Is a visually expressive Virtual Assistant helpful in the kitchen?

Benedikt Boehlke
University of Regensburg
Regensburg, Germany
benedikt.boehlke@student.ur.de

Thilo Hohl
University of Regensburg
Regensburg, Germany
thilo.hohl@student.ur.de

Tobias Lanzl
University of Regensburg
Regensburg, Germany
tobias.lanzl@student.ur.de

Kevin Wach
University of Regensburg
Regensburg, Germany
kevin.wach@student.ur.de

ABSTRACT

This exploratory study delves into the ways we interact with Virtual Assistants today, which tasks they can be used in, and how we perceive them.

Assistants represented by a visual avatar have not been used too frequently; to research potential applications of assistants with faces, we conduct a wizard-of-oz study, in which users are guided through the steps of a recipe with the help of an assistant that looks like a chef and is able to express different emotions.

We found that most of our 13 participants respond positively to our avatar and would use a comparable system again. Analysis of their utterances throughout the study show that the conversation's content and information needs are quite human-like.

From a technical standpoint, the features simulated in our experiment are achievable with today's tools.

An assistant with a visual representation shows great promise and should be tested in various other contexts.

KEYWORDS

Conversational Search, IR, Anthropomorphism, Virtual Assistant, human-like avatar

1 INTRODUCTION

Virtual Personal Assistants, or VAs, have seen a surge in popularity in the last decade. Most major technology companies have released a VA of their own, starting with Siri, which was originally intended by the SRI as a tool to help in completing tasks instead of just searching for information [16]. After Apple acquired and subsequently shipped the system with their devices, competitors released Microsoft Cortana, Amazon Alexa and the Google Assistant over the next five years [4, 6, 11, 14]. These systems are able to answer basic QA tasks to users' satisfaction [12], but are not yet capable of participating in complex conversational search [22]. In this paper, we employ Radlinksi & Craswell's definition of conversational search [18]:

> "A conversational search system is a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user."

None of the aforementioned systems fully employ short- and long-term context and do not take initiative, but are instead activated by buttons or a certain passphrase.
The VAs are controlled by voice and answer by voice as well [4, 6, 11, 14]. Voice input proves most useful in multi-tasking environments where hands and eyes are occupied, like cooking, following an instructional tutorial or driving one's daily commute [24].
Both Amazon and Google also released devices with a slightly different output modality, containing the standard microphone-and-speaker combination, but also featuring a screen with several uses: It was able to more clearly display the status of the VA, show additional information or better visualise results as well as tell the time and show images and videos [21, 23].

This feature could also be used for an important facet of conversation not yet found in commercially available VAs: non-verbal cues. Non-verbal communication includes gestures and facial expressions and is essential to human dialogue [2]. In conversation, non-verbal cues can be used to signal misunderstandings and give feedback. Those aspects extend to a Virtual Assistant: A VA with a face or head could respond faster by showing its current status visually (e.g. thinking, awaiting response) [3] and furthermore, anthropomorphism in a personal virtual assistant is mentioned to increase user engagement and thereby the system's usefulness [5, 20].
For example, a loss of continuity on the VA's side could be communicated either through a long statement ("I am not sure where in the recipe you are right now. Could you help me back on track?") or displaying a 'confused' state in the avatar, which conveys this information much quicker.
The design of the VA's face is a key consideration regarding understanding and emotional connection. A cute face can help tolerate service failure [13], but might not suit every usage context. On the other hand, realistic-looking virtual agents are generally seen as more likeable, trustworthy, intelligent, respectful, calm, extroverted and positive [10].
In this paper, we want to investigate how a virtual assistant with a face and the ability to emote performs in comparison to a 'traditional' display-equipped VA. The context for this study is cooking a recipe, guided by the VA, as it is a complex task that requires the user's full visual attention most of the time. Our goal is to see if comfortable, human-like interaction using an avatar capable of producing facial expressions is more efficient by employing non-verbal cues and communication, but not distracting enough to warrant further developments regarding this context.

2 METHOD

Knowing our temporal and resource limitations, we opted to conduct a qualitative pilot study. We wanted to gain initial insights into the way people interact with a visual avatar in order to find justification for further studies.

