In Proceedings of the 5th International Conference on Human Agent Interaction HAI '17, October, 2017, Bielefeld, Germany, pp. 171–180, (Publisher: ACM, New York, NY, USA).

The Impact of Personalisation on Human-Robot Interaction in Learning Scenarios

Nikhil Churamani, Paul Anton, Marc Brügger, Erik Fließwasser, Thomas Hummel, Julius Mayer, Waleed Mustafa, Hwei Geok Ng, 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

Knowledge Technology, Department of Informatics, Universität Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
5churama@informatik.uni-hamburg.de

ABSTRACT

Advancements in Human-Robot Interaction involve robots being more responsive and adaptive to the human user they are interacting with. For example, robots model a personalised dialogue with humans, adapting the conversation to accommodate the user's preferences in order to allow natural interactions. This study investigates the impact of such personalised interaction capabilities of a human companion robot on its social acceptance, perceived intelligence and likeability in a human-robot interaction scenario. In order to measure this impact, the study makes use of an object learning scenario where the user teaches different objects to the robot using natural language. An interaction module is built on top of the learning scenario which engages the user in a personalised conversation before teaching the robot to recognise different objects. The two systems, i.e. with and without the interaction module, are compared with respect to how different users rate the robot on its intelligence and sociability. Although the system equipped with personalised interaction capabilities is rated lower on social acceptance, it is perceived as more intelligent and likeable by the users.

ACM Classification Keywords

I.2.9 Robotics; I.5.5 Implementation: Interactive systems; I.2.11 Distributed Artificial Intelligence: Intelligent agents;

Author Keywords

human-robot interaction; social robotics; companion robots; personalisation; personalised robots; person identification; person localisation; speech processing; natural language understanding; dialogue management;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HAI '17, October 17–20, 2017, Bielefeld, Germany

© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-5113-3/17/10…$15.00

DOI: https://doi.org/10.1145/3125739.3125756

INTRODUCTION

As research in Human-Robot Interaction (HRI) is advancing, more and more agents are being equipped with enhanced interaction capabilities, allowing them to operate well in human-centred environments. Companion robots, in particular, are designed to possess multi-modal perception capabilities allowing them to pick up verbal as well as non-verbal cues while interacting with humans. One thing which can be seen clearly is that in order to develop agents that work well in a human environment, pure robotic and computing capabilities are not enough [9]. Such robots need to interact as naturally as possible with humans to blend well with their environments also taking into consideration the psychological aspects of an interaction [8].

Companion robots that operate in home environments such as working with children or the elderly should possess the capability to understand the environment and the users to adapt their behaviour accordingly. This makes the robots more aware of their surroundings [8]. A robot that is able to differentiate between different users and remember their preferences while interacting with them is also expected to be perceived as intelligent and likeable by the users [14]. Such robots need to hold a two-way communication with the user. They need to constantly update their knowledge, adding information about the user and the environment in which they operate and use this information to modify their responses.

This study aims to measure the impact of such personalised interaction capabilities on the perceived intelligence and likeability of a human companion robot in learning scenarios. Also, it investigates whether improving the qualitative experience of the user during interaction results in a higher evaluation of the robot's overall competence in a particular task.

To achieve these objectives, the study compares two conditions of the same system in which users are instructed to teach, using natural language, the Neuro-Inspired Companion (NICO) [22] robot (Figure 1) to recognise and recall different objects. Both the conditions realise the Humanoidly Speaking [18, 39] scenario for object learning. While the first condition implements only the learning scenario on NICO, the second one extends

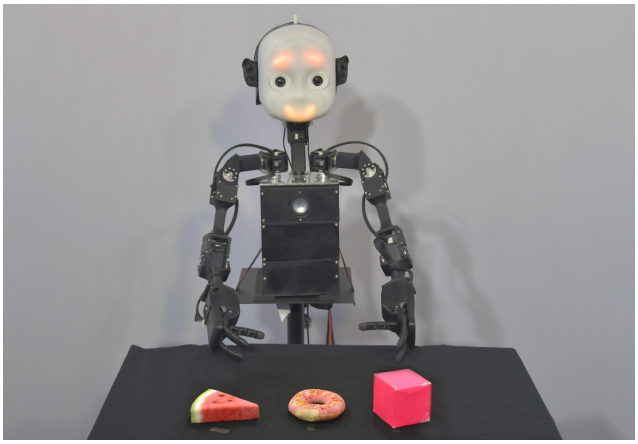


Figure 1: Neuro - Inspired COmpanion Robot (NICO) in the object learning scenario.

it with an interaction module that allows the users to engage in a personalised conversation with the robot where the robot asks them about their name and personal preferences and recalls them when it sees the users again. Both conditions are compared on the basis of how different users rate the robot on its intelligence and sociability.

RELATED WORK

Over the years in HRI research, personalisation has been realised in three different contexts; allowing users to customise the "look-and-feel" of robots [23], designing the robots to increase their friendliness or social acceptance [34], or allowing to develop long-term relationships between the users and the robot [24]. This study focusses on making the users' experience of interacting with the robot pleasant and enriching. Personalisation is achieved by making the robot sensitive to the users, recognising them and using the information shared by them to model a conversation. Since the robot also remembers different users and their preferences, the robot is expected to give an impression of providing special attention and personal recognition to the users.

This study uses the multi-modal sensing capabilities of the NICO robot to realise interactions with the user. The robot is able to use its vision capabilities to identify and engage the user who interacts with the robot using natural language. For the learning scenario, a modified implementation of the Humanoidly Speaking [18, 39] object learning scenario is used, allowing the user to teach NICO to recognise and recall different objects.

