DER FAKTOR MENSCH IN DER MENSCH-MASCHINE-INTERAKTION

Daniel Duran, Sarah Warchhold

Albert-Ludwigs-Universität Freiburg
{daniel.duran , sarah.warchhold}@germanistik.uni-freiburg.de

Kurzfassung: In diesem Beitrag legen wir die Motivation für unser laufendes Forschungsprojekt zur Mensch-Maschine-Interaktion (MMI) dar. Zentrale Frage ist, ob eine natürliche, spontansprachliche Ausdrucksweise der Maschine die MMI verbessert. Dies könnte durch eine geringe kognitive Beanspruchung der Benutzer_innen erreicht werden. In einer ersten Vorstudie testeten wir die Wahrnehmung manipulierter natürlicher Stimmen. Zunächst mit Standardsprache durchgeführt, möchten wir diese Vorstudie mit dialektalen Stimmen wiederholen. Die ersten Ergebnisse zeigen keine wesentlichen Unterschiede zwischen unseren beiden Erwachsenen und der kindlichen Stimme. Sprachsignalmanipulationen der Intonation (bis hin zu einer flachen Pitchkontur) wurden im Vergleich zu spektralen und segmentalen Manipulationen am stärksten als synthetisch wahrgenommen. Die Ergebnisse dieser beiden Vorstudien sollen dann in ein Wizard-of-Oz-Experiment einfließen, in dem wir mit simulierter MMI das Vertrauen in die Stimme einer KI oder eines Agenten untersuchen.

1 Einordnung und Motivation

Wir präsentieren in diesem Beitrag unser laufendes Forschungsprojekt „Der Faktor Mensch in der Mensch-Maschine-Interaktion“, das Aspekte der Computerlinguistik, Psycholinguistik und Kognitionswissenschaft verbindet. Aus einer lingustischen, auf den Menschen fokussierten Perspektive, untersuchen wir grundlegende Mechanismen der menschlichen Sprachwahrnehmung und Sprachproduktion im Rahmen der gesprochensprachlichen Interaktion mit künstlichen Agenten (digitalen Sprachassistenzsystemen). Im Folgenden gehen wir zunächst auf den Stand der Forschung ein und skizzieren, wo unser Projekt ansetzt. Im Abschnitt 2 präsentieren wir Ergebnisse eines ersten Online-Experiments zur Wahrnehmung unterschiedlicher Stimmen. Wir schließen in Abschnitt 3 mit einer kurzen Diskussion der bisherigen Ergebnisse und einem Ausblick auf zentrale Forschungsfragen des Projekts.

Sprachassistenzsysteme (SAS) sind inzwischen allgegenwärtig. Über Nutzungsmuster von SAS in Deutschland berichten u.a. Tas & Arnold [1, S.11] basierend auf einer Befragung von 3184 Konsument_innen im Jahr 2018. Da bereits 26% zu diesem Zeitpunkt angaben einen der verfügbaren Sprachassistenten (Amazons *Alexa*, Apples *Siri*, Googles *Assistant*, Microsofts *Cortana* oder Samsungs *Bixby*) zu nutzen, ziehen die Autor_innen eine Parallele zur Adaptionsrate von Smartphones. Diese entspräche etwa 5 Jahre nach der Einführung erster Modelle. 85% der Befragten hatten außerdem schon ein Gerät mit vorinstalliertem SAS zuhause, 11% sogar einen Smart-Speaker, was laut den Autor_innen einen sprunghaften Anstieg der Adaptionsrate ermögliche. Diesen halten sie allerdings nicht für wahrscheinlich, da die Nutzung sich bisher auf einige wenige Funktionen wie die Ansage des Wetters oder aktueller Sportergebnisse beschränke und selten vorkomme.