2.1 Study Context

The kitchen is an interesting context for using virtual assistants, as it is a relatively closed system, extending maybe to a pantry, but has a great amount of complex contextual data, like user preferences or diets, and processes, such as guiding people through a recipe. VA systems tested in the kitchen could be expanded to work in related fields, like grocery shopping [1]. Related papers have tried several versions of chatbots in the kitchen. Common applications include searching for recipes, answering recipe questions and guiding users throughout the cooking process [19], helping reduce food waste [1, 9], and improving eating habits [1].
Rystedt and Zdybek were quite successful in their NLP-based voice-only approach, but feedback from test participants included suggestions for visualisations of the final dishes and certain recipe steps [19].
We chose to conduct our study in the kitchen and focused on guiding users through a recipe. The recipe we selected, ratatouille with rice-lentil patties, was a vegetarian meal with relatively low cost, high nutritional value, and it involved several different cooking techniques requiring at least intermediate skill.

2.2 Avatar Design

Given the scale of our study, we employed a relatively simple design for the avatar to avoid it falling into the uncanny valley, which dictates that robots that look about 75-90% humanlike are seen as less trustworthy [15]. Giving the assistant a cute, approachable appearance has also been the most frequent choice in related works [3, 13, 20]. We planned on only giving a face to our robot, in order to control the amount of variables, and because facial expressions play a larger role in communication understanding than gestures [17].
We selected five emotional states we wanted to model (see fig. 1). The different states, inspired by Shi et al. [20], were drawn and then animated in Krita. Animation included moving the mouth and eyes of our character in order to simulate talking, but we intentionally kept it quite stylized and low-fidelity to ensure a distinctly non-human design and therefore no risk of falling into the uncanny valley.

2.3 Prestudy

Before we conducted our study, we wanted to know whether our participants could identify our portrayal of those emotions correctly. In a pre-study, we showed six participants our five pictures and asked them to select the fitting emotion from a list of twelve options.
Here we found that most of our pictures were conveying the right emotion. Only the “curious” emotion wasn’t clear to the participants, so we tweaked the face a little bit and added question marks to better convey this emotion, leaning more into the cartoon-like aesthetic.

2.4 Participants

We had 13 participants (5 female, 8 male) aged 19–47 take part in our study. Nine of them were students, two were apprentices and two employees. We had one group of two cook together and the rest of them cooked alone. By doing so, we could also find out how a group interacts with the assistant. Since we just conducted a pilot-study, we used convenience-sampling to find participants. We also asked the participants about their use of virtual assistants, where we got varying answers. Four of them use assistants regularly, four of them rarely and five of them never use assistants.

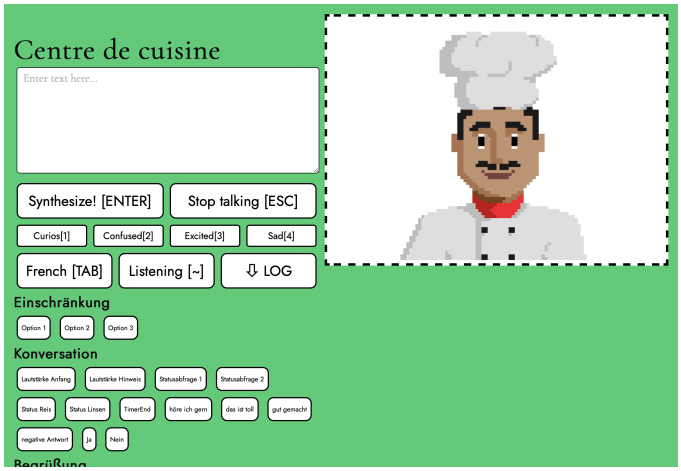

fig. 2: Wizard of Oz Interface.

