Procedia CIRP 97 (2020) 130–135

Contents lists available at ScienceDirect

Procedia CIRP

journal homepage: www.elsevier.com/locate/procir

Human-robot-interaction using cloud-based speech recognition systems

Christian Deuerlein[a,*], Moritz Langer[a], Julian Seßner[b], Peter Heß[a], Jörg Franke[b]

[a] Technische Hochschule Nürnberg, Keßlerplatz 12, 90489 Nürnberg, Germany
[b] Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Factory Automation and Production Systems, Egerlandstr. 7, 91058 Erlangen, Germany

ARTICLE INFO

Keywords:
Human-robot-interaction
Speech control
Cloud-service
Industrial robot

ABSTRACT

Progress in natural speech processing has enabled significantly more powerful speech processing systems, primarily due to the use of machine learning technologies. In order to integrate cloud-based speech recognition systems for human-robot interaction, an interface for the voice control of a lightweight robot was developed. The main contribution of this work is the design and implementation of a software interface to recognize commands via cloud-based speech processing and the subsequently conversion into machine-readable code. Requirements for the evaluation of different cloud-services for the control of robots are determined. Furthermore, the control architecture for the robot is modeled and implemented. An example application, which enables users to control robot movements via speech, is realized as a proof of concept and for additional studies. This application includes the basic features of cloud-based speech processing: intent recognition from utterances, slot filling and dialogue-based interaction. Lastly, the influence of background noise on process safety was examined within an experiment. It turns out that a feasible process reliability can be achieved with the system despite the presence of background noises.

Introduction: human-machine interaction

With the development of collaborative robots (cobots) humans and workers are able to share a common workspace at the same time. In such scenarios, communication is a key factor for the safety of the worker and the acceptance of the robot. To successfully fulfill a task, common goals need to be shared and one entity needs to be aware of what the other is about to do. In order to enable an interaction based on human communication, the human symbols such as words or gestures must be captured and interpreted by the robot system. The interfaces for transmitting information between humans and robots can be based on different senses or sensors. In human teams, information transfer is achieved mostly by speech. It is an intuitive, natural way of communication. Therefore, one major goal of research is the adaption of human-machine communication to human-human communication.

Speech as a user interface in human-robot teams has the advantage that the worker does not have to have eye contact with the communication partner. Furthermore, he does not need to use gestures and consequently can use his hands for other tasks. In order to convert instructions into actions of the robot, the robot system needs suitable processing components, particularly for speech recognition and speech processing. The development of components for processing of human speech can be complex and require considerable computing power. By outsourcing speech processing to cloud-services, developers of robot applications can concentrate on the design of the robot application.

The paper is structured as follows. First, a general overview over the state of research in the fields of speech control for user interfaces, cloud based speech control and their use in human-robot interaction is given. Afterwards we present our requirements for the evaluation of cloud-services, the developed software architecture as well as the control concept for industrial robots. Further, the implementation for a lightweight robot and a smart speaker is presented and evaluated by an experiment regarding background noises.

Related work: speech control – from consumer market to industry standard

Progress in natural speech processing has enabled significantly more powerful speech processing systems in recent years,

* Corresponding author.
E-mail address: christian.deuerlein@th-nuernberg.de (C. Deuerlein).

https://doi.org/10.1016/j.procir.2020.05.214

mainly through the use of Machine Learning (ML) technologies (Graves et al., 2013). Automated human-machine communication can be categorized in Non-task-oriented and Task-oriented (Almansor and Hussain, 2019). In the following implementations of Task-oriented communications are presented.

Speech control as user-interface

Voice input can be found in almost every handheld device. However, there are also developments in specialized areas. *Jorge* et al. developed a speech interface for the control of an augmented reality interface of a nuclear power plant (Jorge et al., 2010). In the paper of *Krapov* et al. speech and head-pointing is used to enable handicapped people to control a computer instead of standard input devices (Karpov et al., 2011).

