Copyright (c) 2020 IEEE. Personal use of this material is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

Voice recognition and processing interface for an interactive guide robot in an university scenario

Marvin Stuede*, Jonas Wilkening, Svenja Tappe and Tobias Ortmaier

Leibniz University Hannover, Institute of Mechatronic Systems
Hannover, 30167, Germany (Marvin.Stuede@imes.uni-hannover.de) * Corresponding author

Abstract: This paper presents a voice user interface consisting of several modules for a mobile service robot, which is used to guide people and provide information on a university campus. The recognition and processing system is based on cloud services to convert from speech to text and vice versa and a dialogue system to allow for natural interaction. An approach to combine these modules with a data management system for meal plan, public transit, and location information is presented. We evaluate the system in different environments, each with their individual reverberation times, proving the functionality under conditions typical for the intended use case. In a user study with 13 participants we show the usability of the system, by letting the participants freely interact with the robot. In 86 % of all cases the desired output can be achieved at least once per user and request. A questionnare shows that most users agree with a good usability of the system.

Keywords: Voice recognition, service robotics, natural language processing

1. INTRODUCTION

Intuitive and natural interaction with robotic systems has gained importance in recent years. Within the field of *social robotics*, various systems have been developed both in research and industry. These systems can be operated via touch or speech input and can therefore output targeted information and serve as an assistant or guide, if they are mobile. For example, the SPENCER project aimed to develop a service robot that could help and guide passengers at an airport [1]. Stricker *et al.* present an interactive mobile robot capable of speech synthesis to guide visitors in a university building [2]. Various works are concerned with controlling mobile robots via certain voice commands, eg. Poncela and Gallardo-Estrella who deal with moving a mobile robot, reading information or writing parameters via voice input [3].

In addition to specific commands, dialogue systems can also be integrated into service robots to enable a more natural form of interaction. Well-known examples for this are the robots Nao and Pepper from SoftBank Robotics, which have several microphones and a speech output and, in the case of Pepper, can recognize human emotions [4]. The robot Jibo (meanwhile discontinued) has a similar purpose as home assistant systems like Alexa or Google Home and is able to perform person-specific interactions [5].
This paper aims to implement such a dialog system functionality for a guiding and information-providing robot in a multi-variate scenario. The robot's objective is to guide and interact with visitors on a university campus. The campus consists of several buildings with multiple floors, containing offices, hallways and foyers and also an outdoor area. Typical for such an environment, some areas are very crowded, resulting in varying background noise level. Users of the system should be able to ask for directions to specific locations and staff's offices, as well as for the meal plan in the canteen and departure times of public transport. An implementation and evaluation of a voice recognition interface consisting of speech-to-text (STT) and text-to-speech (TTS) modules and a natural language processing (NLP) pipeline allowing for dialog based interaction is shown. In contrast to other publications, we evaluate the system in detail with respect to its purpose: in different environments with different reverberation times using several different speakers. In the course of a user study with 13 participants it is checked whether the system correctly processes formulations that have not specifically been provided before.

The remainder of this paper is structured as follows: Section 2 introduces the requirements for the implementation, the mobile robot and different modules of the voice recognition and processing system. Section 3 shows the performance in three different environments with and without ambient noise and the results of a user study in which participants could interact linguistically with the mobile robot. Section 4 summarizes the paper and gives an outlook on future work.

2. MATERIALS AND METHODS

2.1 Definition of requirements

As a basis for hardware selection and software design, we subdivide the interface requirements into three categories:
Scenario specific requirements: As the operational scenario is a university campus in Germany, a focus lies on correctly understanding proper and colloquial (german) names of institutions, facilities and persons. Furthermore, interfacing the university's data management and public transit systems is necessary to provide information about meals and public transport.
Usability specific requirements: There exist a variety of guidelines and principles for user interface design regarding usability, eg. Nielsen's usability heuristics [6]. Although already published in 1994, the heuristics such

This is the author's version of an article that has been published in the ICCAS 2020 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://dx.doi.org/10.23919/ICCAS47443.2019.8971465

Copyright (c) 2020 IEEE. Personal use of this material is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

as *visibility of system status* or *recognition rather than recall*, still apply today because of their broad practicability and are, therefore, taken into account for the interface presented in this work.