The NICO Robot

NICO, the Neuro-Inspired COmpanion [22] (Figure 1), is a teen-sized developmental robot designed and built for neurocognitive research. NICO is equipped with two cameras, modelled as NICO's eyes, which are used to capture the visual scene. The LED markers on its face help in modelling facial expressions which can be used to articulate emotional concepts in a conversation. It can also use its robotic arms for gesticulation and object manipulation. For the study, it

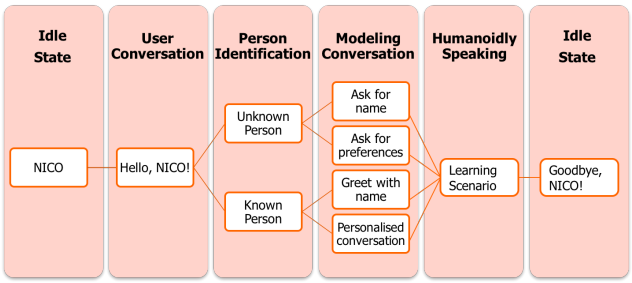


Figure 2: Overview of the Interactive Scenario.

was also equipped with an external microphone to accurately capture the voice of the participant.

Humanoidly Speaking Scenario

The Humanoidly Speaking [18, 39] scenario, originally implemented on the NAO robot, was modified for this study and implemented on the NICO robot. It involves an object learning scenario where users are able to teach objects to the robot using natural language instructions. Objects are placed on a table in front of the robot (See Figure 1) and referred to using their relative positions (left, middle, right) on the table, with respect to NICO. For example, *"The doughnut is in the middle."*

The visual scene of the table is captured and the objects are segmented into three corresponding regions. These segmented images are then fed to a Convolutional Neural Network (CNN) trained to classify different objects. No labels are assigned to the objects, yet. Users are able to assign labels to the initially unknown objects using natural language. Once the objects are taught, NICO is asked to recall and point to these objects using its robotic arms. This allows the system to be scalable as the robot can be trained to identify and recall any object by retraining the CNN classifier.

MODELLING THE INTERACTION SCENARIO

The Humanoidly Speaking scenario acts as the learning scenario for both conditions. In the interaction scenario, the learning is equipped with an interaction module that allows for a natural and human-like interaction with the user. The objective of the interaction module is to engage and motivate the user to teach different objects to NICO. The interaction flow can be seen in Figure 2. The users interact with NICO using natural language, which then models a conversation with them asking for their name and personal preferences. When the interaction module is triggered, NICO attempts to locate, identify and track the users and models a conversation using a state-based dialogue manager.

The different components entailed in realising the interaction scenario (Figure 3) are devised as ROS[1] nodes which interact with each other as well as internally using ROS messages and services. Different components of the interaction model [31] are explained in the following sections.

[1] http://wiki.ros.org [Accessed 16.04.2017]

Vision

Humans use their vision capabilities to look for objects of interest in their visual field such as faces or objects to manipulate. Also, during human-human interactions, recognising other individuals and looking at them allows humans to engage in a more meaningful conversation [7, 11]. Thus, while modelling an interaction with humans, a companion robot should also make use of such capabilities to interact with the users. Detecting and recognising the users and tracking their faces during an interaction would allow the robot to appear interested and engaged in the conversation, evoking a similar response from the user.

Face Detection and Tracking

One of the most common approaches for face detection is applying Haar-like feature-based cascades [42] to locate the face in the visual frame. The vision system [31] attempts to improve the performance of such an approach by augmenting it with template matching. Once a face is found, the algorithm looks only in a specific region of interest determined by the centre of the last detected face. Thus, the area in which the algorithm looks for a face is reduced, speeding up the process of detection with the Haar cascades. For face detection and tracking, a Convolutional Neural Network (CNN) based approach [3] for localising the most prominent face in the visual frame was also attempted but not used due to performance and latency issues.

Haar-like feature-based face detection combined with template matching results in a robust face detection algorithm which provides a strong basis for an effective face tracking. The algorithm tries to track the detected face by keeping it at the centre of the visual frame by rotating NICO's head accordingly. Also, in order to further underline NICO's awareness of the user, it displays a "Happy" emotion using its face LEDs while detecting and tracking a face.

Face Recognition

Recognising the users and differentiating between them becomes a necessary task for a robot aiming to model an interaction with humans. Face recognition is an active field of research [1, 6, 35]. For this study, Local Binary Pattern Histograms (LBPH) [1] and CNN based approaches [35] were tested. Although both approaches gave accurate results, the CNN-based approach turned out to be resource and computation intensive needing at least one second of processing time per frame. Thus, the LBPH-based approach was used for the user study. For face recognition, the classifier was trained using the data set of face images collected from project members and the participants as known faces and the Extended Cohn-Kanade data set [28] for modelling unknown faces.

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 corresponds to a high confidence for one of the known faces while an unknown person yields low confidences for all the known faces. In case a new person is encountered, data for that person is recorded and the classifier is trained again.

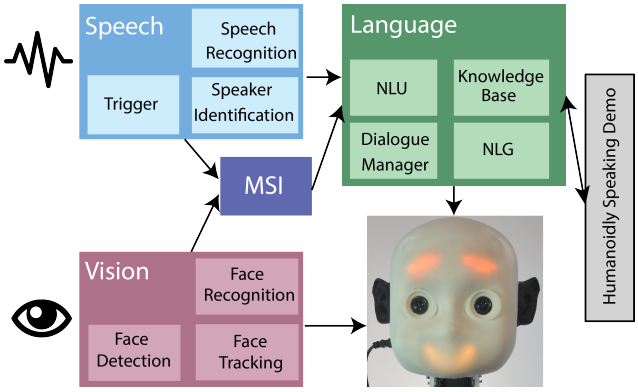


Figure 3: System Architecture for the Interactive Module.

