INCREASING INDUSTRIAL PRODUCTIVITY BY EMPLOYING A SMART SPEECH-BASED QUESTION ANSWERING ASSISTANT

Bojan Kovachki[1,2], *Aleksandar Gjoreski*[1], *Robert Herms*[1],
Matthias Pohl[1], *Martin Wenzel*[1], *Frank Seifert*[2], *Diane Hirschfeld*[1]

[1]*voice INTER connect GmbH, Ammonstraße 35, D-01067 Dresden, Germany,*
[2]*Technische Universität Chemnitz, Straße der Nationen 62, D-09111 Chemnitz, Germany*
{bojan.kovachki, aleksandar.gjoreski, robert.herms}@voiceinterconnect.de

Abstract: This paper proposes a system design for a question-answering assistant, aiming to increase the productivity in industrial environments. The system is characterized by the fact that its components and domain specific knowledge are embedded on a given hardware platform locally. The effects of introducing a prototype-assistant on the productivity and the perceived workload are explored empirically by conducting experiments designed for this purpose. As a result, users were able to answer the given questions faster when using the prototype than when searching manually. Moreover, after employing the prototype, users reported their task to be less demanding, with a positive effect on the human-machine interaction.

1 Introduction

Spoken dialogue systems are getting increasingly commercialized, and devices such as Google Home and Amazon's Echo (Alexa) can be found in many households nowadays [1][2]. Furthermore, automatic speech recognition (ASR) has found its use in many different areas such as health care [3], robotics [4], with an ever growing presence in the industrial world [5]. With the growth of the voice recognition market, the rise in popularity of voice recognition technology and its increased presence in the industry, various solutions that aim to improve the human-machine interaction have been suggested, culminating with multi-modal approaches, which combine inputs such as speech, gesture, gaze and movement [6]. In particular, in working environments, operational benefits in terms of spoken dialogue systems are, for example, workload reduction and an increase in productivity [7].

The concept of productivity, in general, can be defined as the relation between output and input [8]. It is often described as being about the speed and accuracy of carrying out specific tasks. In order to generate more productive environments in terms of individuals means that one has to understand more about the nature of work and how individuals deal with work. In this context, barriers to effective work performance can include distraction and poor support systems [9]. On the other hand, various performance objectives, such as fast and high quality operations, can have a positive effect on the productivity—that is, reducing effort and time by, for example, avoiding re-doing things [8]. A number of personal factors (e.g., physical and mental health) and external factors (e.g., environment and work-related systems) may influence the level of productivity. Typical indicators of productivity include performance measures, e.g., speed of working or accuracy of outputs. Moreover, subjective measures using questionnaires and scales can be employed, also in combination with physiological measures such as brain rhythms [10].

Since today's question answering (QA) systems work on human expert level for some specific tasks [11], one can assume that such an assistant can contribute to a positive effect on the

productivity in daily life as well as in the industry. The impact of QA and ASR systems on productivity has been a research topic for many years, from early works comparing ASR to typed input and single keypresses [12], to later works which shifted the focus towards the problem of question-answering. A recent work studying the ability of users to answer questions using an interactive document retrieval system concludes that the question answering systems have the potential to significantly increase the rate at which users find answers [13].

Although there are many QA systems, most of them are not suitable for industrial environments. Some of them are not able to accept natural speech as input, binding the users to more traditional input methods, which are often unsuitable in those work environments. Other systems are not designed for the constraints of those environments in mind, and are, for example, requiring Internet connection, or are not able to be adapted on a very specific use-case.

In this work we are proposing a design for a system that represents an intelligent industrial assistant which converses with the user in a natural language. The system receives a speech signal as an input, a user utterance in the form of a question, and it is able to comprehend a text corpus to offer an answer. The assistant can be trained for specific use-cases and keeps all of the data locally and solely on the given hardware platforms.

This paper is organized as follows: Section 2 proposes the design for the ASR QA system. Section 3 describes the adaptation process for a specific use-case and an experimental setup to show the effectiveness of the system, while Section 4 presents the results of the experiments. Section 5 concludes the paper.