System specific requirements: The general structure of the system should be modular and comply with the setup shown in Fig. 1. Modularity allows for interchangeability of individual components, such as the speech-to-text (STT) and text-to-speech (TTS) engines, microphone hardware and natural language processing (NLP) system.

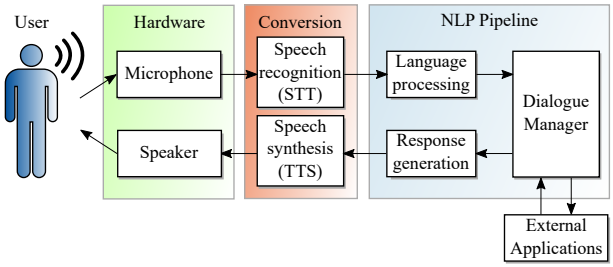

Fig. 1 Structure of the voice recognition interface and dialogue system. The software part is roughly divided between *Conversion* (STT and TTS) and the *NLP pipeline*, including the dialogue manager.

2.2 Mobile guide robot

The experiments are conducted on a mobile robot, equipped with sensors and differential drive to allow for indoor as well as outdoor localization and navigation (see Fig. 2). Localization is achieved by a combination of Visual SLAM and iterative closest point (ICP) with the RGBD-Cameras and 3D-Lidar based on the RTAB-Map method [7]. As a microphone, a four microphone array (ReSpeaker Mic Array v2.0) is used, allowing for on board signal processing techniques such as beamforning, direction-of-arrival or noise reduction. The array is mounted above the tablet and speaker at a height of 1.1 m parallel to the robot's footprint. All sensors are connected to an embedded control computer, running the robot operating system (ROS) under Ubuntu Linux 16.04.

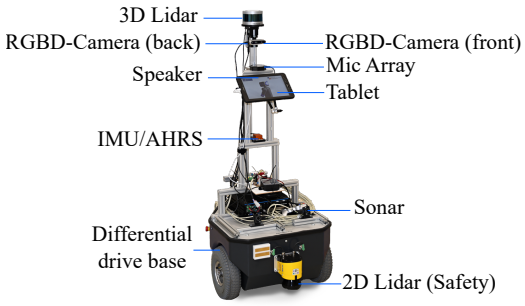

Fig. 2 Main components of the voice controlled mobile robot.

2.3 STT and TTS conversion

As of today, there exist several different solutions for STT-conversion, with deep learning techniques representing the state-of-the-art approach. Only recently, datasets of sufficient size have been made publicly available as part of the Common Voice project [8]. Since this was not yet available during the implementation of this work, especially in german, we have instead looked at various commercial systems. Of the considered systems Amazon Transcribe, Google speech API, Microsoft Bing Voice, IBM Watson Speech-to-Text, Nuance and Wit.ai only Google Speech API supports the german language as well as the possibility to provide additional proper names and a free contingent. Hence, the Google Speech API is used for STT and TTS, each accessed through preemptable action servers within the ROS framework.

2.4 NLP pipeline

To extract the semantic information from the converted speech, computational techniques which understand and learn from human language must be applied. There exist several cloud-based NLP platforms for this task, which usually are based on machine learning algorithms, eg. Google's *DialogFlow*, Facebook's *wit.ai* or IBM *Watson*. The basis of these NLP platforms are *intents* and *entities*. Intents map the user input to responses, based on a given example set of inputs. An entity is used to extract necessary data from the input. For instance, the input "What is the vegetarian dish in the canteen today?" could be mapped to an intent *meal plan* with the entities *date* (today) and *meal type* (vegetarian dish). Based on the findings of Canonico & De Russis [9], we decided to use *DialogFlow* as a platform due to the accessibility from several programming languages and large number of pre-build intents and entities. The pre-build intents can for example be used for smalltalk. We extend the system with the intents and entities shown in Table 1.

Table 1 Selection of the most important intents with examples, entities underlined.

Intent	Entities	Example
meal plan	meal type, date	"What's the dessert tommorow?"
public transport	transport type	"When's the next bus leaving?"
where is	location	"Where is room B12?"
go to	location, time	"Stand in front of the canteen at noon!"
bring me	location	"Bring me to the dean's office!"

