Kevin Kilgour, Rainer Saam, Sebastian Stüker, Matthias Wölfel, Tamim Asfour and Alex Waibel
Institute of Anthropomatics, Karlsruhe Institute of Technology (KIT)
Geb 50.20, Adenauerring 2, 76131 Karlsruhe, Germany
{florian.kraft|kevin.kilgour|rainer.saam}@kit.edu
{sebastian.stueker|matthias.woelfel|asfour|alex.waibel}@kit.edu

***Abstract*— Several real world applications of humanoids in general will require continuous service over a long time period. A humanoid robot operating in different environments over a long period of time means that A) there will be a lot of variation in the speech it has to ground semantically and B) it has to know when a conversation is of interest in order to respond.**

Detailed natural speech understanding is hard in real scenarios with arbitrary domains. To prepare the ground for in-domain dialogs in real day-to-day life open domain scenarios we focus on an intermediate attention level based on conversation concept listening and learning.

With the aid of explicit semantic analysis new concepts from open domain conversational speech are learned together with how to react to them according to human needs. This can entail how the robot performs actions such as positioning and privacy filtering.

The corresponding attention model is investigated in terms of concept error rate and word error rate using speech recordings of household conversations.

***Index Terms*— humanoid robots, attention learning, concept learning, classification, explicit semantic analysis, distant speech recognition**

I. Introduction

The development of human like machines has been a dream and on the mind of mankind for a very long time. While in the past decades robots, a term coined in literature a long time before their technological development [1], have been extensively deployed in factories in order to facilitate the automation process of production, the development of humanoid robots is an active research area with many remaining problems, still needing to be solved.

The collaborative research center SFB588 *Humanoid Robots - Learning and Cooperating Multimodal Robots* [2] aims at the development of a household robot by the name of *Armar* [3], [4]. Armar, currently in its third generation, is intended to support humans at home, in the kitchen and the general household.

In order for a robot in a household environment to be accepted by the user and to maximise its usefulness, it is of utmost importance that the people that encounter the robot can interact with it in a natural manner. Users should not need to learn specific interfaces for controlling the robot or to use special controls that disrupt their normal behaviour. One of the most natural interfaces that can be used to achieve this is communication via speech. In addition to directly receiving instructions from the user, the robot should also be able to passively assess the current situation and context, and to predict the needs and demands of the occupants of the household. Active communication and passive observation of the situations require the implementation of perceptual interfaces, including the techniques of automatic speech recognition, concept detection, and dialog modeling.

When engaging in communication with humans, robots have to follow certain rules, in order to be socially acceptable, for example, the ability to identify and respect human behaviour and workflow by listening to human-human conversations and learning from feedback.

A. Learning by Observing Inter-Human Communication

In order for the robot to provide knowledge and interactions autonomously and to learn behaviour for new domains, we address in this work the task of autonomous concept learning from the routine observation of daily conversations.

For the perceptual speech recognition and the dialog component to work properly, both components have to show a high degree of flexibility. The tasks and concepts encountered in the robot's work environment can vary greatly from household to household. Thus, the robot's interaction components have to autonomously and flexibly adapt to the circumstances of its environment.

In our scenario, the robot will observe conversations between humans and automatically detect the concept in each, if it is already known to him, or will otherwise detect that the conversation was about an as yet unknown concept. In the case of an unknown concept the robot will initiate a short dialog in order to learn the new concept and to receive instructions for appropriate behaviour when the same concept is encountered again in the future. The robot will also learn which conversations are private and adapt its data-keeping methodology accordingly.

Behaviour that can be associated with the detection of a certain concept can be, for example, the repositioning of the robot, such as coming closer or following the user in order

This research is partially funded by the German Research Foundation (DFG) under Sonderforschungsbereich SFB 588: *"Humanoid Robots - Learning and Cooperating Multimodal Robots"*.

Special thanks go also to the people from the robotics group under Professor Dillmann for providing and helping to maintain the robot head with which the experiments were carried out.

to receive instructions or carry out a task, or to just continue its current task, or in other cases to interrupt its current work and to leave the conversing parties alone in order for them to have a private discussion. In the latter case, instead of leaving, the robot might also be allowed to continue its work and to listen to the conversation but should actively forget the perceptual data seen during the conversation, i.e. it should keep no record of it.