Speech

Auditory information processing is an essential part of the system, as communication with NICO is realised using natural language. Results from speech recognition are used as input for the language module. Apart from generating text from speech, the speech signal is also used to identify the user speaking with NICO (see Figure 4). Speaker identification allows to recognise and remember persons solely based on their speech characteristics. Furthermore, a trigger is implemented that allows to initiate and end a conversation with NICO in a natural way using "*Hello, NICO!*" and "*Goodbye, NICO!*" respectively as cue-phrases.

Speech Recognition

Speech recognition is implemented using the DOCKS framework [38] which combines domain-specific knowledge with predictions from Google's cloud-based Automatic Speech Recognition (ASR). This improves the performance of natural language understanding by omitting out-of-domain words. Despite the fact that small and specific language models (LM) result in only limited interaction capabilities, the system [31] combines a small LM containing the cue-phrases and a large LM modelling the full interaction in a cascading manner.

Speaker Identification

One of the most commonly used approaches for speaker identification is using Gaussian Mixture Models (GMMs) on Mel-Frequency Cepstral Coefficients (MFCCs) extracted from the speech-audio signal [26, 30]. Although applying a GMM implementation to a clean training data set yielded good results, problems occurred when facing noisy conditions during live application. MFCCs are known to be sensitive to noise, leading to decreased performance in noisy environments [44]. Using a CNN on Mel-Spectrograms [31] derived from speech audio revealed to be a better choice for the proposed system in terms of accuracy and robustness. As CNNs can be used to learn relevant features directly from the audio signal [15, 29], they avoid various pre-processing steps and the selection of engineered features such as MFCCs. For this implementation, the CNN was trained on a speech data set consisting of samples from project members and the participants.

When the classifier confidence was not high enough (empirically defined threshold), the speaker was classified as unknown.

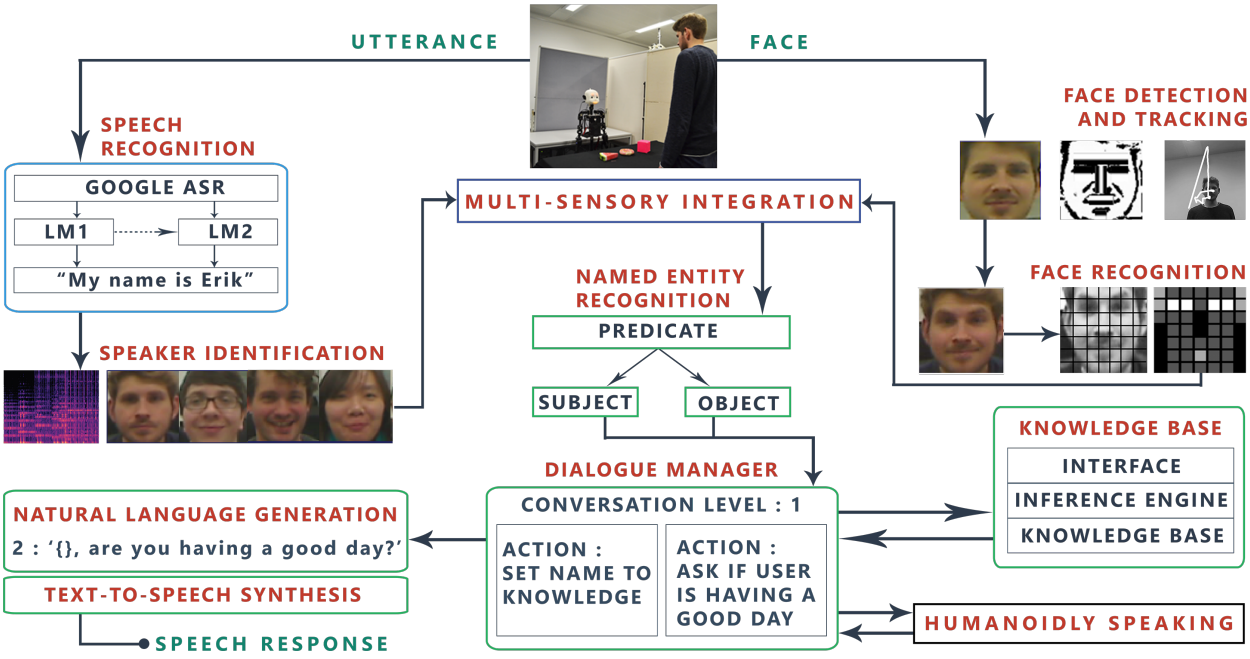


Figure 4: Process overview for the Interaction Module

In such a scenario, the speech recordings from the unknown speaker, together with known speakers, were used to retrain the network.

Multi-Sensory Integration

Making use of different modalities can improve robustness [4] and introduce redundancy which helps to cope with scenarios where only limited information is available (for example, the user is speaking but not in the visual field). The Multi-Sensory Integration component [31] combines predictions from Face Recognition and Speaker Identification using a weighted average of both predictions, generating the final prediction from the system. A higher weight is assigned to face recognition due to the short duration of the sound sample available for classification at the beginning of the conversation as well as the noisy environment.

Language

During an interaction, the robot needs to understand the information provided by the user and to attribute this information in the robot's model of the world. The language module [31] allows NICO to extract this information about the environment and the user through user utterances and formulate appropriate responses. Furthermore, it is also responsible for updating changes in the model about the world as new information is obtained. The language module consists of four components (see Figure 4) namely, Natural Language Understanding - Named Entity Recognition, Dialogue Management, Knowledge Base and Natural Language Generation.