DialogFlow receives the text input from a dialogue management system (Fig. 3) which then receives back the detected intent, proposed text output and parameters such as entities and whether the conversation is finished. Text outputs which require no further system specific data (eg. small talk answers) are then directly forwarded to the TTS engine.

2.4.1 Data management interface

If the intents *meal plan* and *public transport* are detected, the answer is augmented with the requested information via a data management interface based on Representational State Transfer (REST) APIs.

Additionally to providing up-to-the-minute data for public transit and meals, the data management interface accesses the university's room and building management system as part of the *where is*, *bring me* and *go to* intents. The essential entity for this intent is the location

This is the author's version of an article that has been published in the ICCAS 2020 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://dx.doi.org/10.23919/ICCAS47443.2019.8971465

Copyright (c) 2020 IEEE. Personal use of this material is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

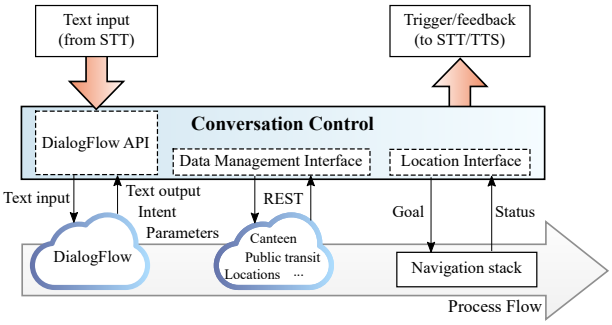

Fig. 3 The dialogue management system in detail. DialogFlow and the REST accessed data services run cloud based, whereas the other components run locally on the robot.

name (room number), which is checked for existence by the data management interface. To ensure unambiguity of location names before existence checking, a set of synonyms is defined for each location name, mapping to the data management compliant format (eg. "Office of Professor Smith" could map to "Room 3403.003.A328").

2.4.2 Location interface

Every existing location has a name, therefore a set of n location names $\{^k\mathcal{L}\}_{k=1..n}$ can be defined, which describes all possible locations. The subset

$$\{^k\mathcal{A}\}_{k=1..p} \subset \{^k\mathcal{L}\}_{k=1..n}, \quad p < n \tag{1}$$

contains all reachable locations, that can be approached by the robot in an autonomous fashion.

The basis to navigate to a specific location, is a topological map m as a 2-tuple consisting of a set of nodes $\{^k\mathcal{N}\}_{k=1..o}$ and edges $\{^{ab}\Delta\}$:

$$m = \left\langle \{^k\mathcal{N}\}_{k=1..o}, \{^{ab}\Delta\}_{\substack{a=1..o \\ b=1..o}} \right\rangle. \tag{2}$$

A database s then implements an injective function

$$s : \{^k\mathcal{A}\}_{k=1..p} \longmapsto \{^k\mathcal{N}\}_{k=1..o}, \quad o \geq p \tag{3}$$

to link the available location names to the map. When the *bring me* intent is detected, the location interface checks whether the location is reachable and asks for confirmation to bring the user to the location (see the detailed procedure in Fig. 4).

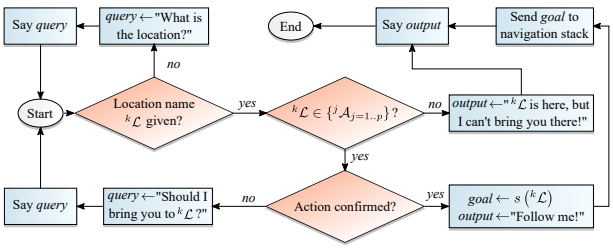

Fig. 4 Program flow of the location interface, in case of the *bring me* intent.

For non-reachable locations, or if the user only asks where a location is (*where is* intent), the system will then only show the location on the campus map and floor plan.

3. RESULTS

3.1 NLP

To evaluate the STT and general NLP pipeline, we use the following five different utterances[1]:

U1 What's tomorrow's one-plate dish at the canteen?
U2 How can you help me?
U3 When's the next bus leaving?
U4 Can you take me to building B4?
U5 Where's Professor Mueller's office?