In summary, our general aim is social integration of the robot by *semantic grounding* at the *concept level*, which enables the robot to detect known concepts from inter-human dialogs and derive appropriate actions, and to detect and learn new, previously unknown concepts and associated actions by conducting a dialog with the user.

B. Related Work

Robot learning from humans by example or demonstration is a very active research area [5]. A lot of work has been devoted to learning specific, single tasks from human examples. Such tasks can be individual movements [6] or movement primitives [7].

In our work we focus on the observation of human interaction and communication instead of visually observing physical movements and imitating them. This kind of problem is strongly related to the problem of learning in dialog, e.g. in order to maintain a database of known persons at a workplace by a robot receptionist [8] or to build a model of social networks [9].

Also related to our task of observing inter-human conversations, is the task of detecting whether the robot is addressed by the user, or whether the user is interacting with somebody else [10]. In our work here, we concentrate on observing inter-human communication instead of situations in which the robot is addressed directly.

In [11] a cascade of HMMs is used for associative learning of language from visual and auditory streams sensed by a mobile robot. While in [11] speech concepts anchored by isolated words have been learned, our work deals with concept learning from real world observations of conversational and spontaneous inter-human speech as a whole. By using state of the art text classification methods we also make use of a large amount of human generated knowledge.

II. Concept Learning and Detection by Explicit Semantic Analysis

This paper distinguishes between: *concepts*, abstract ideas that we hope to find in a conversation; *categories*, predefined sets of human generated data; *word sequences*, the output of our speech recognition system and *documents*, a set of words. In order to learn and detect new speech concepts we transform the word sequence of a conversation observed by the humanoid robot into features representing its concept information. Generally speaking, learning concepts is achieved by partitioning the category feature space; classification is done by finding the most probable partition.

In the following subsection we first describe technically how an analysis of manually (explicitly) defined semantic categories can be performed. Henceforth we refer to learned semantic entities by the term *concept* while the manually predefined semantic entities are referred to as *categories*.

The second subsection describes how concepts are represented using an explicitly predefined feature space and how a new word observation is classified or learned within the new representation.

In practice, our model allows two different learning strategies:

- from word sequences observed in human-human conversations together with confirmation from a dialog
- in dialog with batch mode keyword enumeration

While learning from inter-human conversations uses whole conversation word sequences as given by the speech recognizer, in batch mode new concepts can be learned very fast with almost no training material. Thereby, people are given the ability to tell the robot that, for example, *knife* and *fork* are keywords for the concept *table setting* and let the robot also automatically connect *spoon* with this concept. The robot was equipped with some real world knowledge in form of predefined semantic categories with the help of *Explicit Semantic Analysis*.

A. Explicit Semantic Analysis

In the better known Latent Semantic Analysis [12] a document or sequence of words (from a speech recognizer for example) is represented as a vector in a latent category space. The categories are generated algorithmically and are not necessarily comprehensible to people. In contrast Explicit Semantic Analysis (ESA) developed by Gabrilovich and Markovitch [13] represents the word sequence as a vector in a category space where the categories have been defined by humans. Most commonly the articles in the online encyclopedia Wikipedia are used as the explicit categories. The word sequence is compared to each of the categories and with a tfidf measure (see section II-A.1) the importance of the category to the word sequence is determined.

Because we required a large amount of text to be associated with each category, we chose the categories in the open directory project (ODP)[1] to be our explicit categories. The ODP is a large hierarchically sorted directory of websites which is edited and maintained by human volunteers. It contains links to over 4 million websites sorted into over 500,000 categories. Its hierarchical tree-like structure allows us to associate the text in the websites which are linked to by a category, not only with their corresponding categories, but also with all their ancestor categories. The k categories C with the most associated[2] text are chosen to be the dimensions in our category space $\mathcal{C}_k$ giving us a mapping function:

$$\tau : \mathcal{D} \rightarrow \mathcal{C}_k \subset \mathbb{R}^k \tag{1}$$

$$\tau\left(w_i\right) = \left\langle \upsilon\left(w_i, c_1\right), \upsilon\left(w_i, c_2\right), \ldots, \upsilon\left(w_i, c_k\right)\right\rangle \tag{2}$$