2 Smart Speech-Based Question Answering Assistant

The proposed system consists of two modules, one for automatic speech recognition and one for question answering. The QA module processes only text input, which is the direct output of the ASR module.

2.1 Automatic Speech Recognition

For speech recognition, the vicCONTROL industrial[1] recognition system is used, with the vicSDC application for dialog design [14]. The fact that this system is running completely locally on the target device makes it suitable for the industrial environment. Also, this system supports automatized training and adaptation of the recognition resources, including the statistical language model. The natural language understanding module of vicCONTROL industrial is not used, as the recognized text is fed directly to the question-answering module, unmodified.

2.2 Question-Answering

The question-answering module is based on BERT [15], a state-of-the-art model that has been proven to be successful for question answering tasks, and solutions based on its modifications are topping the current rank-lists[2]. For the exact implementation, the cdQA project[3] was chosen.

The architecture of this module contains two main components: a retriever and a reader. The retriever is based on Facebook's DrQA retriever[4] which relies on sparse, TF-IDF weighted bag-of-words vectors (in this case, unigrams and bigrams), and its purpose is to retrieve the relevant paragraphs of the text corpus where the answer could be located. The selected paragraphs and the query itself are then forwarded to the reader.

[1] https://voiceinterconnect.de/en/viccontrol_industrial

[2] https://rajpurkar.github.io/SQuAD-explorer

[3] https://github.com/cdqa-suite/cdQA

[4] https://github.com/facebookresearch/DrQA

The reader consists of a machine reading comprehension (MRC) neural network model implemented into this question answering dialogue system. The goal of this model is to understand a document (e.g. a user manual) and respond to queries based upon some given context paragraphs.

The reader used in this system is pre-trained on the SQuAD1.1 dataset[5], which consists of more than 100,000 question-answer pairs on over 500 Wikipedia articles. The pre-trained model is provided as part of the cdQA project by HuggingFace[6]. However, as the way the information is presented in the user manuals differs significantly from encyclopaedia articles, additional model adaptation was performed.

3 Study on the Effect of ASR QA on Productivity

In order to verify the effect of our ASR QA system on the productivity of users, performance and subjective measures were derived in a search-task scenario. In the following subsections, the data used for the study, the experimental setup, and the task design are presented.

3.1 Model Adaptation

The search task scenario is related to the operating instructions manual of the Braillo Cut—Automatic Forms Cutter machine, model 6609d-6612d[7]. The QA database is created by extraction of the raw data found in the user manual. The resulting database consists of approximately 19,000 tokens. The data used for the language modelling is a fusion of the QA database, the annotation data (252 QA pairs) and an additional set of 50 questions created for the purpose of speech recognition optimization. Table 1 describes the data used for QA adaptation. Some of the questions have the same answer, i.e. some of the answers are found more than once in this database. After the input file is transformed into a structured document, the QA pairs were annotated with the help of the cdQA-annotator tool[8].

Table 1 – Description of data used for QA adaptation

Description	**Questions**	**Answers**	**All**
Number of Sentences	252	252	504
Number of Tokens	1,994	9,829	11,823
Average number of tokens per sentence	7.97	39.00	23.46

The fine-tuning was done using a TPU[9] in Google Colab[10] and it was completed in a matter of hours. In our case, the following empirically-obtained hyperparameters produce the optimal results: train batch size: 2, learning rate: 2^{-5}, number of epochs: 20. The dropout probability is kept at 0.1.

Regarding language model training for speech recognition, all tokens from the user manual and the QA pairs are used. Additional 50 knowledge-based general questions were included in the language model, as question phrases are rarely found in user manuals.