Speech control in human-robot-interaction (HRI)

Speech is also used in HRI to enhance the acceptance of the robot and improve ergonomics. One challenge in human-robot teams is that human speech often needs grounding. In other words, the meaning of words depends on the context. *Wölfel* and *Henrich* developed a method to map uncertain instructions via a fuzzy logic to control an industrial robot to cope with the grounding problem (Wölfel and Henrich, 2020). *Sharan* et al. use speech input to control an assistive mobile robot platform in real-time (Sharan et al., 2019). *Skraba* et al. purpose a methodology to control a wheelchair via speech, which uses several different cloud service provider to increase speech recognition accuracy (Skraba et al., 2019). *Park* et al. developed a voice control system that enables the control of a robot in an unsafe environment like burning buildings. This is based on a locally executed speech recognition system from Microsoft Kinect (Park et al., 2015). *Zinchenko* et al. developed speech control for surgical applications. The control of relative movements was implemented with an open-source library. The movement distance was determined by the length of the utterance and therefore strongly depends on start- and endpoint-detection of the system (Zinchenko et al., 2017). Within the scope of the work of *Gustavsson* et al., voice control of a lightweight robot for human-robot collaboration was developed. They found that speech control of lightweight robots has great potential but their system lacked of accuracy, caused by the use of local processed speech (Gustavsson et al., 2017).

Performance of cloud based speech processing-systems

The performance of Cloud based Speech Recognition-Systems (SRS) has two main factors. Accuracy and delay under packet loss. *Assefi* et al. investigated the performance of Google Speech Recognition and Apple Siri, regarding network characteristics. They applied network coding on UDP connection to improve the delay while maintaining the accuracy of regular TCP connections (Assefi et al., 2015, 2020). Asian languages are especially demanding for SRS. Therefore, *Dong* et al. focused on developing a system specialized on Southeast Asian languages. They managed to achieve low word error rates with Bahasa Indonesia. For their approach they used different neuronal networks trained on about 10,000 most common words (Wang et al., 2017). *Morbini* et al. compared different automatic speech recognizers regarding the word error rate and found that those systems experience a rapid improvement from each generation (Morbini et al., 2013).

Open challenges of research

Resulting from the fast development of cloud-based speech recognition, those systems are hard to compare. Therefore, one objective is to define appropriate factors to evaluate a suitable system. In addition, we want to investigate the influence of background noises on the performance of the speech recognition. In addition, we want to examine whether rules for the development and design of human-robot assembly cells using smart speaker can be derived from the experiments.

Requirements for the evaluation of cloud-services for the control of robots

The interaction with the robot, based on spoken dialog can be roughly divided into two successive tasks: intent detection of the utterance and intent fulfillment. The intent fulfillment is an action of the robot system in most of the cases. The main focus of HRI developers is on the implementation of the robots' task. As a result, the task of the provider of the speech recognition system shall be the intent-recognition part. Although natural language processing is the core functionality of the speech recognition system, supporting modules and functions are necessary to implement a speech interface:

- Recording device: handles the recording of utterances and the communication with the cloud infrastructure of the provider. This includes wake-word detection and end of utterance detection.
- Dialog management: enables the user to interact with the system on a dialog basis.
- Filtering of background noises, intent verification and validation.
- Speech synthesis module: enables the interface to respond to the user with spoken answers.
- Tools for creating the language understanding model.

The criterion to compare different cloud based speech recognition systems in regard to use them in a speech interface with a robot is whether those supporting modules are provided and how efficient and easy to use they are. In this paper, Amazon Alexa, Google Dialogflow and Microsoft LUIS cloud based speech recognition systems are considered.

The Microsoft LUIS API provides functions to extract the meaning and core information entities from natural language text (Text-to-Speech). The user can define an application specific language model, which includes all the relevant intents for using the application. Since the LUIS API needs a text as input, the intermediate step of converting recorded utterances of the user to text has to be implemented separately. However, Microsoft also provides services for Text-to-Speech conversion and dialogue management as well as Speech-to-Text synthesis. To implement a speech interface based on LUIS, different services have to be chained together. This provides a lot of flexibility but also increases the developing scope (Micosoft, 2019).