In der Forschung zu SAS liegt der Fokus generell oft auf maschinenbezogenen, technischen Aspekten im Bereich der Künstlichen Intelligenz und Informatik (siehe [2, 3]). Wir untersuchen

den Faktor Mensch mit einem Schwerpunkt auf linguistischen, psychologischen, kognitiven und sozialen Aspekten. Die zentrale Frage ist, ob eine natürliche, spontansprachliche Ausdrucksweise der Maschine die MMI verbessert. Dies könnte durch eine geringe kognitive Beanspruchung der Benutzer_innen erreicht werden, wenn diese sich nicht speziell an die technischen Anforderungen der Maschine anpassen müssen, sondern natürlich verbal interagieren können. Die SAS-Entwicklung verfolgt das Ziel, den Benutzer_innen eine möglichst natürliche gesprochensprachliche Interaktion zu ermöglichen. Dabei wird menschliches Kommunikations- und Interaktionsverhalten nachgeahmt, indem auch immer natürlicher klingende Stimmen für die Sprachsynthese entwickelt werden[1]. Grundlage hierfür ist bisher hauptsächlich die Standardsprache, die zwar von den meisten Sprecher_innen problemlos verstanden werden kann, aber von vielen gar nicht oder nicht einheitlich gesprochen wird [4, 5]. Bereits 2008 gab es erste Versuche, dialektale Varietäten in Text-to-Speech-Synthese zu übersetzen [6]. Auf Seiten der Spracherkennung wird in der Schweiz seit etwa einem Jahr daran gearbeitet, einem Sprachassistenzsystem auch schweizerdeutsche Mundart beizubringen (siehe [7]).

Die Evaluation und Bewertung solcher Assistenzsysteme erfolgt meist indirekt durch die objektive Messung des Erfolgs in der Aufgabenerfüllung oder durch subjektive Benutzerbewertungen mit Hilfe von (standardisierten) Fragebögen. Die Auswirkung der Agenten auf die Sprachwahrnehmung, das sprachliche Verhalten der Nutzer_innen oder die kognitivie Beanspruchung während der Interaktion rückt zunehmend in den Fokus aktueller Forschungsarbeiten, in die wir uns einreihen möchten. Trotz der großen technologischen Fortschritte der letzten Jahre gibt es noch zahlreiche ungelöste Probleme, wie z.B. im Bereich der Spracherkennung. Unter Alltagsbedingungen mit Hintergrundgeräuschen oder durcheinander sprechenden Personen verursacht die automatische Spracherkennung auch heute teilweise noch Probleme. Eine praktische Konsequenz in der MMI ist daher, dass Nutzer_innen ihre Sprache oft an die Maschine anpassen, um verstanden zu werden (siehe auch [8, 9]). Vor allem die *Art* der sprachlichen Interaktion ist daher von Interesse und die Frage, welche Interaktion der MMI als Vorbild dienen kann und soll. Beispiele finden sich in *child-directed* [10], *pet-directed* [11] aber auch in *machine-directed speech* [12].

Auch in der Interaktion zweier oder mehrerer Menschen untereinander findet eine automatisierte Anpassung (*Konvergenz* oder *Akkomodation* genannt) z.B. an Dialekte (siehe [13]) oder fremdsprachliche Akzente (siehe [14]) statt. Diese Anpassung verringert die wahrgenommene Distanz zwischen den Gesprächstpartner_innen und verbessert das gegenseitige Verständnis. Hierdurch wird die kognitive Beanspruchung verringert. Gemäß der *Communication Accomodation Theory*, die Giles und Kolleg_innen seit den 1970ern entwickeln [15], kann die Anpassung an Dialogpartner_innen bewusst zur Verringerung der Distanz eingesetzt werden, aber auch zum Gegenteil – einer Abgrenzung durch sprachliche Distanzierung (*Divergenz*). Ein zentraler Aspekt der Akkomodation in frühen Studien war die Interaktion mehrsprachiger Gesprächspartner_innen. Dieser Punkt wurde in folgenden Studien zur Akkomodation lange Zeit nicht betrachtet und soll innerhalb des hier vorgestellten Projekts wieder aufgegriffen werden. Wir gehen dabei der Hypothese nach, dass eine Anpassung der Maschine an die große Variabilität in der alltäglichen Sprachnutzung von Menschen, die auch unterschiedliche Dialekte und Sprachen einschließt, die MMI im Sinne einer effizienteren und natürlicheren Interaktion verbessern kann.

