Titel: Digitale SprachassistentInnen – Ein Reality Check

Einleitung: Digitale SprachassistentInnen, oft auch als *Intelligent Personal Assistants*, *Mobile Assistants*, *Virtual Personal Assistants*, oder *Voice Assistants* bezeichnet [1], fallen in die Kategorie der Natural User Interfaces, da sie im Gegensatz zu Maus und Tastatur einen für den Menschen natürlichen Kanal (i.e. die menschliche Sprache) zur Interaktion mit der Technologie verwenden [2]. Ihr Einsatz hat in den letzten Jahren zunehmend an Popularität gewonnen, wobei Lösungen durchwegs von Technologiegiganten wie beispielsweise Apple, Amazon oder Google getragen werden. Das zunehmende Interesse an sprach-basierter Interaktion ist aber vor allem auch dadurch begründet, dass signifikante Fortschritte bei den den Systemen zugrunde liegenden Technologien, wie der Spracherkennung (*Speech Recognition*), der Sprachproduktion und -ausgabe (*Natural Language Generation* & *Speech Synthesis*), oder dem Sprachverständnis (*Natural Language Understanding*), mittlerweile einen produktiven und weitgehend fehlerfreien Einsatz zulassen. Zumindest wird dieser Eindruck von den jeweiligen Systemanbietern vermittelt, welche ihre Produkte entweder eingebettet in Smartphones oder mittels eigenständigen „smarten" Lautsprechern vermarkten. Dabei wird insbesondere die Nutzung der natürlichen Sprache als eine intuitive Form der Mensch-Maschinen Interaktion hervorgehoben. Ziel dieser Arbeit war es daher, diese von den Herstellern beworbenen Fähigkeiten aktueller Systeme in einem experimentellen Setting unter realen Bedingungen zu vergleichen. Als leitend galt also die Frage nach den wirklich gebotenen Möglichkeiten.

Methode: Mit Hilfe von 32 englischsprachigen StudienteilnehmerInnen (allesamt US English Native Speakers; 8 davon weiblich; durchschnittliches Alter: 18,34 Jahre; SD=0,73) wurde die Performance von vier digitalen SprachassistentInnen in den Kategorien Informationsbeschaffung (i.e. allgemeine Informationen, sportbezogene Informationen, ortsbezogenen Informationen) und Kalenderbedienung evaluiert. Bei den getesteten Systemen handelte es sich um Apple's *Siri*, Microsoft's *Cortana*, Amazon's *Alexa* und Google's *Assistant*. Die Systeme wurden vor Beginn jedes Experiments auf ihre Werkseinstellungen zurückgesetzt um eine Beeinflussung durch eine vorherige Nutzung zu vermeiden. StudienteilnehmerInnen hatten jeweils vier Anwendungsszenarien mit wiederum jeweils zwei Aufgaben zu erfüllen. Jedes Szenario wurde mit einem anderen System durchgeführt. Die Reihenfolge der Szenarien, sowie die den Szenarien zugeteilten Systeme, wurden dabei variiert, um Reiheneffekte zu vermeiden. Im Szenario *Informationen und Neuigkeiten im Sport* wurden zum einen Information zu einem zurückliegenden Event, und zum anderen das Ergebnis eines vor kurzem stattgefundenen Football Spiels abgefragt. Das Szenario *Allgemeine Informationen* enthielt die Definition eines Begriffes sowie eine Wissensfrage. Beim Szenario *Lokale Informationen* wurde die Fahrtdauer zu einem nahegelegenen Ort abgefragt bzw. nach dem nächstgelegenen Ort in einer bestimmten Kategorie gesucht. Im Szenario

Kalenderaktionen mussten StudienteilnehmerInnen schließlich einen Kalendereintrag erstellen, bzw. Informationen zu einem anderen Kalendereintrag abrufen. Um den Vergleich der Systeme zu systematisieren wurde sowohl das korrekte Erkennen der Spracheingabe evaluiert (Note: auf die Berechnung einer exakten Word Error Rate wurde verzichtet, vielmehr wurde evaluiert, ob die Intention der Eingabe erkannt wurde), als auch die korrekte Erfüllung der Aufgabe getestet. Die Durchführung der Experimente fand in einem eigens dafür bereitgestellten Versuchsraum statt, welcher weitestgehend von Hintergrundgeräuschen abgeschirmt wurde.

Ergebnisse: Unsere Ergebnisse zeigen, dass das Erkennen von Spracheingaben (*i.e. Speech Recognition*) bei allen getesteten digitalen SprachassistentInnen bereits auf sehr hohem Niveau liegt. Sowohl Google (92%) als auch Microsoft (96%) und Apple (94%) hatten über 90% der Eingaben korrekt erkannt. Nur Amazon's *Alexa* war mit 86% unter dieser Erfolgsquote. Die Ergebnisse zeigen jedoch auch, dass das korrekte Erkennen der Spracheingabe nicht zwingend zu einer korrekten Aufgabenbearbeitung mit entsprechender Sprachausgabe führt. So konnte Google's *Assistant* mit 87% richtig beantworteten Fragen (i.e. korrekt generierten Sprachausgaben) das beste Gesamtergebnis erzielen. Microsoft's *Cortana* schaffte dies in 71% und Amazon's *Alexa* in nur 61% der Fälle. Apple's *Siri* nahm hier den letzten Platz ein, da nur bei 56% der Versuche ein richtiges Sprachfeedback generiert werden konnte. Dabei ist jedoch hervorzuheben, dass *Siri* zwar oftmals korrekte Ergebnisse lieferte (81% der Fälle), jedoch diese nicht als entsprechendes Sprachfeedback wiedergeben konnte.

Diskussion/Conclusio: Unsere Untersuchung beschäftigte sich mit der Performance aktueller digitaler SprachassistentInnen unter reellen Bedingungen. Die Ergebnisse zeigen, dass die Spracherkennung, also die Umformung einer akustischen Welle in Text, bei den getesteten Systemen auf durchwegs hohem Niveau liegt, das wahre Verständnis dieses Textes und folglich die sprachbasierte Bearbeitung von Aufgabenstellungen dieses Level jedoch nicht erreichen konnte. Zur Relativierung dieses Ergebnisses muss allerdings hervorgehoben werden, dass aktuelle Systeme laufend mit neuen Daten versorgt werden, was zu einer konstanten Verbesserung des Dialogmanagements aber auch des eigentlichen Sprachverständnisses führt.

Quellen:

[1] Gangemi, A., Leonardi, S., and Panconesi, A. (Eds.) 2015. Proceedings of the 24th Inter. Conference on World Wide Web - WWW '15. New York, New York, USA: ACM Press.

[2] López, G., Quesada, L., and Guerrero, L. A. (2018). Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In I. L. Nunes (Ed.), Advances in Human Factors and Systems Interaction (pp. 241–250). Cham: Springer International Publishing.