These utterances are typical for the use case of the robot. None of the proper person or building names were specifically used for training. All utterances are spoken into a vocal microphone by eight different speakers of mixed gender and recorded without disturbing noises. The recordings are then played back in three different environments by a speaker in front of the robot with a sound pressure of averagely 60 dB measured at the center of the robot's microphone array. The environments are an office of 34 m^2 size (carpeting), the outdoor area of the campus (asphalt floor, between different buildings with glass and concrete surfaces) and a hall of 75 m^2 size (PVC floor, concrete walls).

The three environments differ greatly in terms of acoustics, which is also reflected in the reverberation times shown in Fig. 5. Especially the hall environment with $T_{20} \approx 1.5\,\mathrm{s}$ for frequencies $\leq 1000\,\mathrm{Hz}$ is suboptimal in terms of speech intelligibility. The German standard for audibility in rooms [10] recommends a reverberation time of 0.7 s for a room of this size.

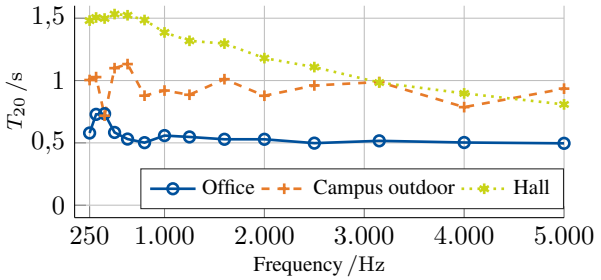

Fig. 5 T_{20} reverberation times for three different environments for the typical frequency range of speech. Determined using the *ITA-Toolbox* [11].

Each recording is played three times, totalling in 120 utterance play backs per environment. In the office and hall environment there was no interference noise during the test, whereas in the outdoor area typical interference noise such as passing cars occurred. The microphone gain is set to the same fixed value for every environment. Additionally, noise suppression and adaptive beam forming is activated.

In order to evaluate the influence of background noise, the same number of measurements are additionally performed in each environment with an additional loudspeaker that plays a background noise and faces away from the robot. The noise, which was recorded in a restaurant and contains indistinct conversation, is

[1] Translated from German

This is the author's version of an article that has been published in the ICCAS 2020 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://dx.doi.org/10.23919/ICCAS47443.2019.8971465

Copyright (c) 2020 IEEE. Personal use of this material is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

restarted for each measurement. The sound level is set to a sound pressure of 60 dB for the hall and office environment. For the outdoor environment the same settings as in the hall are used. Table 2 contains the results, separated by environment and background noise. As a performance metric for the STT module, we use the word error rate (WER) based on LEVENSHTEIN distance. Additionally the mean response time $\overline{r}$ with standard deviation σ gives information about how much time is needed to have the corresponding text available after the end of an utterance. The latency of the NLP pipeline and the TTS engine is in the range of 1 s, which must be added to get the response time of the complete system. The NLP pipeline is evaluated based on the mean of successful intent detections $\overline{s}_{\text{int}}$ and successful total detections $\overline{s}_{\text{tot}}$ (intent and all entities correct).

Table 2 Results from the evaluation in an office (Off.), campus outdoor (Cam.) and hall environment with 120 utterances per run, resulting in 720 requests in total.

	Speech recognition			NLP Pipeline	
Environment	WER	$\overline{r}$	σ	$\overline{s}_{\text{int}}$	$\overline{s}_{\text{tot}}$
Off.	5.8 %	1.23 s	0.39 s	90.0 %	83.3 %
Off. + noise	21.9 %	1.99 s	0.81 s	82.5 %	61.7 %
Cam.	8.5 %	1.40 s	0.72 s	84.2 %	80.0 %
Cam. + noise	8.2 %	1.43 s	0.60 s	93.3 %	78.3 %
Hall	8.5 %	1.40 s	0.72 s	84.2 %	80.0 %
Hall + noise	52.0 %	3.31 s	1.99 s	55.8 %	34.2 %

As it was to be expected, the system works best in environments without ambient noise and low reverberation time with a WER as low as 5.8 % and NLP detection rate up to $\overline{s}_{\text{tot}} = 83.3\,\%$. Ambient noise can greatly impair the result, especially in environments with long reverberation times. Although it must be noted that the sound pressure of the speech output was not increased during the experiments with background noise. People tend to speak louder, when ambient noise is present, therefore the detection rate would presumably be higher in real application. Table 3 shows the detection rate differentiated by utterance. Here it becomes clear that especially U2 and U3 are very well detected and the untrained names of the buildings (U4) and persons (U5) can worsen the result.