Natural Language Understanding

The Natural Language Understanding (NLU) component extracts relevant semantic information from human utterances and sends this tagged information to the Dialogue Manager (DM). Named Entity Recognition (NER) extracts relevant information (predicates) from input sentences assigning them to different information slots. For example, for the sentence:

"My name is Peter and I am from Germany"

It creates information slots such as `['Peter' PER, 'Germany' LOC]` and extracts `Name(person_id, 'Peter')` and `From(person_id, 'Germany')` as predicates. In order to enhance the learning capabilities of NICO, the sub-components of NER which include a POS-tagger and an NP Chunker, are designed to automatically learn from the existing corpora. For this study, a parametric [12] and a non-parametric model [27] were compared and both proved capable of providing learning capabilities without any additional pre-processing steps.

Dialogue Manager

The Dialogue Manager (DM) can be understood as a decision maker that simulates the action selection process to decide what information the robot should seek and how to ask for it. It then sends the corresponding Action ID to the Natural Language Generation (NLG) component. If necessary, information from the data repositories (Knowledge Base) is obtained and a decision on how to respond to the user is made. SMACH[2] is used to implement a finite-state dialogue manager, where each response is a state in the state machine. A total of 32 states were realised for the interaction module. Some examples of state transitions can be seen in Table 1.

Knowledge Base

The purpose of the Knowledge Base (KB) component is twofold. Firstly, it provides a way of modelling pre-existing

[2] http://wiki.ros.org/smach [Accessed: 17.04.2017]

User Utterance	Input Received	Executed State	Action Performed	Transition	Output Utterance
“Hello, Nico!”	ConvLvl: 0 PID: 0 Pred:{‘hello’:‘’}	Hello	If PID is unknown, set AID=2	Set ConvLvl = 1	AID:2; ‘Hello, what is your name?’
‘My name is Erik.’	ConvLvl: 1 PID: 0 Pred:{‘PER’:‘Erik’}	Name	Save name to KB Get PID: 1 (Known person) Set AID = 5	Set ConvLvl = 2	AID: 5; ‘That’s a good name, {Erik}. I will remember it. Are you having a good day?
...	...	...	...	...	...
“Goodbye, Nico!”	ConvLvl: 42 PID: 1 Pred:{‘goodbye’:‘’}	Demo-1	If goodbye, set AID=1	Set ConvLvl = 0	AID:1; ‘OK, {Erik}, I really enjoyed our conversation today.’

Table 1: State transition examples. (***ConvLvl:*** *Conversation level,* ***PID:*** *Person ID,* ***Pred:*** *Predicate(INT, PER),* ***AID:*** *Action ID,* ***KB:*** *Knowledge Base*)

knowledge as well as to incorporate knowledge acquired by the DM during the conversation with the user. Secondly, it provides the link from this knowledge to other components, thus offering support in modelling the structure of the dialogue. The structure of the knowledge base [16] corresponds to a typical rule-based system and consists of three main components, namely a database, an inference engine and an interface to the inference engine. The database contains knowledge as a list of rules. The inference engine processes the information based on the existing knowledge in the rule base and the interface retrieves the relevant information based on custom requests.

Natural Language Generation

Natural Language Generation produces a text representation of what the robot should say and utters the sentence through Text-to-Speech synthesis. NLG has the purpose of turning an action ID (plus provided parameters) into a sentence understandable for humans. This is realised using a template-based approach [33, 40]. In this approach, template sentences (with blanks) for each action are provided which are then filled out with the provided information. There are 114 sentences which are mapped with the respective action IDs created for this component. When an action ID from the DM is received, the corresponding sentence will be uttered using Google’s text-to-speech engine (gTTS [3] ver. 1.1.8).

EXPERIMENTS

To evaluate the impact of the added interaction module, an experiment was designed to investigate whether participants experience a qualitative difference when NICO indulges in a personalised interaction with them in comparison to the baseline scenario. A user study was carried out in order to assess the perceived intelligence and sociability of the robot using specialised questionnaires, which the participants filled in after having completed the experiment.

User Study

The user study consisted of 27 participants (18 male, 9 female) in the age group of 18 to 35 years. These participants were recruited amongst students and employees at the university. The participants were asked to fill in details about their prior experience working with robots which ranged from no experience at all to high experience, with no majority in either group. Participants reported having average to fluent spoken English skills. This was important as the experiments were conducted in English.

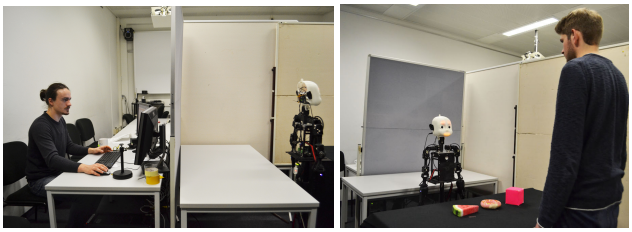
After obtaining an informed consent on data privacy measures and the general procedure of the study, participants were split randomly into two discrete groups. The first group (13 participants) was assigned to the baseline condition (Humanoidly Speaking scenario without personalised interaction capabilities) while the second group (14 participants) was assigned to the interactive condition. The assignment of participants to these respective groups was done by the operator who also ensured that the system is running and could be shut down immediately in case of any contingency. Both the experimenters, who guided participants throughout the experiment, as well as the participants themselves, were blinded from the user group assignment. This double-blinded design was chosen to reduce undesired effects of biases, i.e. the participants’ bias from knowing which condition they were assigned to or the experimenters biasing the participants based on their knowledge about the system. The experimenters, therefore, followed the same protocol while introducing the participants to either of the systems.

The task to be performed by the participants consisted of two interactions. In the first interaction, participants were instructed to teach NICO objects by referring to their positions on the table. After having taught objects, in the second interaction, participants were asked to evaluate if NICO could recall all the learnt objects. Each interaction lasted for about 15 minutes (with at least 10 turns) with a 10 minutes break in between.

