Hey Robot, Why Don't You Talk To Me?

Hwei Geok Ng*, Paul Anton, Marc Brügger, Nikhil Churamani*, Erik Fließwasser, Thomas Hummel, Julius Mayer, Waleed Mustafa, Thi Linh Chi Nguyen, Quan Nguyen, Marcus Soll, Sebastian Springenberg, Sascha Griffiths, Stefan Heinrich, Nicolás Navarro-Guerrero, Erik Strahl, Johannes Twiefel, Cornelius Weber and Stefan Wermter

Abstract— **This paper describes the techniques used in the submitted video presenting an interaction scenario, realised using the Neuro-Inspired Companion (NICO) robot. NICO engages the users in a personalised conversation where the robot always tracks the users' face, remembers them and interacts with them using natural language. NICO can also learn to perform tasks such as remembering and recalling objects and thus can assist users in their daily chores. The interaction system helps the users to interact as naturally as possible with the robot, enriching their experience with the robot, making it more interesting and engaging.**

I. Introduction

With advancements in Human-Robot Interaction (HRI) research, humans are expected to be surrounded by artificial agents in the future, assisting them in day-to-day tasks. An agent, to blend well in its human-centric environment, is desired to be sociable and interactive [1], [2]. The agents also need to take into account the psychological aspects of an interaction when it comes to interacting with humans [3].

This paper aims to demonstrate the effect of improving an agent's interaction capabilities, in terms of making the interaction with human users as natural as possible. It presents an interaction scenario with the Neuro-Inspired Companion (NICO) robot (Fig. 1) [4], holding an engaging conversation with the users. NICO detects and tracks the face of the people talking to it and recognises them. It then asks them about their name and personal preferences and recalls this information when needed.

II. Background

Agents operating in a social environment need to be able to model effective and engaging interactions with the users [5]. Humanoid robots, in particular, have the additional challenge to appear as human-like as possible [6]. Furthermore, they also need to possess capabilities such as vision, speech synthesis, gestures etc. in order to interact naturally with humans. Agents, such as *Kismet* [2], also consider factors like active behaviour, affective appraisal, speech input and body gestures to understand and adapt to the user during an interaction. As the agents get closer to appear and act human-like, carefully avoiding the Uncanny Valley [7], it becomes more and more pleasing and interesting for humans to interact with them. This is expected to improve the users' perception of the robot and they are likely to rate it higher in the context of overall competence.

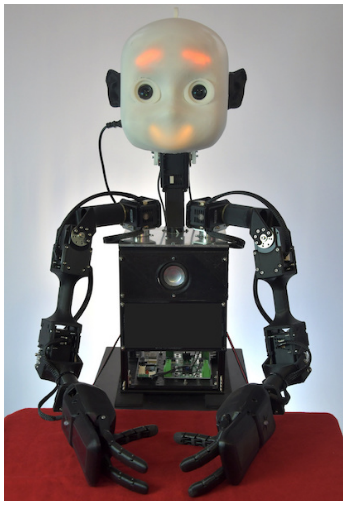

Fig. 1: Neuro-Inspired Companion Robot (NICO).

A. The NICO Robot

NICO, the Neuro-Inspired Companion robot (Fig. 1), is a middle-sized developmental robot designed and built for neuro-cognitive research. It is designed and built to interact with humans as a companion in a natural environment where it can learn from experience and instruction. NICO makes use of its rich sensory, motor and facial expression capabilities such as bimanual manipulation, synthetic speech, facial LED display and cameras for stereo vision. For this work, it was supplemented with an external microphone to capture the voice of the user accurately.

B. Interaction Scenario

The interaction aims to engage and motivate the user to take part in a conversation with NICO. The user interacts with NICO using natural language, which then models a conversation with the user asking for her name and personal preferences. The interaction can be triggered on or off by the user based on when she feels comfortable talking to NICO, and the learning scenario can be triggered as and when needed. When triggered, NICO tries to locate, identify and track the user interacting with it and models a conversation with her based on an advanced state-based Dialogue Manager (DM). The learning scenario for the interaction, the Humanoidly Speaking [8], [9] scenario, originally implemented

[*] All authors are with the Knowledge Technology Institute, Department of Informatics, Universität Hamburg, Germany.
e-Mail: {5ng,5churama}@informatik.uni-hamburg.de

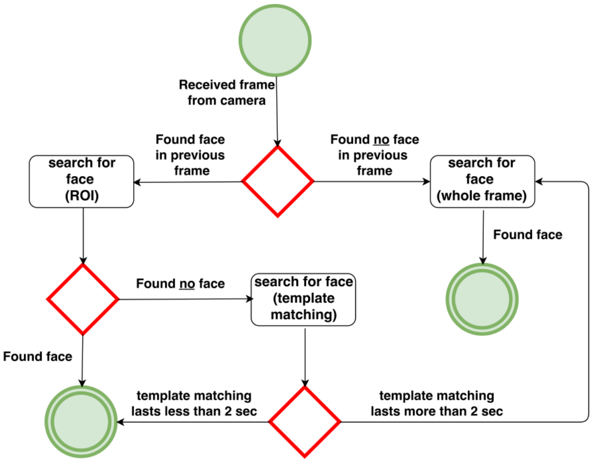

Fig. 2: Algorithm for Face Detection.

on the NAO robot, was modified and implemented using the NICO robot. It involves an object learning scenario where users can teach objects to the robot using natural language instructions.