[1] Das Problem der Dialogsteuerung (*turn-taking*, Rezipientensignale der Maschine etc.) soll hier nicht weiter berücksichtigt werden, auch wenn dies – neben der Wahrnehmung phonetisch-prosodischer Merkmale einer synthetischen Stimme, die hier im Vordergrund steht – die Wahrnehmung von Äußerungen eines künstlichen Agenten beeinflusst.

2 Erste Experimente

In einer Vorstudie wurde die wahrgenommene *Natürlichkeit* verschiedener Stimmen anhand bearbeiteter Aufnahmen experimentell getestet [16]. In Forschungsarbeiten werden verwendete synthetische Stimmen kaum näher beschrieben, sodass unklar ist, inwiefern sich diese Stimmen akustisch-phonetisch von natürlicher Sprache unterscheiden. Daher haben wir getestet, welche Eigenschaften als synthetisch wahrgenommen werden und welche nicht.

Sprachmaterial und Manipulationen: Wir manipulierten Aufnahmen standardsprachlich[2] eingelesener Nonsense-Sätze von zwei Erwachsenen deutschen Muttersprachler_innen (*f*=weiblich und *m*=männlich) und einem Kind (*c*) auf spektraler, segmentaler und Intonationsebene. Alle Manipulationen wurden mit Praat vorgenommen [17]. Semantisch unvorhersagbare bzw. *sinnlose* Sätze, wie etwa Chomsky's oft zitiertes Beispiel „Colorless green ideas sleep furiously" [18], werden traditionell zur Erfassung der Verständlichkeit (*intelligibility*) oder der Qualität von Sprachsynthesesystemen verwendet (vgl. [19, 20]). Durch die Verwendung semantisch leerer Stimuli soll die Aufmerksamkeit der Proband_innen von der Bedeutung auf die Form gelenkt werden. Auf spektraler Ebene veränderten wir den Schwerpunkt des Spektrums und blendeten das ausgeschnittene Segment unterschiedlich gut in das Sprachsignal ein (2 Stufen). Dafür wurden die Spektren der Frikativsegmente [s] und [z] mit einem Bandpass im Bereich 300–8000 Hz herausgefiltert. In der Manipulationsstufe *f2* wurde das gefilterte Signal mit dem Faktor 0.2, und in der Manipulationsstufe *f3* mit dem Faktor 0.8 multipliziert und entsprechend mit dem ursprünglichen Signal wieder zusammengefügt. Dadurch wurde das *centre of gravity* des Spektrums dieser Frikative verschoben. Segmental extrahierten wir aus einem im Experiment nicht verwendeten Satz des/der jeweiligen Sprecher_in, ein Schwa [ə] und die Artikel *der*, *die*, *das* und fügten sie in unterschiedlicher Anzahl in die Nonsensesätze ein (2 Stufen). In der Manipulationsstufe *s3* wurde jedes dieser Segmente im Stimulussatz ausgetauscht, und in der Manipulationsstufe *s2* nur jedes zweite. Eine Anpassung der Intensität oder der Prosodie an die Zielumgebung wurde bewusst nicht vorgenommen. Für die intonatorische Manipulationsebene führten wir für die erste Manipulationsstufe *p1* in Praat eine einfache Resynthese (PSOLA) des Sprachsignals mit der ursprünglichen Intonationskontur durch. Weiterhin wurde die Intonationskurve leicht abgeflacht (Höhe- und Tiefpunkt näher an den Mittelwert gezogen, Stufe *p2*) und zuletzt monotonisiert (auf den Mittelwert geglättet, Stufe *p3*). Mit der Auswahl der Manipulationsarten wollten wir gängige Artefakte in der Sprachsynthese testen, wie beispielsweise die fehlerhafte Verkettung eines Sprachsignals oder die mangelhafte Wiedergabe spektraler Eigenschaften. Alle Manipulationen wurden isoliert vorgenommen. So entstanden insgesamt 288 Stimuli (3 Sprecher_innen, 12 Sätze, 8 Varianten).