The main difference between the baseline condition and the interaction-enabled condition was based on how NICO interacted with the users. For the baseline condition, participants were expected to directly teach or make NICO recall objects, whereas the interactive condition involved a personalised interaction prior to object learning or recalling. As the users had never interacted with NICO, they were not recognised in the first interaction and personal information, such as their name, nationality etc. was sought. NICO then learnt to recognise the participants based on their appearance as well as their way of speaking and remembered the personal information obtained, which could later be recalled in the second interaction.

After completing both interactions, the participants were asked to fill out a questionnaire recording their evaluation of the system. Only after the participants had filled in the questionnaire, they were informed about their group assignment.

[3] https://pypi.python.org/pypi/gTTS [Accessed: 19.04.2017]



(a) The operator monitoring the system. (b) User interacting with NICO.

Figure 5: Experiment Set-up.

Experiment Set-up

The experiment set-up consisted of an artificially well-lit room in order to exclude effects of changing natural lighting conditions. Three toy objects (a melon, a doughnut and a box) were placed on a table covered with a black sheet simplifying object segmentation and recognition. During the interaction, the participants were instructed to stand approximately one meter in front of NICO. The operators were positioned behind an artificial wall separating them from the participants in order to avoid any communication between the two during the experiment. The above-explained experiment set-up can be seen in Figure 5 showing the operator as well as a participant interacting with NICO. Once the participants had finished interacting with NICO, the questionnaire was presented in an electronic form and the data obtained was pseudonymised and stored in the database. Any video or audio data recorded during the user study was deleted after the experiment.

Questionnaire

To compare the performance between the two conditions, the devised questionnaire consisted of three parts: The GODSPEED test [5], questions based on the UTAUT model [41] and some additional questions (MISC) measuring the performance of the interaction module (see Table 2). All questions were presented in a random order to the participants with some questions also scale-inverted to avoid any reporting bias. All questions were based on a 5-point Likert scale [25].

- GODSPEED: The GODSPEED test [5] is a widely used test to measure how participants perceive a robot. It measures several attributes including perceived intelligence which was a focus of this study. More specifically, it measures *Anthropomorphism, Animacy, Likeability, Perceived Intelligence* and *Perceived Safety*.

Question	Negative answer	Positive answer
How much does the robot pay attention to you?	Not much	A lot
Did the robot keep you engaged by paying attention at you while talking?	Not at all	Always
How likely does the robot remember you?	Very unlikely	Very likely
How likely will the robot remember you next time?	Very unlikely	Very likely
How confident was the system in remembering you?	Not confident	Very confident

Table 2: MISC questions.

- UTAUT: The Unified Theory of Acceptance and Use of Technology (UTAUT) [41] is a model for evaluating the acceptance of the technology. It is based on four key aspects, namely *Performance Expectancy, Effort Expectancy, Social Influence* and *Facilitating Conditions*. A customised version of the UTAUT questionnaire was used in the experiment. In addition, positive and negative variants for the same questions were randomly chosen.
- MISC: Focussing on attributes specific to the interaction enabled system, some miscellaneous (MISC) questions were created concerning NICO's ability to be attentive and to remember the participants.

Results

As the two conditions were evaluated by different sets of participants, the choice of statistical analysis for significance was limited to two independent sample tests. In order to evaluate the ordinal responses in Likert scale from a small number of participants, an appropriate test should be able to handle the multivariate ordinal aspect of the collected data and have enough power to detect a statistical difference with small sample sizes. While the debate over whether ordinal scales could be viewed as interval scales has been going on [20], recent studies in medical research have heavily criticised approaches using parametric statistical tests on ordinal data [19, 20]. Other studies have suggested that parametric tests on ordinal data should only be used as a pilot analysis under normality assumption [2, 10].

Thus, a multivariate variant of the Mann-Whitney U test was chosen for this study to ensure the validity of the result. The test was carried out based on an implementation from the SpatialNP R-Package [36]. The continuous assumption of the data is only considered for trend analysis to support investigating the direction of changes as the multivariate tests do not explicitly provide information about trends.

Furthermore, post hoc analysis on each individual question was performed using the Wilcoxon rank-sum test on two independent samples from the Hotelling R-package [13]. This choice of tests also influences the way missing data was handled. In contrast to replacing missing data with mean values [17], since all questions in GODSPEED and UTAUT questionnaires are considered as multidimensional variables, all questions in one dimension are excluded if one of their counterparts contains missing data.

For the GODSPEED questionnaire, all the five aspects measuring the perception of the robot were evaluated (Figure 6). The standardised statistics Q^2 and Degree of Freedom (DoF) for all tests on five dimensions are reported in Table 3 with significant results highlighted. Significant changes in the way participants perceived the robot were detected in two out of five dimensions: *Likeability* ($Q^2 \approx 12.7, p\text{-}value = 0.026$) and *Safety* ($Q^2 \approx 9.1, p\text{-}value = 0.027$) as well as some evidence for *Intelligence* ($Q^2 \approx 9.85, p\text{-}value = 0.079$).

In this multivariate setting, the test statistics could not determine the direction of influence and thus, post hoc analysis had to be carried out on each of the questions [32]. The trend analysis (Figure 6) suggests that the interaction-enabled

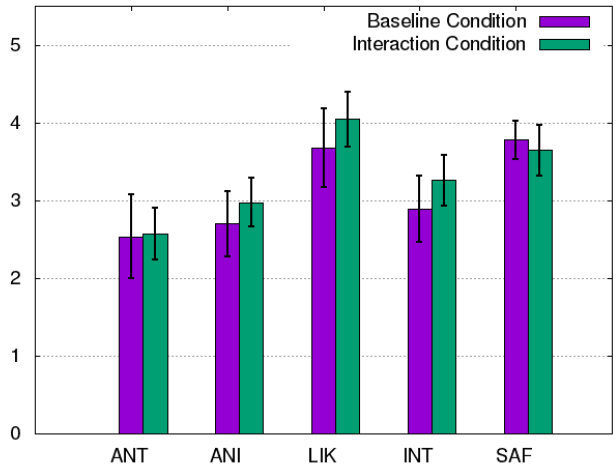


Figure 6: Comparison of average values and trend analysis for the GODSPEED Questionnaire including 95% confidence interval.
ANT: Anthropomorphism, ANI: Animacy, LIK: Likeability, INT: Perceived Intelligence, SAF: Perceived Safety