[1] http://www.dmoz.org/

[2] the most general categories were disregarded

Because the categories c_j can be considered documents in the text classification sense we can use text classification methods to find the similarity υ between a word sequence w_i and a category c_j ($w_i, c_j \in \mathcal{D}$, the set of all possible documents). Documents (word sequences or categories) have lots of properties that can be useful in deciding how similar they are to one another. By far the most important property is its body of text out of which term counts (in our case word counts) are extracted and a feature vector is built.

1) Term-Frequency Inverse Document Frequency (TFIDF) Metric: A standard method of generating a feature vector $\overrightarrow{f_j}$ from a document c_j is to first extract a set of n terms from the sum of the text of all the documents and then weight them according to their occurrence in c_j. For each c_j we have a

$$\overrightarrow{f_j} = (f_{1,j}, f_{2,j}, ..., f_{n,j})$$

where $f_{i,j}$ denotes the weight of term i in document j.

Once a set of terms $|\mathcal{T}| = n$ has been decided upon a TFIDF function can be used to generate a document feature vector $\overrightarrow{f_j}$. In our case the number of terms is limited by the vocabulary of our speech recognizer. The function calculates each component of $\overrightarrow{f_j}$ from its term frequency $\text{TF}_{i,j}$ in c_j and the inverse document frequency IDF_i of $t_i \in \mathcal{T}$. The term frequency component measures how often a word occurs in a document. Let $\#(t_i, c_j)$ be the number of times t_i appears in c_j.

$$\text{TF}_{i,j} = \log \#(t_i, c_j)$$

The inverse document frequency measures how discriminative a word is. Words that appear in few documents have a high inverse document frequency and words in a lot of documents have a low inverse document frequency [14].

$$\text{IDF}_{i,j} = \log\left(\frac{|C|}{\#(C, t_i)}\right)$$

$\#(C, t_i)$ is the number of documents containing t_i. The logarithm of the quotient is used to blunt the effect of extremely rare words, which might only appear in one or two documents. Putting these together gives us.

$$f_{i,j} = \text{TFIDF}(t_i, c_j) = \log \#(t_i, c_j) \cdot \log\left(\frac{|C|}{\#(C, t_i)}\right) \quad (3)$$

As is, the TFIDF function does not take into account the length of a document. A term appearing once in a short document is more relevant than if it were to appear in a longer one.

One way to solve this is to normalise the vector generated by the TFIDF function.

$$f_{i,j} = \frac{\text{TFIDF}(t_i, c_j)}{\sqrt{\sum_{h=1}^{|\mathcal{T}|} \text{TFIDF}(t_h, c_j)^2}} \quad (4)$$

The same feature extraction method is applied to the new document or word sequence w that is to be mapped to the category space.

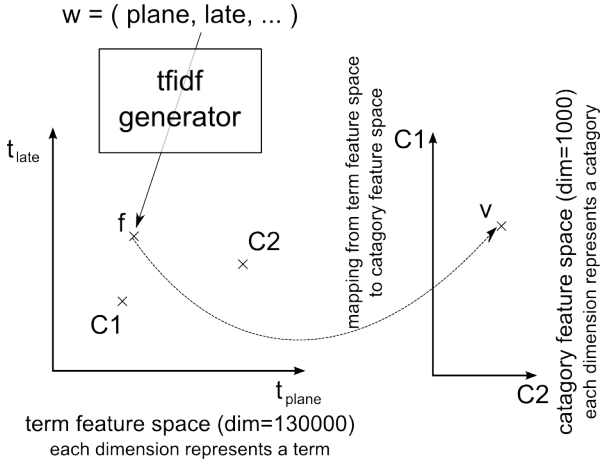

Fig. 1. Explicit Semantic Analysis performed on a short sequence of words. A sequence of words is first of all converted to a sparse term feature vector using tfidfs and then mapped to a vector in the category feature space by comparing it to the term feature vectors of each category.

$$f_{i,w} = \log \#(t_i, w) \cdot \log\left(\frac{|C|}{\#(C, t_i)}\right) \quad (5)$$

$$\phi_{\mathcal{T}}(w) = \overrightarrow{f_w} = (f_{1,w}, f_{2,w}, ..., f_{n,w}) \quad (6)$$

Because the selection of our terms $\mathcal{T}$ will remain constant we will henceforth refer to $\phi_{\mathcal{T}}$ simply as ϕ.