Table 3 Successful detections differentiated by utterance.

	U1	U2	U3	U4	U5
$\overline{s}_{\text{int}}$	68.6 %	87.1 %	96.4 %	96.4 %	77.9 %
$\overline{s}_{\text{tot}}$	57.1 %	87.1 %	96.4 %	53.6 %	66.4 %

3.2 User study

3.2.1 Design

In order to evaluate how well the system is suitable for the intended use with the intended users, an adequacy evaluation [12] is carried out in the form of a user study. The study consists of three parts: free interaction with the robot for two minutes, interaction under specifications and a questionnaire survey. The free interaction serves to familiarize the participant with the system. In addition, possible linguistic interactions with the robot are to be demonstrated in order to improve conversation. The only preliminary information given to the participant for this part is how speech recognition is activated via the tablet.

The interaction under specifications serves primarily to evaluate the speech interface with regard to the implemented functionalities. In addition, further training data can be generated to improve the NLP pipeline. The instructions for this part are shown in Table 4 The study was conducted in the hall environment and in German language, all answers and specifications were translated accordingly.

Table 4 Instructions for the second part of the user study, divided into three categories (canteen, location, public transit).

ID	Task
A1	Request the menu of the canteen.
A2	Specify the menu of the canteen on the dish.
A3	Specify the menu of the canteen on the day.
A4	Specify the menu of the canteen on the day and the dish.
B1	Ask where an arbitrary location on campus is.
B2	Ask where the "imes"[2]is.
B3	Ask to be guided to one of these places: canteen, porter, room A328
C1	Ask to leave the campus by bus or tram.
C2	Ask to leave the campus at a specific time.

The questionnaire includes 18 questions to obtain general information about the participants and the subjective quality of the language interface. The language interface is evaluated using a five-level Likert scale (*Strongly agree* to *strongly disagree*).

3.2.2 Evaluation

In the first part of the study, an average of 5.2 interactions with the system were carried out by each user. Through the free interaction further potential tasks of the robot could be captured. In each case, at least two test persons asked about the weather, where certain lectures take place, which events are planned and what the current location is. In Fig. 6 the percentage distribution of the success of the requests in the second part of the study is shown.

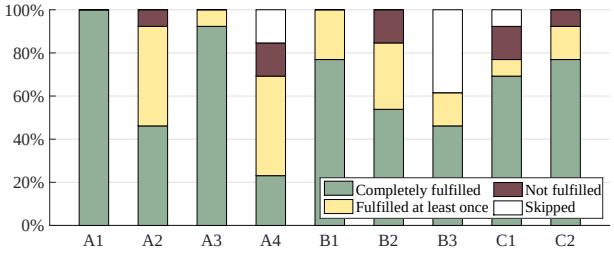

Fig. 6 Results from the second part of the user study. The labels correspond to the IDs in Table 4.

Canteen related requests (A1-A3) could be fulfilled at least once[3]for each participant in over 90 % of the cases. The greatest potential for error lies in the specification by meal type and time (A2, A4), if the word *canteen* is not used or untrained terms (e.g. *food stand*) are used. The

[2]Colloquial term for our institute

[3]Partially fulfilled, at least once in the case of more than one trial.

This is the author's version of an article that has been published in the ICCAS 2020 proceedings. Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://dx.doi.org/10.23919/ICCAS47443.2019.8971465

Copyright (c) 2020 IEEE. Personal use of this material is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

query (B1) for a self-selected location was successfully carried out by all participants in at least one case. In all cases, the correct intention was recognized, although STT could not always provide the correct location name. Every participant who did not skip question B3 could be led successfully to one of the given locations in at least one attempt. For evaluation purposes, the specified locations were defined as points within the same room.