2.5 Technical Implementation

We opted to use a Wizard-of-Oz design to research this idea, as it was the natural way to test the concept without creating a full voice-recognition-enabled assistant.
Our developers constructed a website to aid the Wizard in conducting our study (fig. 2).
It is currently hosted on the author’s university homepage:
https://homepages.uni-regensburg.de/~hot59097/CookHub_v1/
To keep interactions as consistent as possible, we decided to synthesise the voice of our assistant. SpeechSynthesis API can be found in all modern browsers and was originally envisioned to read out content for people with vision impairments.
On the right side of the screen a framed output window of the avatar is fixed in place. It is the part our participant sees while on the video call to the “program”.
The left portion of the screen is divided into three sections: a text input field, action buttons, and utterance shortcuts. When the `Synthesise` button is hit, text from the input field is read out while the avatar is animated. This process is refined by the preselection of emotion states which can be triggered by the number keys or by clicking the according buttons. If the avatar is told to speak, the selected emotion’s animation is played and then remains on screen for two further seconds after the synthesizer is done.
The utterance shortcuts are a quick way to find answers to participants’ questions and guarantee repeatability of the experiment across four different conductors. Utterances were sorted by recipe

Image					
State	Idle	Excited	Curious	Confused	Sad
Occurrence	Standard expression	User achieves something	Listening to user question	Loss of state	User rejects advice, User fails task

fig. 1: Emotional states and their according faces.

steps, containing the whole step, its substeps, and required ingredients. Ingredients for the whole recipe and their amounts were read out in the beginning of the study. In addition, options were given for greeting and introduction, general conversation ("yes", "no", positive and negative reactions, praise, and status requests). Another category of shortcuts were limitations, which were played whenever user requested information or actions the avatar was not capable of.

fig. 3: Study Setup.

2.6 Procedure

Before we started with the study, we gave the participants a list of the ingredients to buy them. The study was conducted in German, which was every participants' native language. For the actual experiment, we started a video call and invited our participant. After joining the call, we greeted them and instructed them to interact with the assistant as they would with a person showing them a recipe. We then shared the avatar screen and muted our microphones so the participants could only see and hear the avatar (see fig. 3).
Before they started the cooking process, a quick sound check was conducted, ensuring the participants could always hear the assistant and vice versa.
After that we started recording the call and started talking to the participants through the avatar by greeting them and telling them the list of ingredients they needed for the recipe. Then we let the participants start cooking by telling them the first step. From now on we answered all their questions and told them each step of the recipe while they were cooking. If they didn't interact with the avatar for about two minutes, we asked them if we could help them. In this way, our experiment contained mixed initiative, but the bulk of the conversation was driven by the user, as they had to fulfil tasks in their own time. After they finished the recipe, we concluded the cooking and stopped streaming the avatar after it said its goodbyes. Then we conducted a short interview with our participants, containing nine questions about the general experience, problem solving while cooking, the avatar's expressions, functionality and optional future usage. We closed the study with a disclaimer that the former interaction was not with a real program, but a Wizard-of-Oz experiment.

3 RESULTS

3.1 Information Needs

We transcribed the recordings of the experiments and created a coding-scheme for the annotation (see fig. 4). For the scheme we used the findings of Frummet et al.[7] and [8] and added some more categories that fit our needs.
The categories we added ourselves were *Conversation and Remarks*, *Gratitude*, *Step*, *Acknowledgement*, *Invalid* and *Repeat*.
Conversation and Remarks were the participants' statements which were not related to the cooking process or recipe, but were conversation or banter directed towards the assistant. (e.g. "Have you also eaten yet today?", "Do you know Siri?")
Gratitude contains all user statements thanking the avatar or the system in general. (e.g. "Thanks a lot!")
All user requests to move on to the next cooking step were labelled as *Step* (e.g. "What do I have to do next?")
Acknowledgement describes statements like "Okay" or "Yes".
Invalid contains all user quotes that didn't fit into any of the other categories (e.g. "I'm feeling like an Otto here!").
Repeat includes statements where the participant wanted to hear something again (e.g. "Can you say that again?").
The most frequent information need was to tell our Bot that its statements had been understood by the user. The next most popular was a question for the next step in the recipe, which can be expected, as it is the main mode of interaction with our system. It's nice to see that expressions of gratitude rank quite high in frequency, next to conversation and remarks, which show that the innate human need to communicate also applies to interactions with our assistant.