2) Cosine Similarity Metric: TFIDF vectors are built to be able to compare documents with each other. This requires a similarity metric. The cosine similarity metric is an easy and fast metric to compute. It is defined by the angle τ between the two vectors f_1 and f_2 that are to be compared.

$$\text{cossim}(f_1, f_2) = \cos(\varphi) = \frac{f_1 \cdot f_2}{|f_1||f_2|} \quad (7)$$

Since the TFIDF vectors are often already normalised ($|f_1| = 1$ and $|f_2| = 1$) the denominator part of this definition can be ignored. Also most TFIDF vectors are sparse, leading to very few nonzero terms in the numerator of the definition. This allows the cosine similarity metric to be calculated very fast making it an ideal function for υ which we can now define as.

$$\upsilon(w_i, c_j) = \phi(w_i) \cdot \phi(c_j) \quad (8)$$

B. Learning and Remembering Methodology

When the robot is told to learn a new concept or receives confirmation that the words that it picked up from a particular conversation belong to a new concept, it maps these words w to a concept vector v_q in category space and stores them (see Figure 1).

$$v_q = \tau(w) \quad (9)$$

Future word sequences $v_m = \tau(w_m)$ are compared to each of the previously stored concept vectors using the cosine

similarity metric thereby identifying the most similar learned concept.

$$v_s = argmax_{v_r} \frac{v_r \cdot v_m}{|v_r||v_m|} \quad (10)$$

Because word sequences will often not correspond to any of the previously learned concepts we introduce a pseudo concept *off-concept* which is returned when the similarity between the word sequence and the most similar concept is less than a threshold. Experimentation on a small text-only development set showed $\frac{|v_s|_{l2}}{2}$ to be a reliable concept dependent threshold. The component that performs this concept classification is hereafter referred to as the Explicit Semantic Topic Analyser (ESTA).

III. TOPIC BASED ATTENTION MODELING

Learning a concept representation using explicit semantic analysis given a word sequence was technically described above.

In the following we describe the whole attention and behaviour cycle. All attention model components are illustrated by figure 2.

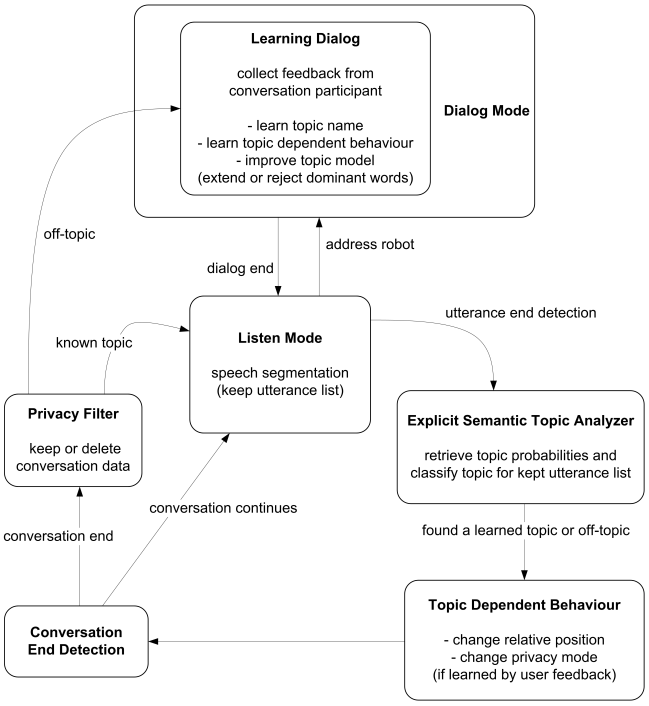

Fig. 2. Attention model schematic

A. Listen mode

The cycle starts in the listen mode. In this state word sequences are generated by our speech recognition system from acoustic observation of the humanoid's environment. Speech segments are kept in an utterance list, as well as permanently written to a database with connection to the corresponding audio data. As soon as a new speech segment is added we check whether a speaker is directly addressing the robot by speaking certain keywords. If the robot is addressed we enter the dialog mode. In all other cases the whole utterance list is forwarded to the Explicit Semantic Topic Analyzer.