[5] https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/

[6] https://github.com/huggingface/transformers

[7] https://braillo.com/wp-content/uploads/2017/12/Braillo-Cut-Manual.pdf

[8] https://github.com/cdqa-suite/cdQA-annotator

[9] https://cloud.google.com/tpu/docs/tpus

[10] https://colab.research.google.com/notebooks/intro.ipynb

3.2 Task Design

The goal of the task is to compare the user productivity by human search in user manuals regarding industrial devices provided in text form, with the employment of a question-answering speech assistant. In this context, the human performance (i.e., time required to complete a task) as well as the user experience in terms of workload is of interest—these parameters have been found to be important determinants of productivity [10][16][17][18].

The task is composed of two phases, whereby 10 out of 20 questions are randomly chosen for the first phase and the remaining 10 questions for the second phase. Hereafter, the phases are given at a glance:

- Phase 1: Subjects are required to read a question and then to find the answer in the PDF document (Section 3.1) presented using a desktop computer which is equipped with a mouse and a keyboard. Common search functions provided by the PDF reader (Okular version 1.6.3) or scrolling in the document are allowed. Once a subject believes to have found the answer, he or she should say 'stop' and the time span required starting at the search procedure is logged.

- Phase 2: Speech is used as the input at the same desktop computer, where now the ASR-QA system is employed—subjects are provided with a headset (Sennheiser PC 3 Chat Headset) including a close talk-microphone. First, a preliminary test was conducted with a question not being used in the real test in order to familiarize the users with the system. First, subjects had to utter the ASR wakeup word ('hey assistant') and then read the question aloud. The answer is then displayed on the screen via a web interface as shown in Figure 1. Then, subjects are required to read and verify the answer in terms of its correctness—which should be part of the procedure in order to make this search comparable to phase 1. Note that the subjects do not know beforehand whether the answer of the system is correct. After verification, the subject has to say 'stop' and the time is logged from the start of the subject's speech, i.e., from the beginning of the wakeup word uttered.

In total, 15 20-to-55-years old participants (employees with academic background in terms of economics or engineering; 11 male and 4 female) performed the task in an office environment. At the end of each phase the NASA-TLX [19] was used to obtain the subjective experience of the users engaged in this human-machine interaction scenario. This experiment lasts about 45 min for each subject.

4 Results

During the experiment, all participants were able to use the speech recognition system with no observed difficulties. There were misrecognitions noticed, but they were overcome after retrials. The time for retrials was also included in the measurement. It was noticed that some minor misrecognitions did not affect the system's question-answering performance.

After completing the experiments, short interviews with the participants were conducted. At those interviews the participants have confirmed our expectations and reported feeling less pressure when using the system in the second phase. Also, many participants noted the performance of the system was beyond their initial expectations. Fourteen of the fifteen participants reported that in real-world case they would prefer to have the speech question answering system to search for information.