Da Erfahrung mit einer synthetischen Stimme zu Gewöhnungseffekten führt [21], haben wir selbst Stimuli erzeugt, statt ein in Anwendungen verbreitet eingesetztes System zu verwenden. Die Perzeption synthetisierter Sprache hängt von verschiedenen Faktoren wie den akustisch-phonetischen Eigenschaften des Sprachsignals, den kognitiven Anforderungen der Testaufgabe oder der Erfahrung mit synthetischen Stimmen ab [22]. Mit unseren Stimuli sollte gezielt getestet werden, welchen Einfluss Sprecher_innenidentität (weiblich, männlich oder Kind) und unterschiedliche akustische Merkmale auf die Wahrnehmung von Natürlichkeit bzw. Künstlichkeit haben.

Teilnehmer_innen und Ablauf: Die mit Psychopy [23] implementierte Onlinestudie wurde auf der Plattform Pavlovia[3] durchgeführt. In einer passiven Perzeptions- bzw. Bewertungsaufgabe wurden die Stimuli auf einer 5-Punkte-Skala von *eher natürlich* bis *eher synthetisch* bewertet. Unter den 27 Proband_innen war eine Person mit einer Hörbeeinträchtigung. Daher

[2] Die Vorstudie soll noch mit Dialektvariation durchgeführt werden.

[3] `https://pavlovia.org/`

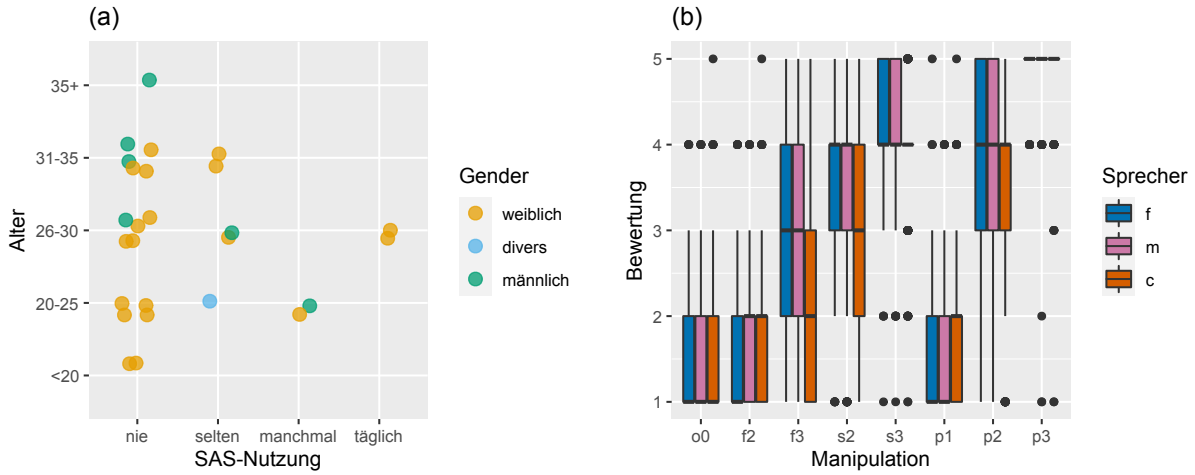

Abbildung 1 – (a) SAS-Nutzung nach Alter und Gender (Proband_innen). Die Option *täglich (lang)* wurde nie gewählt und ist hier nicht dargestellt. (b) Bewertung nach Manipulationsarten und Sprecher_innen von 1=*eher natürlich* bis 5=*eher synthetisch*. *o0* bezeichnet die unveränderten Aufnahmen.