Google Dialogflow is a tool to build conversational speech interfaces. To use it, a Dialogflow Agent, which includes the language model for the application, has to be configured. This agent can then be used in combination with a Google Assistant able device. It handles all the tasks relevant to recording the utterances, wake-word detection and speech synthesis. The agent handles the intent detection and the extraction of slot values. Once an intent is detected, the agent sends out a request to the intent fulfillment. This module has to provide HTTPS endpoints in order to send the requests for every intent (Google, 2019).

Amazon Alexa works similar to Dialogflow but provides the developer with more options to configure the language model. Together with the dialog model, the language model forms the Alexa Skill which gets invoked by an Alexa enabled device. In the skill configuration, the developer can set up intent confirmation and

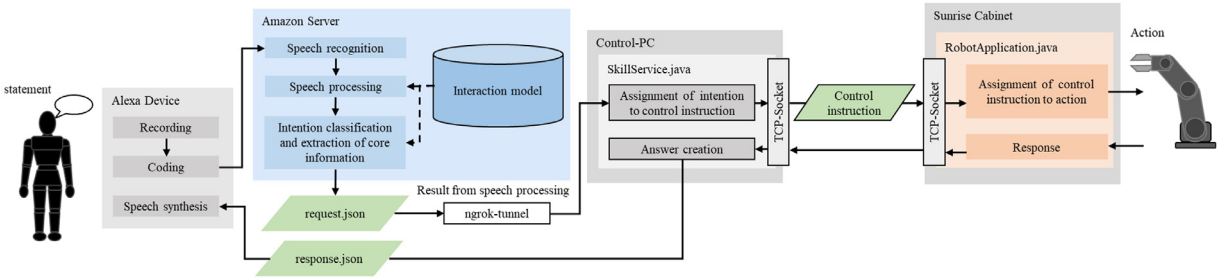

Fig. 1. System architecture of the implementation. From left to right: Alexa Device, Amazon Server with Alexa Skill and interaction model, Control PC with Skill Service, robot controller Sunrise Cabinet.

slot value validation for every intent and slot. This enables to verify intents by asking the user to confirm if the system has assigned the utterance to the correct intent, before sending the request to the intent fulfillment. The request is sent to a single HTTPS endpoint which handles the requests for all intents. With the Alexa Skill Kit, Amazon also provides a Software Development Kit to build the fulfillment service (Amazon, 2019).

To select a suitable cloud based speech recognition system, comparison criteria are required. One commonly used metric for evaluating such systems is the word-error-rate. This metric is generated by comparing results of the speech recognition system with a correct transcript of the utterance. Every deviating word in the result of the system affects the word-error-rate. Although the correct recognition of single words in an utterance is important for speech recognition systems, in human-robot-interaction it is more important to correctly detect the intent of an utterance is. The performance of intent detection can be quantified with the F1-score. The results of the intent detection of a system can contain, true positives, false positives, true negatives and false negatives. From these results, the F1-score can be calculated.

It turns out that evaluating the word-error-rate as well as the F1-score, for the intent detection of different systems is quite complex and time consuming. Every system has to be configured with comparable interaction models and be tested with the same domain of utterances. Even then, the results do not provide a general conclusion. Since cloud based systems are changed and improved constantly, the results of such experiments are very short-lived (Braun et al., 2017). Therefore, the usability, the adaptability and the scope of the modules provided by the cloud-service system are the main decision criteria for choosing a suitable service. They have a significant influence on the architecture of the voice interface to the robot controller.

Amazon Alexa meets the requirements to provide a platform to efficiently build the module for the intent recognition task of the HRC applications best, thus it was chosen for the interface in this paper. With the “Echo” series, Amazon provides the user with off the shelf hardware for recording utterances and speech synthesis. The clear separation of intent detection and intent fulfillment into two software modules, one running in the cloud and one running locally on a control-PC makes those modules more maintainable.

Implementation: speech control of a lightweight-robot using smart speaker

The advantages of using smart speaker, such as the continuous improvement of speech processing by the provider, who has access to the voice data of millions of users and can rely on a well-developed infrastructure, are the main reasons for using a smart speaker as input device. Furthermore, no additional hardware is needed for the voice input. The smart speaker is installed permanently within the cell. If the worker needs to wear additional equipment like a headset, there is a risk of forgetting to wear them. In addition, wearing e.g. a headset for long periods of time can be uncomfortable.