The following sections detail different components of the interaction module and their contribution to the overall interaction.

III. The Model

NICO processes both visual and auditory information and uses this information to respond accurately to the user using synthetic speech. The different tasks entailed in realising the interaction scenario are devised as ROS[1] nodes, which interact with each other, as well as internally using ROS messages and services.

A. Face Detection and Tracking

Humans use their vision capabilities to look for objects of interest in their visual field such as faces, objects to manipulate etc. Thus, modelling an interaction with humans, a humanoid robot should also make use of its vision capabilities to interact naturally with the user. The robot must be able to locate the user in the visual frame and track the user's movement. This is implemented in NICO using the built-in 2 MegaPixel cameras modelled as the NICO's 'eyes'.

The face detection algorithm (Fig. 2) makes use of Haar-like features based cascades [10] and improves their performance using template matching, which searches for the face template in the new frame, derived from the last detected face. Once a face is found, the algorithm looks only in the region of interest, defined by the position of the face but with double its width and height. Therefore, the algorithm allows for some tolerance in movement between consecutive frames. This also reduces the area in which the algorithm has to look for a face thus speeding up the process of detection with the Haar cascades.

These improvements in the conventional Haar-like feature based face detection and template matching result in a more efficient and robust face detection algorithm which provides a strong basis for an effective face tracking mechanism.

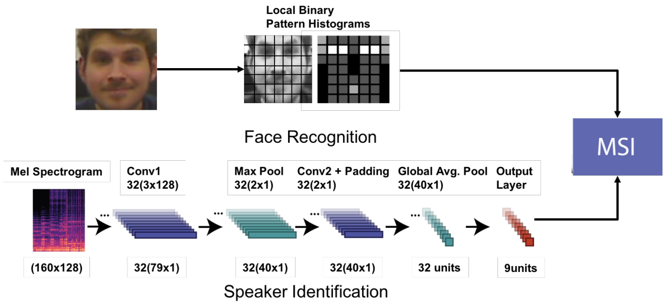

Fig. 3: Person Identification using Multi-Sensory Integration (MSI) based on Local Binary Pattern Histograms (LBPH) for Face Recognition and a Convolutional Neural Network (CNN) for Speaker Identification.

The face tracking algorithm tries to keep the detected face in the centre of NICO's visual frame. The distance of the face to the centre of the frame is calculated and used to compute the angle with which NICO's head needs to be rotated so as to align the centre of the face with the centre of the visual frame.

B. Person Identification

An important attribute of a natural human conversation is to recognise the person we are interacting with and treating her in a personalised manner. Thus, differentiating between different persons becomes a necessary task for a robot aiming to model an interaction with human users.

This is achieved using a Multi-Sensory Integration (MSI) approach (Fig. 3) where the results from face recognition and speaker identification are combined together using a weighted average resulting in the final prediction of the system.

1) Face Recognition: Face recognition is an active research field [11]–[13] and there are a number of effective approaches for solving this problem adequately. For this implementation, a Local Binary Pattern Histograms (LBPH) [12] approach was used. The classifier was trained using the dataset of face images collected from the project team members.

The LBPH algorithm uses the face detection results i.e. the grey-scale face image of the user and computes a binary pattern histogram [12] for a window around each pixel in the image. This histogram is then compared to the histograms computed from the training images giving a distance measure for each label. The label with the shortest distance is assigned as a prediction for the input face image. Once a face is recognised, a confidence measure is associated with each prediction which defines how trustworthy the prediction can be interpreted to be. Thus a known person would correspond to a high confidence for one of the known faces while an unknown person would yield low confidences for all the known faces. In case a new person is encountered, data for that person (300 frames of face images) is recorded, and the classifier is trained again using the entire dataset.