wurden nur die Daten von 26 Proband_innen ausgewertet. Darunter waren 19 weibliche, sechs männliche und eine nicht-binäre Person. Alle haben Deutsch als Muttersprache erlernt. Zwei Personen gaben an, unter 20 zu sein und eine Person war älter als 35. Zur weiteren Anonymisierung haben wir das Alter für unsere Hauptzielgruppe (20 bis 35) in Gruppen von 5 Jahren zusammengefasst. So fallen in die Gruppe der 20 bis 25 Jährigen sieben Personen, in die Gruppe der 26 bis 30 Jährigen neun und in die Gruppe der 31 bis 35 Jährigen ebenfalls sieben Personen. 16 Personen gaben als höchsten Bildungsabschluss ein Studium an, sieben einen Gymnasialabschluss, zwei einen Realschulabschluss und eine Person einen Hauptschulabschluss. 11 Proband_innen hatten einen linguistischen Hintergrund in Ausbildung oder Studium. Von unseren ausgewerteten Proband_innen gaben 17 an, nie Sprachassistenen zu nutzen, fünf selten und zwei manchmal. Lediglich zwei Personen gaben an, täglich für kurze Zeit einen Sprachassistenten zu benutzen. Die fünfte Antwortmöglichkeit *täglich (lang)* gab niemand an, vgl. Abbildung 1 (a). Im Einleitungstext der Onlinestudie baten wir darum, Kopfhörer zu benutzen. 6 Personen gaben an, dennoch keine Kopfhörer benutzt zu haben.

Ergebnisse und Limitation: Die spektralen Manipulationen der Frikative wurden insgesamt eher als natürlich bewertet, die segmentalen eher als synthetisch. Die Manipulation der Intonationskontur zeigte die größte Bandbreite an Bewertungen, vgl. Abbildung 1 (b). Für die statistische Auswertung modellierten wir die Bewertung in Abhängigkeit von Sprecher_in , Manipulationsart und anderen Faktoren (siehe Tabelle 1) mit *linear mixed models* (in R [24] mit den Paketen *lme4* [25] und *lmerTest* [26]). Mit *Backward Elimination* (Funktion *step*) erstellten wir ein finales Modell dessen Koeffizienten in Tabelle 1 zusammengefasst sind (Random effects sind die Probanden ID, Trial und Satz Nummer). Der AIC-Wert des finalen Modells ist zwar kleiner als der des größeren Startmodells, der Modellvergleich mit ANOVA zeigt aber einen nicht signifikanten Wert von 0.128 für $\Pr(> \chi^2)$. Wir sehen positive Estimates für alle Manipulationsstufen, das höchste Estimate für *p3* mit 3.328, das kleinste für *p1* mit 0.09. Die Reaktionszeit zeigt einen kleinen, aber signifikanten Effekt. Bei größerer Reaktionszeit tendierten die Proband_innen eher dazu, die Stimuli natürlicher zu bewerten. Männliche Probanden haben im Gegensatz zu den weiblichen die Stimuli eher natürlich bewertet. Die Stimuli wurden bei stärkerer SAS-Nutzung eher als synthetisch bewertet. Möglicherweise werden bei häufigerer SAS-Nutzung akustische Artefakte eher als Hinweis auf Synthese interpretiert. Zudem hatten wir nur wenige Teilnehmende, die SAS häufiger nutzen. Auch der Nonsensegehalt der Sätze war nicht gleichmäßig. Ein generelles Problem bei dieser statistischen Auswertung ist, dass die Bewertungen nicht normalverteilt sind. Dies war uns von vorneherein klar, da wir auch an den