GODSPEED Dimension	Q^2	DoF	*p-value*
Anthropomorphism	8.9606	5	0.110
Animacy	6.0206	6	0.421
Likeability	**12.7130**	**5**	**0.026**
Intelligence	9.8482	5	0.079
Safety	**9.1397**	**3**	**0.027**

Table 3: Test results for the Mann-Whitney U test on the GODSPEED Questionnaire.

system performed better. Additionally, univariate, one-sided Wilcoxon-Mann-Whitney tests on each individual question for Likeability and Perceived Intelligence confirm this observation. Under the alternative hypothesis that the interaction-enabled system performs better, significant changes were reported in two individual questions for Likeability ($U = 46.5$, *p-value* ≈ 0.04 & $U = 39$, *p-value* ≈ 0.01) and also two questions for Perceived Intelligence ($U = 52$, *p-value* ≈ 0.08 & $U = 49$, *p-value* ≈ 0.05). For Perceived Safety, while significant changes could only be found on combinations of questions and not on every individual question, a strong implication can be derived ($U = 102$, *p-value* ≈ 0.16) under the alternative hypothesis that the interactive system performs worse. This suggests that the baseline system seemed safer to use to the participants.

UTAUT Dimension	Q^2	DoF	*p-value*
Performance Expectancy	4.4579	5	0.485
Effort Expectancy	5.9265	5	0.313
Social Acceptance	**13.4710**	**5**	**0.019**
Facilitating Condition	4.3774	5	0.496

Table 4: Test results for the Mann-Whitney U test on the UTAUT Questionnaire.

The trend analysis on UTAUT evaluations (Figure 7) reveals a higher rating for the baseline system than the interactive one.

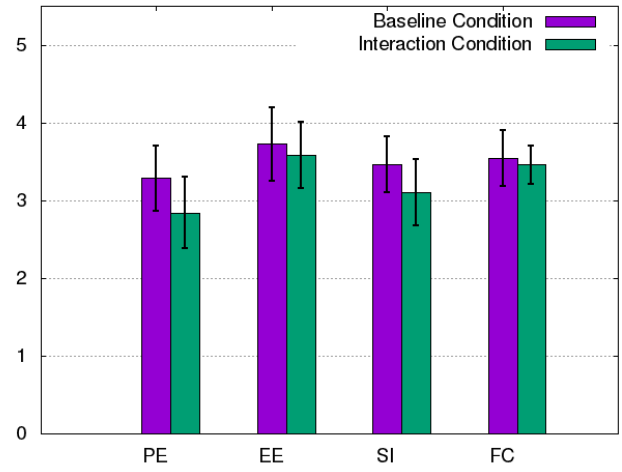


Figure 7: Comparison of average values and trend analysis for the UTAUT Questionnaire including 95% confidence interval.
PE: Performance Expectancy, EE: Effort Expectancy, SI: Social Influence, FC: Facilitating Conditions

Test statistics for each dimension of UTAUT evaluation are reported in Table 4 with one significant difference detected on the Social Influence dimension. Post hoc analysis for each individual question in Social Influence under the alternative hypothesis that the interactive system has less influence on the participant, results in evidence for three out of five possible changes ($U = 124$, *p-value* ≈ 0.05, $U = 122$, *p-value* ≈ 0.06 & $U = 125.5$, *p-value* $= 0.04$).

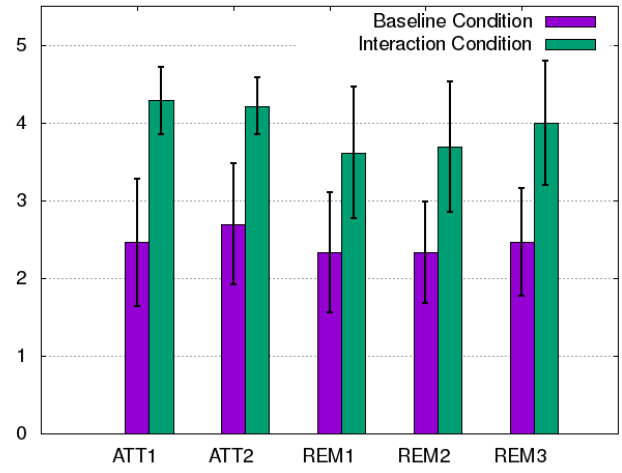


Figure 8: Comparison of average values and trend analysis for the MISC Questionnaire including 95% confidence interval.
ATT1: Pay Attention, ATT2: Keep Engaged, REM1: Remember, REM2: Remember Next Time, REM3: Remembering Confidence

The same analysis was done for the MISC questionnaire, although for this questionnaire, only the univariate Wilcoxon test was carried out since each dimension of MISC contained only one question. The results (Table 5) validated the pilot analysis with all the measured attributes of the interactive system rated higher by the participants (Figure 8). All measurements

MISC Dimension	U	*p-value*
Pay Attention	**33.5**	**0.002**
Keep Engaged	**35.5**	**0.002**
Remember	**42.5**	**0.025**
Remember Next Time	**37.5**	**0.013**
Confidence remembering	**33.0**	**0.003**

Table 5: Test results for the univarite one-sided Wilcoxon-Mann-Whitney U test on the MISC Questionnaire. The alternative hypothesis being the new system is better.

achieved statistical significance and hint that the system with the interaction module met all functional requirements.