[1] http://wiki.ros.org [Accessed 16.03.2017]

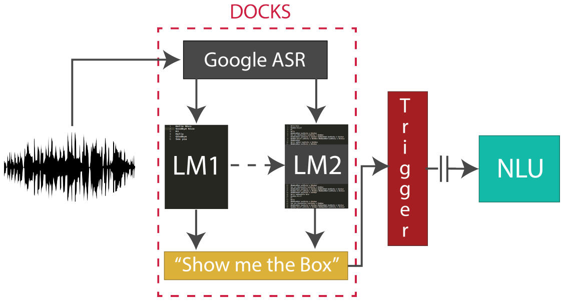

Fig. 4: Speech Processing based on the DOCKS [17] framework using two Language Models (LM).

2) Speaker Identification: Recognising persons using their voice, in addition to face recognition helps to identify a user, even when the user is not in the visual frame.

A Convolutional Neural Network (CNN) is used to perform speaker identification on Mel-Spectrograms derived from overlapping windows of the speech audio signal. As CNNs can be used to learn relevant features directly from the audio signal [14], [15], they avoid various pre-processing steps and the selection of engineered features such as Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are known to be rather sensitive to noise which can lead to a decrease in performance when dealing with noisy environments [16]. Furthermore, MFCCs might neglect important information by drastically decreasing data dimensionality. Considering spectrograms or even raw audio as input and allowing the model to derive relevant features from the input, more fine-grained characteristics with respect to the speakers' way of speaking can be captured. The proposed CNN performs 2D convolution over both frequency and time, thereby allowing to detect relevant structures in both domains. For this implementation, the CNN model is trained on a speech dataset consisting of samples from project team members.

When the confidence is not high enough (empirically defined threshold), the current speaker is classified as being unknown. In this case, the speech recordings for the unknown speaker, together with known speakers, are used to retrain the network. This allows NICO to learn new speakers during the interaction scenario.

C. Speech Processing

Speech processing is realised using the DOCKS framework [17] which provides the benefit of combining domain-specific knowledge with predictions from Google's cloud-based Automatic Speech Recognition (ASR). This is achieved by selecting the best matching hypothesis based on the Levenshtein distance between the phoneme sequences derived from Google ASR output and those derived from a domain specific sentence list or grammar. This avoids the problems that might arise due to recognition of words that are out of the domain, thus enhancing the performance of natural language understanding and dialogue management.

Integrating the cue-phrases (for example "*Hello NICO*") in a large Language Model (LM) would hinder the system from robustly detecting these phrases in a conversation. On the other hand, small and specific Language Models result in only limited interaction capabilities. The proposed system (Fig. 4) combines both a small Language Model (LM1) containing the cue-phrases and a large Language Model (LM2) modelling the full interaction. When recognising speech, the small language model is used first and only if the best matching phrase in the small Language Model is detected with a high confidence, the hypothesis resulting from this model is selected and used. Otherwise, the best hypothesis resulting from the large Language Model is applied.

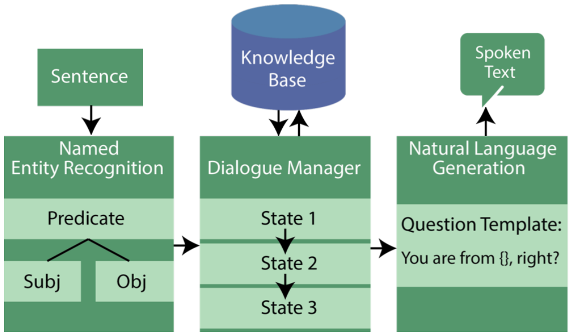

Fig. 5: Conversation Modelling based on Named Entity Recognition, State-Based Dialogue Management and Natural Language Generation.

D. Conversation Modelling

The ability of a robot to understand human utterances and in turn give proper responses to the user plays a central role in the human-robot interaction scenario. Conversation Modelling provides means to extract information about those concepts from human utterances and formulate appropriate responses. Conversation Modelling is also responsible for updating changes in the model about the world as new information is obtained.

Named Entity Recognition (NER) is used to extract relevant information i.e. entities and classes from the input sentences. For example, for the sentence:

"My name is Marcus.",

NER creates the information slots `['Marcus' PER]` which can then be derived to yield the predicate `Name(person_id, 'Marcus')`. When recognising entities, words are checked and referred to a list containing all possible entities.

Apart from responses containing specific information which are catered well by the NER, the system is also able to understand generic responses such as *yes* and *no* as well as names that are not on the list. This is achieved by implementing a context-free grammar setting to allow the expression of agreement and disagreement as well as all possible names as parameters in the sentence.

Upon receiving the semantics of the user's utterances from NER in the form of predicates, the *Dialogue Manager (DM)* checks for the relevance of these semantics to the current state of the dialogue. If necessary, information from the

Knowledge Base (KB) is obtained and a decision on the response to be given to the user is made. A finite-state DM is implemented using the SMACH[2] python library with 32 states, where each dialogue is a state in the state machine.