B. Explicit Semantic Topic Analyzer

The Explicit Semantic Topic Analyzer interface provides the dialog and speech recognition system with learning and concept classification functionality.

Topic classification is performed for the whole utterance list in the classification case while the learning case updates or saves a category distance vector. The classification outputs previously learned concept names or off-concept within a couple of seconds.

C. Topic Dependent Behaviour

If an already learned concept was found the robot triggers connected behaviour. Currently we implement position changes to enable the robot to approach humans, or to leave the scene. Arbitrary actions could potentially be connected here. We further support a privacy mode which can also be attached to a concept in the dialog feedback. The robot can also have no action attached to a concept which leads to a continuation of the current task.

D. Conversation end detection

Conversation end detection leads to the deletion of an utterance list kept in listen mode, while if not explicitly deleted the database keeps record of what was said. A conversation end can be detected by various methods. A trivial realisation just uses the time stamp of the last utterance to detect a long speech pause. If the conversation continues we continue with speech recognition in the listen mode. Otherwise we change the state to privacy filter.

E. Privacy filter

Privacy mode means that after the conversation end we discard conversation data such as recognition hypotheses and raw audio recordings if the privacy flag attached to the classified concept is true. All off-concept cases lead to a learning dialog, whereas other cases lead to a start over in listen mode.

F. Dialog mode

As a result of a dialog request in the off-concept case our robot tags the observed conversation (word sequence) with a concept name and anchors the new concept by saving the association vector. During this off-concept learning dialog the robot repeats dominant words recognized in the previous conversation as a sanity check for mis-recognition. It requests settings for action triggers and the privacy mode to be used when the newly learned concept is encountered again.

The dialog conversation can also learn further concepts in batch mode rather than by observing human-human conversations. Thus the user can predefine a number of relevant behaviour dependent concepts in a quick way. This feature also allows us to set the robot into a controlled *"testing-scenario'* in which we can, together with a set of prerecorded user conversations, perform experiments with by varying environmental parameters.

IV. EXPERIMENTS

We first describe the evaluation data, followed by the description of our speech recognition system used in the listen and dialog mode, evaluation metrics and finally present the results on remembering learned concepts.

A. Data collection

For evaluation of classification performance within the end-to-end system learning cycle we recorded 28 conversations of 10 speakers next to our robot's head within the household domain. The recorded conversations with a total duration of 80 minutes contained concepts about

1) setting a table
2) arranging a private meeting
3) cleaning and tidying up
4) asking where a specific person is currently located
5) further concepts (i.e. weekend experiences, holidays etc.).

The recorded conversations were carried out by non-native English speakers and contain a significant amount of disfluencies. Each observed conversation participant was recorded with a lavalier microphone while the robot head's microphones were listening in parallel from a short distance. To compare the effect of different distances on the same conversations and the channel difference between the lavalier and robot head microphones, we later played back the close talk recordings through a loudspeaker and let the robot head and the close talk microphones listen again from further distances.

The ESTA was trained on text extracted from all the websites linked to in the ODP. When cleaned this data set consists of just over 16 GBytes of text, roughly 9 GBytes of which came from the 5,000 categories that were selected for use. For terms the 130,000 word vocabulary from the ASR language model was used.