DISCUSSION

The primary objective of this study is to measure how an agent's capability to model personalised interactions affects its acceptability and perceived intelligence for humans. The interactive system is assigned two principal tasks: interacting with users gathering information about them and transitioning its capabilities to the object learning baseline system. Both functional and perceptual changes of the interactive system are evaluated based on the responses from users on three questionnaires. GODSPEED measures the users' perception of the robotic system, UTAUT evaluates the users' acceptance of technology, and MISC consists of five additional questions measuring the functional aspects of the system. Overall, results from all questionnaires reflect significant differences in the users' awareness towards the two systems. The realisation of functional aspects in all measured dimensions of MISC indicates a major improvement of the interaction-equipped system in keeping the users engaged and remembering their information across interaction sessions. These functional changes, in turn, lead to perceptual changes in the users which are reflected in the results of GODSPEED and UTAUT questionnaires. For GODSPEED, evidence for *Likeability* and *Intelligence* imply a positive development in humans' perception of the object learning scenario without having to improve any aspect of the learning task itself. However, the interactive system also appears to be less safe as indicated by the shift in *Perceived Safety* in favour of the baseline system. This observation is reflected in both the direction of the tests and the trend analysis (Figure 6). This can be attributed to the fact that the interactive system, in essence, models much more complex interactions with the user which could evoke some reservations in the mind of the user [21].

For UTAUT, both statistical tests and trend analysis detect a drop in the evaluation of Social Influence. An explanation for this could be the Acquiescence Bias effect [43] which would mean that the users who were assigned to the baseline system had very limited interaction with the robot and thus, they might have been biased to give a high rating to the baseline system. Also, even though the participants rate the interactive system as more intelligent and likeable, they would not recommend using it in a real-world context. This could be due to the longer and more complex interactive sessions and the noisy environment. This can be compared to a similar situation [17] where the quality of dialogues of the system was considerably affected by the speech recognition module and users had to repeat their responses on multiple occasions. Consequently, the whole system was considered undesirable to operate in a real-life context. Another aspect that could account for the low acceptance on Social Influence is the domain of conversations. Different states of the conversation are detected by the Named Entity Recognition component, which is in turn built and evaluated based on the training data set of the Conference on Natural Language Learning (CoNLL) [37]. While this approach enables an automatic method to measure the isolated module to the state-of-the-art approach, its domain is limited to the three main categories of the CoNLL challenges: Person (PER), Organisation (ORG) and Location (LOC). In order to slightly increase the range of possible topics, two more categories have been added (food and drinks), given that the toy objects to be learnt included a melon and a doughnut. However, the total number of possible topics is still considerably low.

In general, this study endorses the reliability of GODSPEED and UTAUT questionnaires for similar studies measuring the social impact of robotic systems. A homogeneity in statistical results of all questions in the same dimension is witnessed: no two questions in the same significant dimension reflect different shift directions. More importantly, the fact that multivariate Mann-Whitney U tests are able to detect significances and confirm the trend analyses, indicates that multivariate non-parametric tests should take higher priority in similar analyses.

CONCLUSION

This study described an experiment to investigate the impact of personalised interaction capabilities of the NICO robot on how the users perceive it in terms of intelligence and sociability. A baseline system which is capable of learning objects through natural language interaction was compared to a system with added interaction capabilities [31]. The results from the user study show that participants perceived NICO to be more intelligent and likeable when it involved them in personalised conversations. This is an important attribute to possess for agents that are designed to work in close human proximity, particularly in home scenarios working with children or the elderly (see Section 1 - 2). It would allow the robots to be accepted as an integral part of the home, they are designed to serve. Despite the overwhelmingly positive impressions of the interactive NICO, it is perceived to be less safe and perhaps because of this reason, has a weaker social influence on the participants. It would be important to further investigate these factors in order to improve the interaction capabilities of the robot although the study tries to explain some of the possible causes. Overall, the study presents evidence for improved impressions of the robot for the users when it is able to model personalised and engaging conversations with them.

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. The authors also thank Sven Magg for the guidance in experimental setup and analysis as well as Matthias Kerzel for the support with the NICO robot.