	Estimate	Std. Error	df	t value	Pr($>\|t\|$)
(Intercept)	1.491	$7.192 \cdot 10^{-2}$	$9.757 \cdot 10^{1}$	20.727	$< 2 \cdot 10^{-16}$
speaker[m]	$5.006 \cdot 10^{-3}$	$4.226 \cdot 10^{-2}$	$2.566 \cdot 10^{3}$	0.118	0.905716
speaker[c]	$-1.902 \cdot 10^{-1}$	$4.244 \cdot 10^{-2}$	$2.454 \cdot 10^{3}$	-4.480	$7.80 \cdot 10^{-6}$
man[f2]	$1.451 \cdot 10^{-1}$	$5.920 \cdot 10^{-2}$	$1.806 \cdot 10^{3}$	2.452	0.014311
man[f3]	1.172	$5.765 \cdot 10^{-2}$	$2.235 \cdot 10^{3}$	20.324	$< 2 \cdot 10^{-16}$
man[s2]	1.954	$5.951 \cdot 10^{-2}$	$1.672 \cdot 10^{3}$	32.829	$< 2 \cdot 10^{-16}$
man[s3]	2.552	$5.890 \cdot 10^{-2}$	$1.850 \cdot 10^{3}$	43.322	$< 2 \cdot 10^{-16}$
man[p1]	$9.222 \cdot 10^{-2}$	$5.889 \cdot 10^{-2}$	$1.861 \cdot 10^{3}$	1.566	0.117519
man[p2]	2.137	$5.853 \cdot 10^{-2}$	$1.945 \cdot 10^{3}$	36.511	$< 2 \cdot 10^{-16}$
man[p3]	3.328	$5.841 \cdot 10^{-2}$	$1.989 \cdot 10^{3}$	56.970	$< 2 \cdot 10^{-16}$
RT_{rel}	$-8.411 \cdot 10^{-3}$	$2.271 \cdot 10^{-3}$	$4.744 \cdot 10^{3}$	-3.703	0.000215
gender[divers]	$2.243 \cdot 10^{-1}$	$2.343 \cdot 10^{-1}$	$3.258 \cdot 10^{1}$	0.957	0.345465
gender[männlich]	$-2.182 \cdot 10^{-1}$	$9.098 \cdot 10^{-2}$	$3.890 \cdot 10^{1}$	-2.398	0.021359
SAS[selten]	$2.734 \cdot 10^{-1}$	$9.386 \cdot 10^{-2}$	$2.555 \cdot 10^{1}$	2.913	0.007331
SAS[manchmal]	$1.488 \cdot 10^{-1}$	$1.519 \cdot 10^{-1}$	$2.618 \cdot 10^{1}$	0.979	0.336305
SAS[täglich]	$6.316 \cdot 10^{-1}$	$1.524 \cdot 10^{-1}$	$2.603 \cdot 10^{1}$	4.144	0.000321
speaker[m]:gender[divers]	$1.739 \cdot 10^{-1}$	$1.508 \cdot 10^{-1}$	$4.765 \cdot 10^{3}$	1.154	0.248670
speaker[c]:gender[divers]	$-7.418 \cdot 10^{-2}$	$1.513 \cdot 10^{-1}$	$4.764 \cdot 10^{3}$	-0.490	0.624039
speaker[m]:gender[männlich]	$1.355 \cdot 10^{-1}$	$6.793 \cdot 10^{-2}$	$4.618 \cdot 10^{3}$	1.995	0.046130
speaker[c]:gender[männlich]	$3.215 \cdot 10^{-1}$	$6.798 \cdot 10^{-2}$	$4.618 \cdot 10^{3}$	4.730	$2.32 \cdot 10^{-6}$

Tabelle 1 – LMER Analyse: Fixed effects. *response* steht für die Bewertung von 1 bis 5, *man* steht für die Manipulationsart, RT_{rel} ist die relative Raktionszeit (gemessen ab Ende des Stimulus). Formel des resultierenden Modells (in R-Notation): response ~ speaker + man + RT_{rel} + gender + SAS + (1|participant) + (1|sentence) + (1|trial) + speaker:gender

Extremen interessiert sind und eine durchschnittliche Bewertung in der Mitte (3) uninteressant wäre. Die Sprecher_innenidentität scheint keine große Rolle zu spielen. Tendenziell wurde die Kinderstimme eher als natürlich bewertet, auch wenn der Effekt nur klein ist. Wir hatten für die Kinderstimme einen (deutlicheren) Unterschied zu den Erwachsenen und insbesondere zur SAS-typischen weiblichen Stimme erwartet.