We implement cloud based speech control for a lightweight robot for further experiments. The four main components of the system are shown in Fig. 1: Alexa Device, Amazon Server, Control PC and Sunrise Cabinet. The configuration and preparation of functions and actions take place in the individual modules. The configurations depend on each other. Therefore, a change of the interaction model results in a corresponding change of the module for intention fulfillment.

System architecture of the speech-interface

The Alexa-enabled device represents the speech interface with humans on the hardware side. It has a continuous connection to the Alexa Server and can evoke configured Alexa Skills. The interface is implemented within the development console.

The next module runs on the Amazon Server. It is used to configure which user intentions are possible in the context of the skill and which utterances can be used to express them. Each intention can be associated with core information in the interaction model. The task of the skill is the classification of user statements with intentions, predefined in the interaction model and to extract core information from the utterance.

The Skill Service runs on the Control PC and is the component that fulfils the intention expressed by the user. The Skill Service must receive the results of the voice-processing, running in the cloud and send an appropriate control command to the robot controller. The Skill Service receives the result of the speech processing as a JSON object and evokes a corresponding function to fulfill the contained intention. The function then creates a TCP socket for the robot controller and sends a matching control command. The Alexa Skill kit for Java enables the implementation of the skill service as a servlet. The open-source servlet engine “jetty” is used as the web server for the execution of the servlet. While the skill service is running, the jetty server waits for requests of the Alexa speech processing on a defined TCP port. This port is made accessible from the web via an ngrok tunnel.

A KUKA iiwa robot is controlled via the Sunrise Cabinet, which executes the robot application. The robot application receives the control commands from the skill service and causes the robot to perform corresponding actions.

Example application

The example application intends to demonstrate the possibilities of using Alexa for human-robot interaction. The control of linear relative movements and circular movements by speech is chosen as example functions. All essential features of the dialog-based voice control based on Alexa are integrated in this control. The expectation of the user is a relative movement in a certain direc-

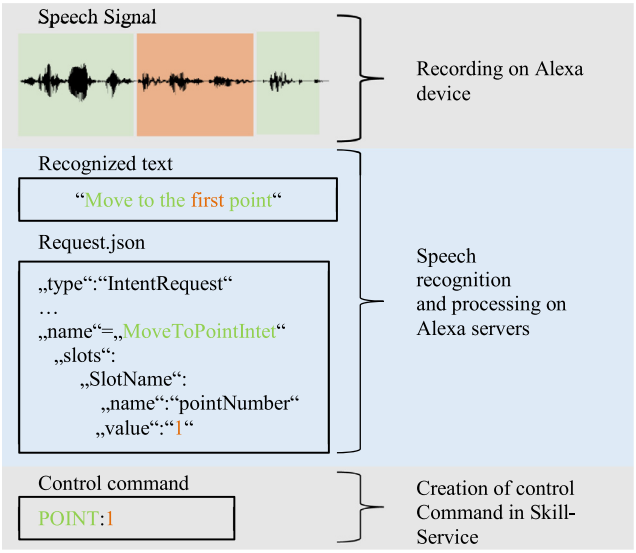

Fig. 2. Information transformation from utterance to machine command.

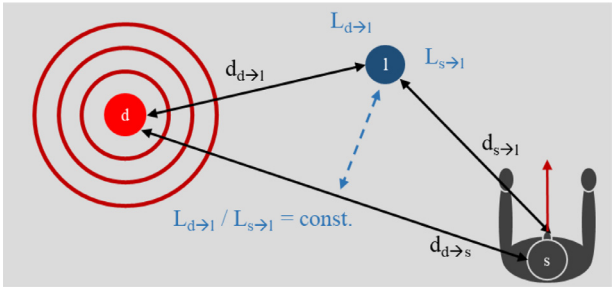

Fig. 3. Layout of disturbance sound (d), smart speaker (l) and voice input (s).

tion. The amount by which the tool center point is to be relatively moved is core information of intent.