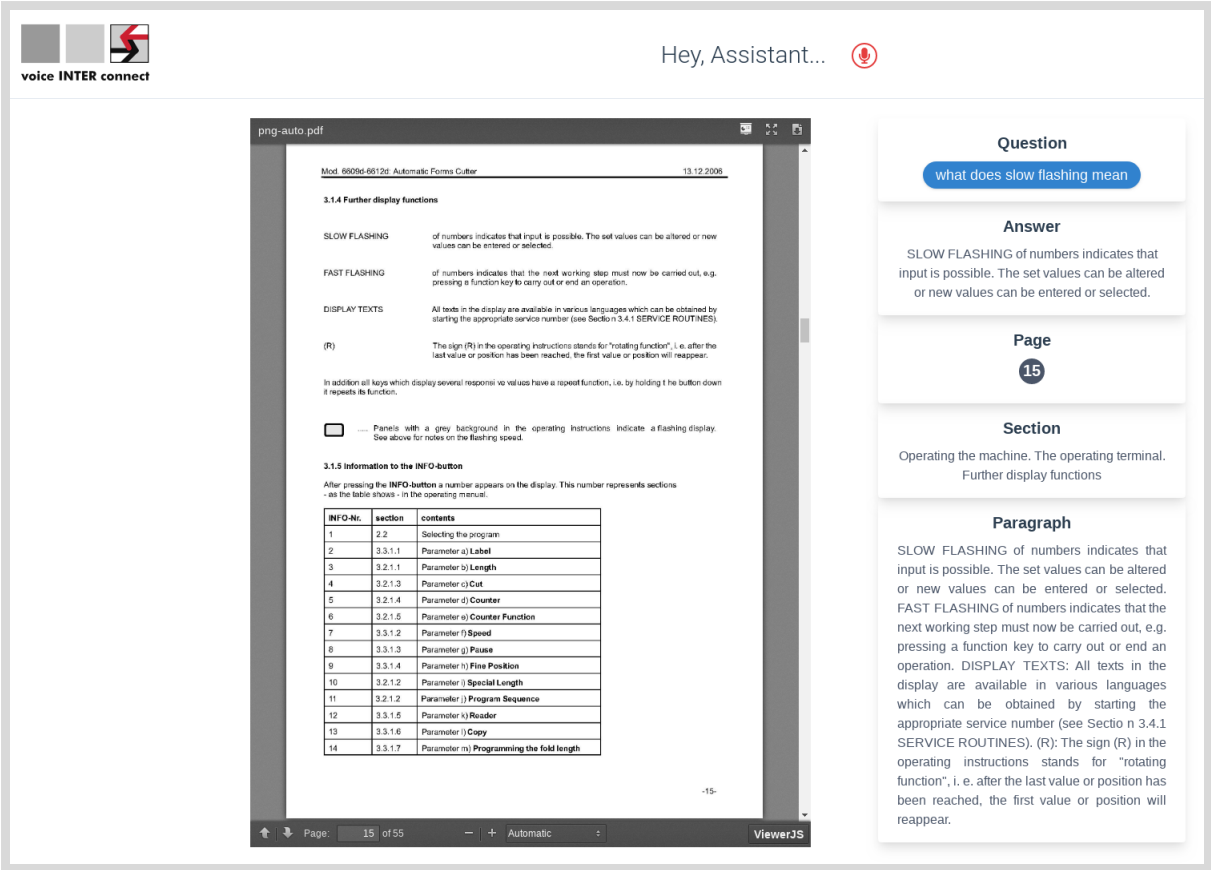

Figure 1 – The graphical user interface used during the experiments

4.1 Performance

Regarding the average performance measures per subject under the two task phases, Figure 2 presents the means and 95% confidence intervals. One can see that the distributions in terms of human search ($min = 26.47$, $max = 107.22$, $\mu = 52.45$, $\sigma = 19.46$) and the QA system ($min = 9.71$, $max = 26.41$, $\mu = 16.56$, $\sigma = 5.36$) are well separated. As a consequence, when using the QA system, less time is needed for retrieving the answer to the question given. Moreover, the variation in time decreases by introducing the system. The difference between means is 35.89 s, whereby a statistical significance can be observed at the significance level of $\alpha = 0.01$ (paired t-test).

4.2 NASA-TLX

Figure 3 presents the results obtained from the NASA-TLX questionnaire. The grades from 0 to 20 were averaged for each subscale separately, for both task phases. From the results it can be seen the phase 1 was perceived to be more demanding in general, across all subscales. The frustration and effort were also reported higher. In the second phase, the participants reported considerably lower demands, frustration, and effort, while grading their own performance higher. From the obtained results it can be concluded that the proposed speech QA system is perceived to be less complicated, i.e. less demanding for usage than the manual search.