B. Automatic Speech Recognition System

All experiments in this work were performed with the help of the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [15]. The acoustic model used in our experiments utilises 3-state sub-phonetically tied semi-continuous Hidden Markov Models composed of 16,000 quinphone models over 4,000 codebooks with a maximum of 64 Gaussians per model. The preprocessing stacks 15 frames of 15 mel scaled warped Minimum Variance Distortionless Response (wMVDR) cepstral coefficients[16]. The resulting feature vector is reduced to 42 dimensions using linear discriminant analysis. The model was trained on 140 hours of transcribed speech data composed of European Parliamentary Plenary Sessions [17], conference talks given by non-native speakers from the Translanguage English Database (TED) [18], and broadcast news recordings. The training procedure consisted of merge-and-split training on samples extracted with the help of existing forced alignments using one global semi-tied covariance (STC) transformation [19], followed by two iterations of viterbi training to compensate for wrong alignments. The models were then further improved by several iterations of minimum mutual information estimation (MMIE) training.

To jointly compensate for additive noise as well as reverberation we used a feature enhancement technique, based on particle filtering and multi-step linear prediction, which has previously demonstrated significant reductions in word error rate on real data. Even though the acoustic environment is quite different from the experiments performed in [20] we have decided to use the same enhancement setup, as the free parameters have been demonstrated to be similar for very different acoustic environments. As the acoustic training material differs to [20], a retraining of the clean speech model within the PF was necessary for For optimal performance.

The language model used is an interpolation of 4-gram language models trained on transcripts of news texts from the Gigaword Corpus and data collected from the World Wide Web, for a general English transcription task. The models were built and interpolated using the SRI Language Modelling Toolkit [21]. The resulting language model was pruned to slightly more than $6x10^7$ 3-grams and 4-grams.

C. Evaluation Metrics

We evaluated our system using word error rate (WER) and task error rate (TER). TER measures how often the robot incorrectly identifies a concept.

D. System Evaluation

To evaluate how well the system performs in a real world setting we initially taught the robot 4 concepts that it might encounter in a household setting. This was done having a person who was not present at the data collection tell it that for example *table*, *spoon*, *coffee*, *plate* and *set* are important words for the concept *table setting*. The other concepts *arrange meeting*, *cleaning* and *locate person* for which data was recorded were taught in an analogous manner.

After this learning phase the original recorded audio from both tested microphones (see *near* column in table I) was passed directly to the speech recognition component of our system.

Its performance was also evaluated on distant speech. To do this the audio from the lavalier microphone was played through a speaker system at distances of 60cm and 120cm and recorded from the robot's microphones.

As can be seen in table I the performance of our speech recogniser deteriorated from 37.0% WER to 69.6% WER when we increased the distance of the robot to the speaker from nearby to 120cm. The use of a special distance microphone decreased our WER at 120cm to 58.4%. Although the task error rate also steadily got worse with increasing distance the error rate varied a lot more between the tasks. While *"arrange meeting"* was almost never correctly detected in the hypotheses from the speech recogniser, *"table setting"* was, even at 120cm, correctly identified 80% of the time. When tested on the references only one instance of *arrange meeting* was incorrectly identified. This may be because the speech recogniser incorrectly recognised the

Source	near		60cm		120cm	
	TER	WER	TER	WER	TER	WER
References	7.1	-	-	-	-	-
Lavalier mic	28.6	37.0	39.3	52.0	57.1	69.6
Distant mic	39.3	42.9	50.0	50.4	57.1	58.4
Meeting	20 / 80	38.9	100.0	55.5	100.0	64.2
Locate Person	0 / 0	36.9	60.0	57.9	80.0	73.6
Cleaning	20 / 80	34.5	80.0	45.0	80.0	51.1
Table Setting	0 / 0	34.4	20.0	51.0	20.0	56.3
Off-concept	0 / 0	41.0	12.5	52.4	25.0	61.8

TABLE I

TASK AND WORD ERROR RATES (IN %) MEASURRED ON THE LAVALIER AND DISTANT MICROPHONES AT VARIOUS DISTANCES. THE TER OF INDIVIDUAL CONCEPTS WAS MEASURED ON THE HYPOTHESES FROM THE DISTANT MIC IN THE 60CM AND 120CM COLUMNS. IN THE NEAR COLUMN THE LEFT VALUE IS THE TER MEASURED ON THE REFERENCES; THE RIGHT VALUE IS MEASURED ON THE HYPOTHESES FROM THE LAVALIER MIC

semantically important vocabulary in these conversations. Another interesting observation is that most errors were not incorrect classification (i.e. *"arrange meeting"* miss classified as *"table setting"*) but rather Off-concept errors (*"arrange meeting"* miss classified as *"Off-concept"*). This explains why the Off-concept error rate is consistently low. An interesting conclusion is that at word error rates below 40% concepts can be correctly identified when they are well defined by their keywords.