Sequential data transformation

The main task of the speech interface is to transform the intent of the user into machine readable code. The conversation begins with the utterance of the user, which is recorded by the Alexa device. Afterwards, speech recognition takes place and the intend of the statement is classified. For this purpose, a corresponding intent is stored in the interaction model, to which the statement is assigned. The core information of the intention is then extracted from the utterance. These and other Meta data are sent to the skill service which then creates the control command for the robot application. The whole process can be seen in Fig. 2.

Experimental setup: influence of background noises on process reliability

Regarding the use of smart speaker in robotic applications, one mayor Problem is disturbance noise. Therefore an additional goal of this paper is to investigate the influence of disturbance noise on the accuracy of the intent recognition.

Experiment preparation

The test layout can be seen in Fig. 3. The sound ratio regarded to in this paper, is considered as the quotient of the sound level of the disturbance sound measured at the smart speaker $L_{d\to l}$ and the sound level of the voice input measured at the smart speaker $L_{s\to l}$.

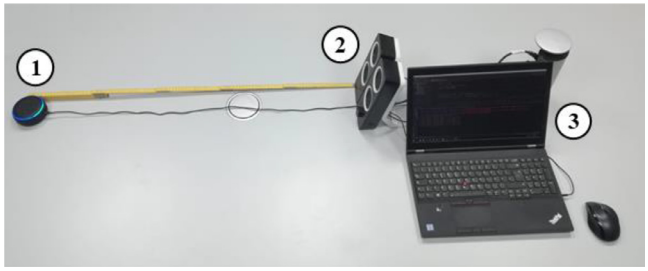

Fig. 4. Experimental setup for background noise testing. Distance from speaker to Echo Dot is set to 1 m.

It is measured in decibel and is 0 dB if the sound levels are equally high. If the disturbance sound level is lower than the voice input sound level, the ratio is negative. On a straight line connection between speaker and interference source, isolines run perpendicular, with a constant ratio. The goal of the experiment is to find setups where the smart speaker is still able to correctly fulfill the intend. With this information it is possible to measure the sound ratio in a real assembly cell and immediately determine if the smart speaker will be able to successfully interpret the voice input although background noises are present.

Experiment implementation

For the experiment, we use recorded disturbance noises from an assembly environment and a record of a spoken instruction from a human to the smart speaker. Both records are unified to have the same initial average sound level. The disturbance noise is amplified to generate different sound level ratios. Afterwards, both records are conflated in one file and played from one device. Therefore, it is assured that the sound ratio is not distorted due to inaccurate experimental setups of the speaker and the microphone. The Echo Dot 3rd generation (1) is set up 1 m away from the speaker (2) and the system volume of the control PC (3) is adjusted, that the initial playback of the undisturbed intention is about 60 dB sound level (see Fig. 4). This is the sound level resulting from a normal conversation in 1 m distance. To ensure that the smart speaker is waiting for instructions we play the wake-word and the skill invocation without any disturbance. After the disturbed intent is played, the skill is terminated. The manipulation of the ratio as well as the playback of all sounds is automated with a python script using the *pydub* library. The timestamp when the playback of the sample begins as well as the regarding noise amplification is tracked in a log file. The intent invocation is also logged in a separate file, implemented in the *jetty servlet*.

We expect that there is a certain sound level ratio from which on the speaker cannot correctly interpret the intention anymore. This results in a drastic drop in the detection rate. If the limit ratio is found, it can be used to design the placement of the speaker within an assembly cell.

Results: success-rate of intent recognition

Fig. 5 sums up the experimental results. A total of 555 samples are played during the experiment. One test set consists of 37 samples with different noise amplifications and the experiment was repeated 15 times. Two of the samples are not valid due to interrupts during the experiment. 553 valid samples are used to determine the accuracy with different noise amplifications. If the smart speaker fails to assign the voice input to an intent, the skill does not evoke any intention handling and as a result there is no log entry of the *jetty servlet*. The gap in the log thus indicates an unsuccessful attempt. The reference sound level is the level of the speech