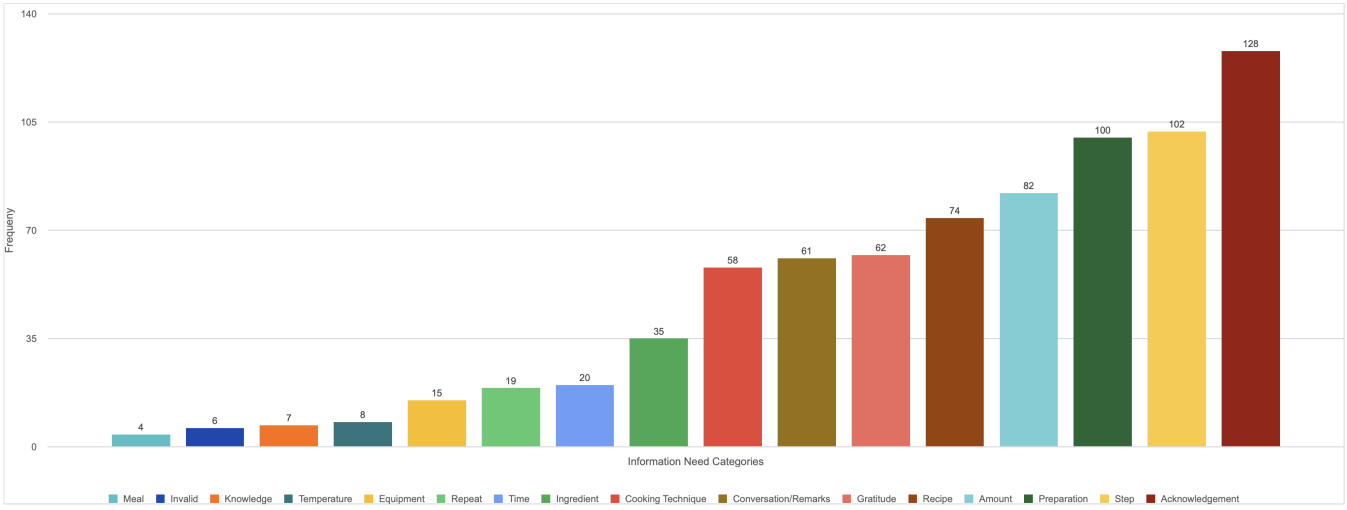

fig.4: Distribution of Information Needs.

3.2 Interview Results

Cooking with the assistant was received positively with little complaints. For many participants, it was a new and exciting experience and went more smoothly and quickly than expected.

Throughout the study, problems arose seldomly, the most prominent being the speed of the assistant's voice, which was too fast for some participants. Most problems could be resolved with the assistant repeating steps or details, but the rate of information should be reduced and the steps should be separated further for easy following.

Three-quarters of users were convinced by this text-free way of cooking a recipe, while one-quarter would have liked to see steps or key information presented alongside the avatar as text, which could be read quicker than asking.

Surprisingly, a majority of participants perceived our assistant as entertaining, which was never an intended feature, but a welcome one.

The avatar's face and its emotional display encountered mixed reception, with 80% appreciating the design, describing it as "friendly" and "good-looking", while 20% either didn't notice it or thought of it as an unnecessary gimmick.

Despite our Wizard interface, response times only seemed adequate to half of our users. This area requires some attention, as all of our Wizards stated being quite comfortable using the tool towards the end of our studies.

The given feedback towards the avatars functionality was mostly about the avatars visual design and the wish for even more gesture like a thumbs up, fireworks or similar.

Most users would use our system again given little improvements. Feature suggestions included the aforementioned text display, music player functionality, different voice selection and more extensive background knowledge.

4 DISCUSSION

As with most student studies, more time and resources could have improved this study. We would have loved to increase scale and conduct A/B tests to compare a voice-only assistant with our avatar-enabled version, but we are content with the qualitative insights gained in this format. We faced some problems with the sound synthesis in selected browsers across different operating systems, but they did not impact the study, as Wizards were able to work around them. For further optimisation of the avatar, possible implementations could contain a optional textual representation of recipe steps, as well as animations showing how to perform certain cooking techniques or more difficult steps. We considered different layouts, including text alongside the avatar, but decided against it to minimise variables.

5 CONCLUSION AND FUTURE WORK

In conclusion we can say that our experiment fulfilled its purpose of clarifying whether further studies in this area will be reasonable or not, showing that most of the participants indeed enjoyed being guided through cooking a meal by a visual voice avatar. Even though around 50% of the participants would have wished for more regular or faster answer times and even more than half would have preferred even more avatar expressions like gestures or screen effects, the wide majority of participants rated the experience very positively and would use it again in the future. They also responded quite positively to the avatar design in its distinct, pixel-art style.