V. CONCLUSIONS AND FUTURE WORKS

In this paper we presented our work towards giving humanoid robots adaptive behaviour which learns from and reacts to our needs.

A concept learning and remembering cycle was proposed and implemented, by which humanoid robots are able to learn unseen concepts from conversations and attach behaviour to that new concept according to a human suggestion. The term *learning* implies here that an anchored concept will be recognised again, even if a completely different vocabulary is used. The technical realisation of intelligent concept learning and matching was discussed using explicit semantic analysis. By applying this learning cycle in the day-to-day life of humanoid robots, we will enable them to ground the meaning of concepts by letting them autonomously reactivate behaviour. In the proposed model, behaviour is attached by human suggestion. As a result we can think of an incrementally improved autonomous adaptation to human needs over long time periods, leading to integration of humanoid robots into our functional as well as social awareness.

The concept detection could be improved by analysing the lattice output of the speech recognition system and taking the word's acoustic confidences into consideration. Continuous research on feature enhancement in our distant speech frontend will incrementally extend the distance at which concepts can be reliably detected.

REFERENCES

[1] Karel Čapek, "R.u.r—rossum's universal robots," 1921.

[2] "The SFB 588 website," www.sfb588.uni-karlsruhe.de.

[3] T. Asfour, K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Genova, Italy, December 2006, pp. 169–175.

[4] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe Humanoid Head," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Daejeon, Korea, December 2008, pp. 447–453.

[5] Aude Billard and Roland Siegwart, Eds., *Robotics and Autonomous Systems—Special Issue: Robot Learning From Demonstration*, vol. 47, Elsevier, June 2004.

[6] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," in *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*. 2009, pp. 763–768, IEEE.

[7] M. Pardowitz and R. Dillmann, "Towards life-long learning in household robots: the piagetian approach," in *Proceedings of 6th IEEE International Conference on Development and Learning*. 2007, IEEE.

[8] Hartwig Holzapfel, "Knowledge acquisition in dialogue with the interact receptionist robot," in *Proceedings of the Second Swedish Language Technology Conference (SLTC-08)*, 2008, pp. 47–48.

[9] Felix Putze and Hartwig Holzapfel, "Islenquirer: Social user model acquisition through network analysis and interactive learning," in *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, December 2008.

[10] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proceedings of 6th International Conference on Multimodal Interfaces (ICMI 2004)*, October 2004.

[11] S. Levinson, K. Squire, R. Lin, and M. McClain, "Automatic language acquisition by an autonomous robot," *AAAI Spring Symposium on Developmental Robotics*, 2005.

[12] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, 1990.

[13] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007, pp. 6–12.

[14] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," 1987.

[15] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one pass-decoder based on polymorphic linguistic context assignment," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01)*, Madonna di Campiglio Trento, Italy, December 2001, pp. 214–217.

[16] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[17] Christian Gollan, Maximilian Bisani, Stephan Kanthak, Ralf Schlüter, and Hermann Ney, "Cross domain automatic transcription on the tc-star epps corpus," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, PA, USA, March 2005, pp. I–825-I–826, IEEE.

[18] Lori F. Lamel, Florian Schiel, Adrian Fourcin, Joseph Mariani, and Hans G. Tillmann, "The translanguage english database (ted)," in *Proceedings the Third International Conference on Spoken Language Processing (ICSLP 94)*, Yokohama, Japan, September 1994, pp. 1795–1798, ISCA.

[19] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," Tech. Rep., Cambridge University, Engineering Department, February 1998.

[20] M. Wölfel, "Enhanced speech features by single channel joint compensation of noise and reverberation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 2, Feb. 2009.

[21] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*. ISCA, 2002.