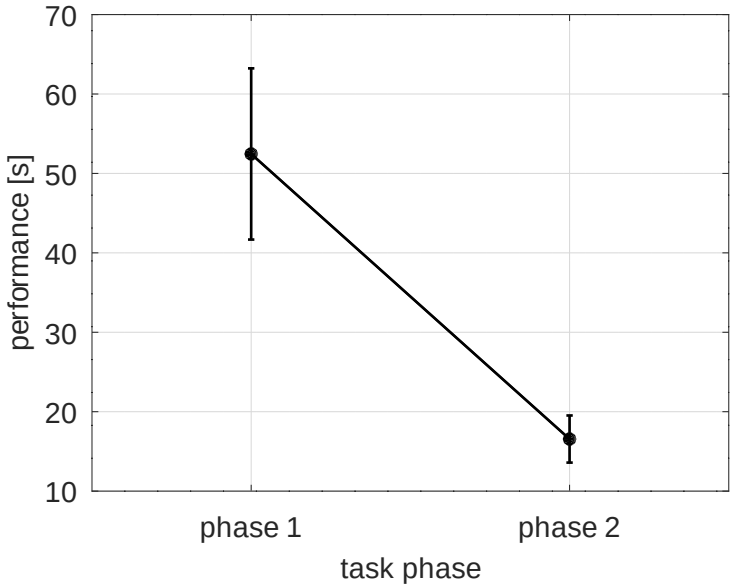

Figure 2 – Means and 95% confidence intervals regarding the average performance in seconds per subject for each task phase

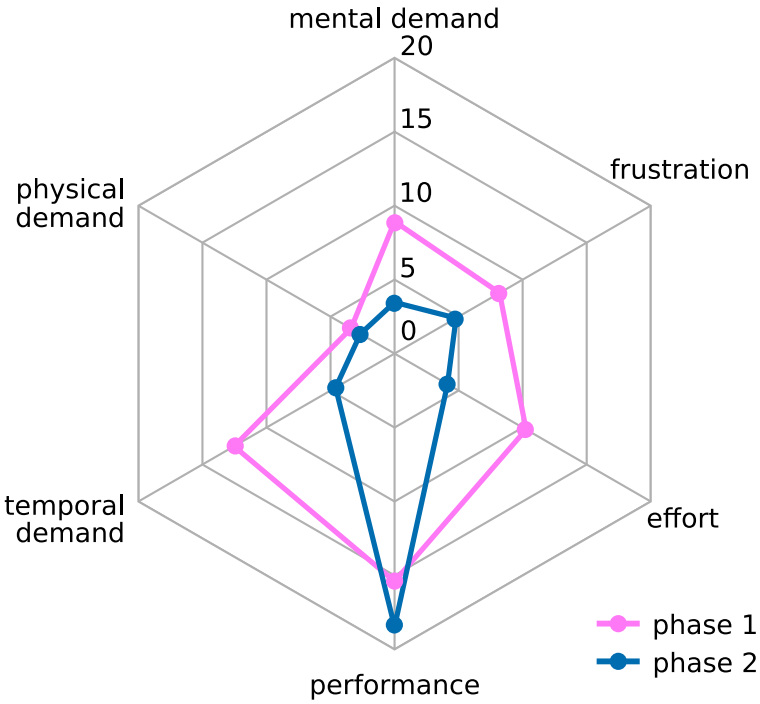

Figure 3 – Mean of the grades given on the NASA-TLX questionnaire for each task phase

5 Conclusion

We presented the system design of an industrial assistant which covers the technologies from the field of automatic speech recognition as well as question answering. In order to verify the effect of using such an assistant on the user's productivity, an assistant-prototype has been developed which is aimed for the employment in a search task scenario in which users are required to find an answer to a question using a digital text document. This paper reports the design of the search task in which two phases are contained: 1) human (manual) search by using a mouse and a keyboard; 2) speech as input for the assistant-prototype. Experimental results show a positive effect on the productivity when the assistant-prototype is used, which could be derived by performance and subjective opinion measures.

For the future, we plan to conduct further experiments, in particular with noisy conditions that affect both the user as well as assistance that is based on speech recognition.

References

[1] BOHN, D.: *Amazon says 100 million Alexa devices have been sold – what's next?* theverge.com, 2019. URL `https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp`. Accessed: 2020-01-19.

[2] MCCARTHY, P. X.: *Conversational AI – A new wave of voice-enabled computing*. forbes.com, 2019. URL `https://www.forbes.com/sites/paulxmccarthy/2019/12/17/conversational-ai---a-new-wave-of-voice-enabled-computing/`. Accessed: 2020-01-19.