The most frequent information needs and especially the high frequency of conversational elements support our own observation that, apart from the timing issue, contents and intentions of the conversation with our assistant are quite natural and human-like.

As expected, almost all of the participants suggested the possibility of choosing from a variety of different recipes, even taking another step forwards and proposing a functionality for the avatar that helps plan trips to the grocery store.

On a technical level, the features simulated in our Wizard-of-Oz-design and the ones suggested by users are quite achievable given appropriate funding.

In summary, further research would be interesting. Exploring different contexts with more participants could show new applications.

REFERENCES

[1] Prashanti Angara, Miguel Jiménez, Kirti Agarwal, Harshit Jain, Roshni Jain, Ulrike Stege, Sudhakar Ganti, Hausi A Müller, and Joanna W Ng. 2017. Foodie fooderson a conversational agent for the smart kitchen.. In *CASCON*. 247–253.

[2] Dane Archer and Robin M Akert. 1977. Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of personality and social psychology* 35, 6 (1977), 443.

[3] Aryel Beck, Lola Cañamero, and Kim Bard. 2010. Towards an Affect Space for robots to display emotional body language. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 464 – 469. https://doi.org/10.1109/ROMAN.2010.5598649

[4] Jez Corden. 2017. A brief history of Cortana, Microsoft's Trusty Digital assistant. https://www.windowscentral.com/history-cortana-microsofts-digital-assistant

[5] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[6] Darrell Etherington. 2014. Amazon Echo is a $199 connected speaker packing an always-on Siri-style assistant. https://techcrunch.com/2014/11/06/amazon-echo/

[7] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. (2019).

[8] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What Can I Cook with these Ingredients?"-Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–32.

[9] Elena Gong, Nanxi Li, and Yang Yue. [n.d.]. SmolKat: A Smart Kitchen Assistant. ([n. d.]).

[10] Jennifer Hyde. 2013. *Design of avatars as conversational partners.* Ph.D. Dissertation. Trinity College.

[11] SRI International. 2020. The history of Apple's Siri. https://www.sri.com/hoi/siri/

[12] Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 241–250.

[13] Xingyang Lv, Yue Liu, Jingjing Luo, Yuqing Liu, and Chunxiao Li. 2021. Does a cute artificial intelligence assistant soften the blow? The impact of cuteness on customer tolerance of assistant service failure. *Annals of Tourism Research* 87 (2021), 103114. https://doi.org/10.1016/j.annals.2020.103114

[14] Matthew Lynley. 2016. Google unveils Google assistant, a virtual assistant that's a big upgrade to google now. https://techcrunch.com/2016/05/18/google-unveils-google-assistant-a-big-upgrade-to-google-now/

[15] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[16] Erica Naone. 2009. TR10: Intelligent software assistant. *Technology Review, Mar.-Apr* (2009).

[17] Marieke Peeters, Vivian Genaro Motti, Helena Frijns, Siddharth Mehrotra, Tugce Akkoc, Sena Büşra Yengec, Oguz Calik, and Mark Neerincx. 2016. Design and development of a physical and a virtual embodied conversational agent for social support of older adults. In *eNTERFACE 2016: Summer workshop*. CTIT, 21.

[18] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[19] Beata Rystedt and Mia Zdybek. 2018. Conversational agent as kitchen assistant.

[20] Yang Shi, Xin Yan, Xiaojuan Ma, Yongqi Lou, and Nan Cao. 2018. Designing emotional expressions of conversational states for voice assistants: Modality and engagement. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[21] Google Support. 2021. Google Nest and home device specifications. https://support.google.com/googlenest/answer/7072284#zippy=%2Cgoogle-nest-hub

[22] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2187–2193.

[23] May Walton. 2017. Amazon Echo Show: Alexa-powered touchscreen speaker launches June 28. https://arstechnica.com/gadgets/2017/05/amazon-echo-show-price-specs-release-date/

[24] Jennifer Zamora. 2017. Rise of the chatbots: Finding a place for artificial intelligence in India and US. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*. 109–112.