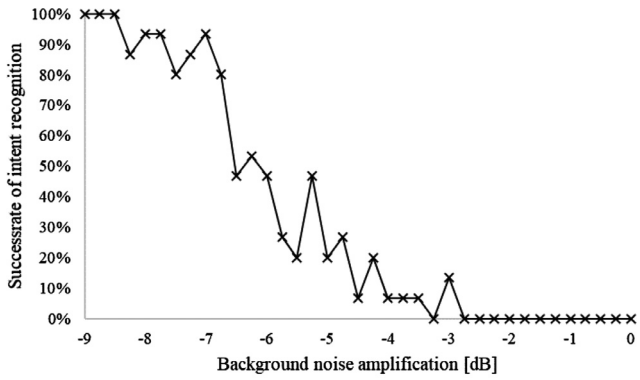

Fig. 5. Success-rate of the smart speaker with different sound level ratios.

recording. Noise amplification is given in a relative value compared to the speech. Consequently, if the sound level of the noise and the speech are equal, the amplification is 0.00 dB. If the noise is quieter than the speech input, there is a negative amplification. The noise amplification reaches from a minimum of −9.00 dB to a maximum of 0.00 dB in 0.25 dB steps.

First failed intention recognitions occur with -8.25 dB amplification. Two intent recognitions failed of a total of 15 samples (87% succeeded intent recognitions). From -6.75 dB (80% succeeded intent recognitions) to -6.50 dB (47% succeeded intent recognitions) there is a drop of 33% regarding the accuracy performance. When the amplification reaches -2.75 dB the system is not able to detect any intention correct anymore.

Discussion: limit-ratio of background noise to utterances

Contrary to expectations, there is no abrupt drop in accuracy. Within the experiment, the biggest drop of the success rate is at -6.75 dB. The decrease of the accuracy is about −17% per 1 dB noise amplification in the range from -8.75 dB to -2.75 dB. With regard to the use of smart speaker for human-robot interaction, this means that the utterance must be at least 8.75 dB louder than the background noise. Only then it can be assured, that the right intent will be evoked. This ratio depends only on the sound level of the background noise and the sound level the intend. As a result, it can be used for the setup of different robot cells. Practical use of the experimental results:

With the limit ratio of -8.75 dB, we are able to assess if a smart speaker in an assembly cell will be able to detect the spoken intents of the worker without further testing. Therefore, the subsequent steps need to be followed:

- Setup the assembly cell and position the smart speaker.
- The worker utters the intention while no disturbance sound is present.
- The undisturbed intention is recorded and the sound level is determined.
- The disturbance noise sound level is measured.
- If the sound level of the noise is smaller than the sound level of the workers' utterance minus 8.75 dB, the smart speaker will recognize the intention.

Conclusion and future work

In this paper the use of smart speaker for communication in HRI scenarios is discussed. The comparison of such systems by common performance criteria as the word-error-rate or F1-score is very time-consuming. Moreover, the comparison is only valid for a short period of time due to the rapid improvement of the systems. Another point in the context of HRI is that in most cases the focus of the developer is on the robots task with its challenges and safety issues and not on the implementation of the speech input. Therefore, the main criterion for the comparison of speech processing systems is a simple implementation. In addition, a software architecture for the control of a lightweight robot with speech is presented. To evaluate the architecture, the implementation with an *Amazon Echo Dot* and a *Kuka iiwa* is realized. With this setup a worker is able to control the linear and circular movement of the robot in a two dimensional plane. The intent model can be expanded to cover more functionality. The key findings of this work are:

- Because the main focus of HRI applications is on the safety and process realization, the used speech recognition system needs to be easy to use. The precision of the system experiences a fast change and is not valid for a long time.
- Background noise is the limiting factor for the smart speaker.
- The sound ratio between the human utterance and the background noise should not exceed -8.75 dB to grant safe intent recognition with this implementation.

Voice control for lightweight robots can contribute to a more intuitive and natural interaction between man and machine. The use of smart speaker offers several advantages. The technology can be used to access a highly developed voice processing system with a well-developed infrastructure. Through the use of machine learning and enormous learning data from users all over the world, a constant improvement takes place. In addition, the integration of the smart speaker in the robotic cell avoids the need for additional hardware, which can lead to inconveniences when worn for long periods of time.