The *Knowledge Base (KB)* provides a way of modelling pre-existing knowledge as well as to incorporate knowledge acquired by the DM during the conversation. The KB consists of three main components: a database for knowledge representation, an inference engine, and an interface to the inference engine. Facts are divided into universal and case specific facts. The universal facts allow persisting information about already known persons as well as knowledge about capital cities of different countries across multiple runs of the system, which forms the base of NICO's knowledge about the world. As the conversation evolves, case specific facts can be incorporated in the form of dialogue-based contextual knowledge and user specific information.

The DM receives text predicates, a person ID, and a confidence level associated with that person ID (Section III-B) as input. The next state is then executed based on the conversation level and an action ID with additional context is passed on for natural language generation as the output.

Natural Language Generation (NLG) then produces a text representation of what the robot should say and utters the sentence through Text-to-Speech synthesis. This is realised using a template based approach, which provides an easy-to-write and easy-to-understand way to generate sentences [18], [19]. There are 114 sentences which are mapped to the respective action IDs. When an action ID is received from the DM, the respective sentence is uttered using the Google text-to-speech engine.

IV. Conclusion and Future Work

The paper describes an interactive robot, as presented in the video, which can operate in a human-centred environment and hold engaging and interesting conversations with humans. The interaction scenario involves tracking the users' face, recognising the users based on visual and auditory information and engaging them in a personalised dialogue. The robot can also be taught to assist humans by performing different tasks such as learning and recalling objects. It would be worthwhile to investigate if such a socialising capability would make it interesting and inviting for the users to interact with the robot in various learning scenarios.

Current work investigates the impact of such an interaction ability on the perceived intelligence, social acceptance, and likeability of the robot. Preliminary results suggest that the users perceive such a robot as more intelligent and likeable than a regular mechanical button-driven robot, rating it higher in terms of its ability to engage them in a conversation.

Acknowledgement

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169), the European Union under project SECURE (No 642667), and the Hamburg Landesforschungsförderungsprojekt CROSS.

References

[1] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The Cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, C. L. Nehaniv, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 52–87.

[2] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[3] C. L. Breazeal, "Sociable machines: Expressive social exchange between humans and robots," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[4] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "NICO – Neuro-Inspired COmpanion: A Developmental Humanoid Robot Platform for Multimodal Interaction," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Lisbon, Portugal: IEEE, 2017, to appear.

[5] G. Trajkovski and S. G. Collins, *Handbook of Research on Agent-Based Societies: Social and Cultural Interactions*, 1st ed. IGI Global, 2009.

[6] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 177 – 190, 2003, Socially Interactive Robots.

[7] M. Mori, K. F. MacDorman, and N. Kageki, "The Uncanny Valley," *IEEE Robotics Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[8] X. Hinaut, J. Twiefel, M. B. Soares, P. Barros, L. Mici, and S. Wermter, "Humanoidly speaking–learning about the world and language with a humanoid friendly robot." *International Joint Conference on Artificial Intelligence Video competition*, 2015.

[9] J. Twiefel, X. Hinaut, M. Borghetti, E. Strahl, and S. Wermter, "Using Natural Language Feedback in a Neuro-Inspired Integrated Multimodal Robotic Architecture," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. New York, NY, USA: IEEE, 2016, pp. 52–57.

[10] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. Kauai, Hawaii, USA: IEEE, 2001, pp. 511–518.

[11] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[12] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," in *European Conference on Computer Vision (ECCV)*, ser. LNCS, vol. 3021. Prague, Czech Republic: Springer, Berlin, Heidelberg, 2004, pp. 469–481.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, 2015, pp. 815–823.

[14] S. Dieleman and B. Schrauwen, "End-to-End Learning for Music Audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 6964–6968.

[15] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker Identification and Clustering Using Convolutional Neural Networks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Salerno, Italy: IEEE, 2016, pp. 1–6.

[16] X. Zhao and D. Wang, "Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, BC, Canada: IEEE, 2013, pp. 7204–7208.

[17] J. Twiefel, T. Baumann, S. Heinrich, and S. Wermter, "Improving Domain-Independent Cloud-Based Speech Recognition with Domain-Dependent Phonetic Post-Processing," in *AAAI Conference on Artificial Intelligence*, vol. Twenty-Eighth. Québec City, Québec, Canada: AAAI Press, 2014, pp. 1529–1535.

[18] E. Reiter and R. Dale, "Building Applied Natural Language Generation Systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.

[19] K. van Deemter, M. Theune, and E. Krahmer, "Real Versus Template-Based Natural Language Generation: A False Opposition?" *Computational Linguistics*, vol. 31, no. 1, pp. 15–24, 2005.

[2] `http://wiki.ros.org/smach` [Accessed: 16.03.2017]