All participants also filled out the questionnaire survey. 12 of the 13 participants (avg. age 28.15 years) have already interacted with digital assistants before. For the following statements, the median of all responses was used.

The first six questions considered the quality of the spoken language. In the median, the clarity, intelligibility, naturalness and speed of the speech output were rated as "strongly agree". Another four questions considered the adequacy of the information output. The "agree" rating was given for the information output to be correct, relevant, complete and in the right quantity. Also with "agree" was evaluated that the system confirms to have understood the entered information correctly and which information is currently being processed. The logical, task-oriented and expected structuring of the dialogues was rated with "agree" as well.

4. CONCLUSION AND OUTLOOK

We presented a voice recognition and processing interface for a mobile tour guide robot to be used in a university environment. A modular structure was presented using cloud-based services for converting speech to text (and vice versa) and the dialogue system DialogFlow as part of an NLP pipeline. The dialogue system was extended by various intents and entities for the given use case of information output and guiding functionality.

The evaluation of a total of 720 requests in three different environments with and without interference noise shows a direct correlation between the environment (and noise) and the quality of the voice recognition. A WER between $5.8\,\% - 8.5\,\%$ and total detection rate of correct intents and entities of over $80\,\%$ can be achieved without ambient noise. The greatest negative influence on the recognition rate, especially in the case of ambient noise, is the use of less common words that are not used for training. Automatically incorporating all building and room names into the TTS as well as NLP-system could therefore further increase the detection rate. This is also reflected in the user study carried out with 13 participants as a proof-of-concept of the system. Inquiries, which use pre-trained terms (eg. A1, A3, B2, B3) could be detected with a recognition rate of up to $100\,\%$, which is much more robust than requests that give the user more freedom in wording (eg. A2, A4). The parameters of the microphone array, such as gain, noise suppression or adaptive beam forming were kept constant in the course of the experiments. A possible extension of the system would be to adaptively modify these settings depending on the current location, ie. the type of the environment. The corresponding information can already be provided by the SLAM method and an approach like this promises a further increase in robustness.

REFERENCES

[1] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. A. I. Ramírez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. J. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang, "Spencer: A socially aware service robot for passenger guidance and help in busy airports," in *Field and Service Robotics*, ser. Springer Tracts in Advanced Robotics, vol. 113, Cham, 2016, pp. 607–622.

[2] R. Stricker, S. Muller, E. Einhorn, C. Schroter, M. Volkhardt, K. Debes, and H.-M. Gross, "Interactive mobile robots guiding visitors in a university building," in *2012 IEEE Ro-Man*, Paris, France, 2012, pp. 695–700.

[3] A. Poncela and L. Gallardo-Estrella, "Command-based voice teleoperation of a mobile robot via a human-robot interface," *Robotica*, vol. 33, no. 1, pp. 1–18, 2015.

[4] *Softbank robotics*, 2019. [Online]. Available: `https://www.softbankrobotics.com/emea/en/index`.

[5] *Jibo*, 2019. [Online]. Available: `https://www.jibo.com/`.

[6] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, B. Adelson, Ed., New York, NY: ACM, 1994, pp. 152–158.

[7] M. Labbé and F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 110, no. 3, p. 346, 2018.

[8] Mozilla Foundation, *Common voice*, 2019. [Online]. Available: `https://voice.mozilla.org`.

[9] M. Canonico and L. de Russis, "A comparison and critique of natural language understanding tools," in *CLOUD COMPUTING 2018*, IARIA, 2018, p. 120.

[10] *DIN 18041:2016-03:hörsamkeit in räumen (transl. audibility in rooms)*, 2016.

[11] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer, "The ITA-toolbox: An open source MATLAB toolbox for acoustic measurements and signal processing," 43th Annual German Congress on Acoustics, Kiel (Germany), 6 Mar 2017 - 9 Mar 2017, Mar. 6, 2017.

[12] M. King, "Evaluating natural language processing systems," *Communications of the ACM*, vol. 39, no. 1, pp. 73–80, 1996.

This is the author's version of an article that has been published in the ICCAS 2020 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://dx.doi.org/10.23919/ICCAS47443.2019.8971465