REFERENCES

1. Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. 2004. Face Recognition with Local Binary Patterns. In *European Conference on Computer Vision (ECCV) (LNCS)*, Vol. 3021. Springer, Berlin, Heidelberg, Prague, Czech Republic, 469–481.
2. I. Elaine Allen and Christopher A. Seaman. 2007. Likert Scales and Data Analyses. *Quality Progress* 40, 7 (2007), 64–65.
3. Pablo Barros, German I. Parisi, Cornelius Weber, and Stefan Wermter. 2017. Emotion-Modulated Attention Improves Expression Recognition: A Deep Learning Model. *Neurocomputing* 253 (2017), 104–114.
4. Pablo Barros and Stefan Wermter. 2016. Developing Crossmodal Expression Recognition Based on a Deep Neural Model. *Adaptive Behavior* 24, 5 (2016), 373–396.
5. Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2008. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2008), 71–81.
6. Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. 1997. Eigenfaces Vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7 (1997), 711–720.
7. Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter F. Dominey, and Jocelyne Ventre-Dominey. 2012. I Reach Faster When I See You Look: Gaze Effects in Human–Human and Human–Robot Face-to-Face Cooperation. *Frontiers in Neurorobotics* 6, 3 (2012).
8. Cynthia Lynn Breazeal. 2000. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Dissertation. Massachusetts Institute of Technology.
9. Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. 1999. The Cog Project: Building a Humanoid Robot. In *Computation for Metaphors, Analogy, and Agents*. Number 1562 in LNCS. Springer Berlin Heidelberg, 52–87.
10. James Dean Brown. 2011. Likert Items and Scales of Measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter* 15, 1 (2011), 10–14.
11. Allison Bruce, Illah Nourbakhsh, and Reid Simmons. 2002. The Role of Expressiveness and Attention in Human-Robot Interaction. In *IEEE International Conference on Robotics and Automation (ICRA)*, Vol. 4. IEEE, Washington, DC, USA, 4138–4142.
12. Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.
13. James Curran. 2017. Hotelling: Hotelling's T^2 Test and Variants. (2017). `https://cran.r-project.org/web/packages/Hotelling/index.html`
14. Kerstin Dautenhahn. 1995. Getting to know each other – Artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems* 16, 2 (1995), 333 – 356. Moving the Frontiers between Robotics and Biology.
15. Sander Dieleman and Benjamin Schrauwen. 2014. End-to-End Learning for Music Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, 6964–6968.
16. Bruce Frederiksen. 2008. Applying expert system technology to code reuse with Pyke. *PyCon: Chicago* (2008).
17. Manuel Giuliani, Ronald P.A. Petrick, Mary Ellen Foster, Andre Gaschler, Amy Isard, Maria Pateraki, and Markos Sigalas. 2013. Comparing Task-Based and Socially Intelligent Behaviour in a Robot Bartender. In *ACM on International Conference on Multimodal Interaction (ICMI '13)*. ACM, Sydney, Australia, 263–270.
18. Xavier Hinaut, Johannes Twiefel, Marcelo Borghetti Soares, Pablo Barros, Luiza Mici, and Stefan Wermter. 2015. Humanoidly speaking–Learning about the world and language with a humanoid friendly robot. *International Joint Conference on Artificial Intelligence Video competition* (2015).
19. Ulf Jakobsson. 2004. Statistical Presentation and Analysis of Ordinal Data in Nursing Research. *Scandinavian Journal of Caring Sciences* 18, 4 (2004), 437–440.
20. Susan Jamieson. 2004. Likert Scales: How to (Ab)Use Them. *Medical Education* 38, 12 (2004), 1217–1218.
21. Frédéric Kaplan. 2004. Who is afraid of the Humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics* 01, 03 (2004), 465–480.
22. Matthias Kerzel, Erik Strahl, Sven Magg, Nicolás Navarro-Guerrero, Stefan Heinrich, and Stefan Wermter. 2017. NICO – Neuro-Inspired COmpanion: A Developmental Humanoid Robot Platform for Multimodal Interaction. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Lisbon, Portugal. In Press.
23. Rachel Kirby, Jodi Forlizzi, and Reid Simmons. 2010. Affective social robots. *Robotics and Autonomous Systems* 58, 3 (2010), 322 – 332. Towards Autonomous Robotic Systems 2009: Intelligent, Autonomous Robotics in the UK.

24. Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A longitudinal field experiment. In *7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 319–326.

25. Rensis Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* 22 (1932), 55.

26. Wenyong Lin. 2015. An Improved GMM-Based Clustering Algorithm for Efficient Speaker Identification. In *International Conference on Computer Science and Network Technology (ICCSNT)*, Vol. 1. IEEE, Harbin, China, 1490–1493.

27. Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1 (ETMTNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 63–70.

28. Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, San Francisco, CA USA, 94–101.

29. Yanick Lukic, Carlo Vogt, Oliver Dürr, and Thilo Stadelmann. 2016. Speaker Identification and Clustering Using Convolutional Neural Networks. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, Salerno, Italy, 1–6.

30. Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka. 2012. Speaker Identification and Verification by Combining MFCC and Phase Information. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 4 (2012), 1085–1095.

31. 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. 2017. Hey Robot, Why Don't You Talk to Me?. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Lisbon, Portugal. In Press.

32. Kathrin Pollmann. 2014. *Does the Human-Like Behavior of a Robot Evoke Action Co-Representation in a Human Co-Actor?* MSc Thesis. Technische Universiteit Eindhoven, Eindhoven, The Netherlands.

33. Ehud Reiter and Robert Dale. 1997. Building Applied Natural Language Generation Systems. *Natural Language Engineering* 3, 1 (1997), 57–87.

34. Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. 2016. "Teach Me - Show Me" End-User Personalization of a Smart Home and Companion Robot. *IEEE Transactions on Human-Machine Systems* 46, 1 (Feb 2016), 27–40.

35. Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 815–823.

36. Seija Sirkia, Jari Miettinen, Klaus Nordhausen, Hannu Oja, and Sara Taskinen. 2013. SpatialNP: Multivariate Nonparametric Methods Based on Spatial Signs and Ranks. (2013). `https://cran.r-project.org/web/packages/SpatialNP/index.html`

37. Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Conference on Natural Language Learning at HLT-NAACL (CONLL '03)*, Vol. 4. Association for Computational Linguistics, Edmonton, Alberta, Canada, 142–147.

38. Johannes Twiefel, Timo Baumann, Stefan Heinrich, and Stefan Wermter. 2014. Improving Domain-Independent Cloud-Based Speech Recognition with Domain-Dependent Phonetic Post-Processing. In *AAAI Conference on Artificial Intelligence*, Vol. Twenty-Eighth. AAAI Press, Québec City, Québec, Canada, 1529–1535.

39. Johannes Twiefel, Xavier Hinaut, Marcelo Borghetti, Erik Strahl, and Stefan Wermter. 2016. Using Natural Language Feedback in a Neuro-Inspired Integrated Multimodal Robotic Architecture. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York, NY, USA, 52–57.

40. Kees van Deemter, Mariët Theune, and Emiel Krahmer. 2005. Real Versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics* 31, 1 (2005), 15–24.

41. Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478.

42. Paul Viola and Michael Jones. 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, Kauai, Hawaii, USA, 511–518.

43. Dorothy Watson. 1992. Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness. *Sociological Methods & Research* 21, 1 (1992), 52–88.

44. Xiaojia Zhao and DeLiang Wang. 2013. Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Vancouver, BC, Canada, 7204–7208.