However, the use has also some drawbacks. For further implementations with regards of industrial applications, a secure handling of the user data must be ensured. Furthermore, the time required for transmitting and processing the speech input must be reduced. Only then speech control can become suitable for safe real-time applications. These are the keys to the long-term establishment of such systems. One possible approach is to use offline solutions as Mozilla DeepSpeech or Rasa.

Acknowledgments

This work was conducted within the project MRK&MoCap4Robots (13FH008IX6), funded by the German Federal Ministry of Education and Research (BMBF). We are thankful for the support and funding from the BMBF.

References

Almansor, E.H., Hussain, F.K., 2019. Survey on intelligent chatbots: state-of-the-art and future research directions. In: Proceedings of the 12th International Conference on Computational Intelligence in Security for Information Systems (CISIS), 13-15, p. 534.

Amazon, 2019. Alexa Skill Kit: https://developer.amazon.com/en-GB/docs/alexa/ask-overviews/build-skills-with-the-alexa-skills-kit.html, accessed: 2019-11-24.

Assefi, M., Wittie, M., Knight, A., 2015. Impact of network performance on cloud speech recognition. In: Proceedings of the 24th International Conference, p. 1.

Asseffi, M., Liu, G., Wittie, M.P., Izurieta, C., 2020. An Experimental Evaluation of Apple Siri and Google Speech Recognition. Department of Computer Science, Montana State University.

Braun, D., Hernandez, A., Matthes, F., Langen, M., 2017. Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics, p. 174.

Google, 2019. Dialogflow API reference: https://dialogflow.com/docs/reference/agent, accessed: 2019-11-24.

Graves, A., Mohammed, A.-r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p. 6645.

Gustavsson, P., Syberfeldt, A., Brewster, R., Wang, L, 2017. Human-robot collaboration demonstrator combining speech recognition and haptic control. In: Proceedings of the 50th CIRP Conference on Manufacturing Systems, p. 396.

Jorge, C.A.F., Mól, A.C.A., Pereira, C.M.N.A., Aghina, M.A.C., et al., 2010. Human-system interface based on speech recognition: application to a virtual nuclear power plant control desk. Prog. Nucl. Energy 52, 379.

Karpov, A., Ronzhin, A., Kipyatkova, I., 2011. An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. In: Human–Computer Interaction, Part II. Springer-Verlag, Berlin-Heidelberg, p. 454.

Micosoft, 2019. Language Understanding (LUIS) documentation: Learn how language understanding enables your applications to understand what a person wants in their own words. https://docs.microsoft.com/en-gb/azure/cognitive-services/luis, accessed: 2019-11-24.

Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., et al., 2013. Which ASR should I choose for my dialogue system? In: Proceedings of the SIGDIAL Conference, Metz, France, pp. 394–403.

Park, S., Kim, Y., Matson, E.T., Jang, H., et al., 2015. An intuitive interaction system for fire safety using a speech recognition technology. In: Proceedings of the 6th International Conference, p. 388.

Sharan, S., Nguyen, T.Q., Nauth, P., Araujo, R., 2019. Implementation and testing of voice control in a mobile robot for navigation. In: Proceedings of the IEEE/ASME, p. 145.

Skraba, A., Kolozvari, A., Kofjac, D., Stojanovic, R., et al., 2019. Development of cyber-physical speech-controlled wheelchair for disabled persons. In: Proceedings of the 22nd Euromicro Conference, p. 456.

Wang, L., Tong, R., Leung, C.-.C., Ni, C., et al., 2017. Cloud-based automatic speech recognition systems for southeast Asian. In: Proceedings of the Languages Proceedings of the International Conference on Orange Technologies (ICOT), Singapore. Kent Ridge Campus, National University of Singapore, p. 147.

Wölfel, K., Henrich, D., 2020. Grounding of uncertain force parameters in spoken robot commands. In: Advances in Service and Industrial Robotics. Springer International Publishing, p. 194.

Zinchenko, K., Wu, C.-.Y., Song, K.-.T., 2017. A study on speech recognition control for a surgical robot. IEEE Transactions on industrial Informatics 13 (2), 607.