[3] JOHNSON, M., S. LAPKIN, V. LONG, P. SANCHEZ, H. SUOMINEN, J. BASILAKIS, and L. DAWSON: *A systematic review of speech recognition technology in health care*. *BMC Medical Informatics and Decision Making*, 14(1), p. 94, 2014. doi:10.1186/1472-6947-14-94.

[4] ALONSO-MARTÍN, F. and M. A. SALICHS: *Integration of a voice recognition system in a social robot*. *Cybernetics and Systems*, 42(4), pp. 215–245, 2011. doi:10.1080/01969722.2011.583593.

[5] VAJPAI, J. and A. BORA: *Industrial applications of automatic speech recognition systems*. *International Journal of Engineering Research and Applications*, 6(3), pp. 88 – 95, 2016.

[6] KËPUSKA, V. and G. BOHOUTA: *Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)*. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 99–103. 2018. doi:10.1109/CCWC.2018.8301638.

[7] MÖLLER, S.: *Quality of telephone-based spoken dialogue systems*. Springer Science & Business Media, 2004.

[8] TANGEN, S.: *Understanding the concept of productivity*. In *Proceedings of the 7th Asia-Pacific Industrial Engineering and Management Systems Conference, Taipei*, pp. 18–20. 2002.

[9] CLEMENTS-CROOME, D. and L. BAIZHAN: *Productivity and indoor environment*. In *Proceedings of Healthy Buildings*, vol. 1, pp. 629–634. 2000.

[10] CLEMENTS-CROOME, D.: *Consciousness, well-being and the senses*. In *Creating the productive workplace*, pp. 42–52. Taylor & Francis, 2006.

[11] FERRUCCI, D., E. BROWN, J. CHU-CARROLL, J. FAN, D. GONDEK, A. A. KALYANPUR, A. LALLY, J. W. MURDOCK, E. NYBERG, J. PRAGER, N. SCHLAEFER, and C. WELTY: *Building Watson: An overview of the DeepQA project*. *AI Magazine*, 31(3), pp. 59–79, 2010. doi:10.1609/aimag.v31i3.2303. URL `https://www.aaai.org/ojs/index.php/aimagazine/article/view/2303`.

[12] MARTIN, G. L.: *The utility of speech input in user-computer interfaces*. *International Journal of Man-Machine Studies*, 30(4), pp. 355 – 375, 1989. doi:10.1016/S0020-7373(89)80023-9.

[13] SMUCKER, M. D., J. ALLAN, and B. DACHEV: *Human question answering performance using an interactive document retrieval system*. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX 12, p. 3544. Association for Computing Machinery, New York, NY, USA, 2012. doi:10.1145/2362724.2362735.

[14] KRALJEVSKI, I., M. POHL, A. GJORESKI, U. KOLOSKA, J. WÖHL, M. WENZEL, and D. HIRSCHFELD: *Design and deployment of multilingual industrial voice control applications*. In P. BIRKHOLZ and S. STONE (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 205–211. TUDpress, Dresden, 2019.

[15] DEVLIN, J., M. CHANG, K. LEE, and K. TOUTANOVA: *BERT: Pre-training of deep bidirectional transformers for language understanding*. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805. 1810.04805`.

[16] ZHANG, X.: *A study of the effects of user characteristics on mental models of information retrieval systems*. Ph.D. thesis, National Library of Canada= Bibliothèque nationale du Canada, 1998.

[17] HANEDA, M., S. TANABE, and N. NISHIHARA: *The effect of traffic noise on productivity*. In *Healthy Buildings Conference Proceedings*, vol. 1, pp. 253–256. 2006.

[18] SEPPANEN, O., W. J. FISK, and Q. LEI: *Room temperature and productivity in office work*. In *Healthy Buildings Conference Proceedings*, vol. 1, pp. 243–247. 2006.

[19] HART, S. G. and L. E. STAVELAND: *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*. In P. A. HANCOCK and N. MESHKATI (eds.), *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139 – 183. North-Holland, 1988. doi:10.1016/S0166-4115(08)62386-9.