3 Diskussion und Ausblick

Bereits seit längerem wird dafür plädiert, die Bewertung von Sprachsynthesesystemen mit Hilfe einer interaktiven, „realistischeren“, d.h. alltagsnahen Aufgabe durchzuführen, anstelle von passiven Perzeptions- oder Bewertungsaufgaben [19, 21]. Im nächsten Schritt wollen wir daher in einem spielbasierten Wizard-of-Oz-Experiment[4] mit simulierter MMI das Vertrauen in die Stimme einer KI oder eines Agenten/Avatars untersuchen. Die verwendete Stimme soll dabei zwischen den in den Vorstudien ermittelten Extremen zwischen *natürlich* und *synthetisch* variieren. Durch eine kollaborative Aufgabenstellung wollen wir auch Sprachproduktionsdaten erheben und gleichzeitig die Interaktion mit dem System betrachten. Konkret möchten wir in der Studie eine Quizrunde mit einer Navigationsaufgabe kombinieren. So soll die Vertrauenswürdigkeit verschiedener synthetischer Stimmen getestet werden. Ein Indiz für Vertrauen in die Stimme/die KI könnte ganz basal die Befolgung von Ratschlägen im Rahmen der Aufgabe sein (vgl. [28]). Wir stellen die Hypothese auf, dass nicht-standardsprachliche Stimmen einer

[4] Aufgrund der anhaltenden Pandemie mit Kontaktbeschränkungen wurde bereits die Vorstudie als Onlineexperiment durchgeführt, was wahrscheinlich auch unsere weiteren Experimente betrifft. Da wir hierdurch weniger experimentelle Kontrolle haben, wollen wir in Zukunft so weit wie möglich sicherstellen, dass die Experimente unter vergleichbaren Bedingungen zuhause durchgeführt werden. Hierzu wollen wir den Experimenten das von [27] vorgeschlagene Kopfhörerscreening voranstellen.

Maschine die Erwartungshaltung des Menschen brechen und sich dadurch auf die kognitive Beanspruchung der Benutzer_innen auswirken. Bei als natürlicher wahrgenommenen Stimmen erwarten wir eine reduzierte kognitive Beanspruchung was die Interaktion mit SAS verbessert (z.B. im Hinblick auf die Effizienz von Aufgabenerfüllungen oder auf das Vertrauen der Benutzer_innen in die Maschine).

Aus Forschung und Entwicklung im Bereich der SAS, insbesondere der Sprachsynthese, steht heute ein umfangreiches Repertoir an Evaluations- und Testverfahren zur Verfügung – auch wenn der Fokus traditionell oft auf der segmentalen oder lexikalischen Verständlichkeit (*intelligibility*) liegt (vgl. z.B. frühe Beschreibungen in [19, 20, 22, 21]). Mit der zunehmenden Verbreitung von SAS im Alltag rückten aber auch andere psycholinguistische und kognitive Aspekte in den Fokus der Forschung. So ist z.B. der Aspekt der kognitiven Beanspruchung (und dadurch verursachter Ablenkung) durch SAS insbesondere im Automobil auch von großem praktischen Interesse [29, 30, 31]. Auch die Natürlichkeit synthetischer Stimmen oder der verbalen Interaktion spielt eine immer größere Rolle (wie z.B. die Pausengestaltung [32]). So tragen z.B. eine an die Benutzer_in angepasste Intonation zum Vertrauen bei [28], oder gezielt eingesetzte Häsitationen zur pragmatischen Markierung von Unsicherheit [33].

Das ideale Vorbild für die MMI könnte nach Moore [34] nicht, wie zurzeit angenommen, die Interaktion zwischen kompetenten Muttersprachler_innen, sondern eher die Interaktion zwischen Nichtmuttersprachler_innen oder die zwischen Menschen und weniger intelligenten Interaktionspartnern wie Hunden sein. Dies gilt es zu ergründen und damit eine Basis für künftige Grundlagenforschung sowie Anwendungsentwicklung zu schaffen.

4 Danksagung

Das vorgestellte Projekt an der Albert-Ludwigs-Universität Freiburg wird gefördert durch die Vector Stiftung.

Literatur

[1] TAS, S. und R. ARNOLD: *Nutzung von Sprachassitenten in Deutschland*. In *Sprachassistenten: Anwendungen, Implikationen, Entwicklungen: ITG-Workshop, Magdeburg, 3. März 2020. Abstractbook*, S. 11–12. Otto von Guericke University Library, 2020. doi:10.25673/32572.2.

[2] WARD, N. G. und D. DEVAULT: *Challenges in building highly-interactive dialog systems*. *AI Magazine*, 37(4), S. 7–18, 2016.

[3] HÄB-UMBACH, R.: *Sprachtechnologien für Digitale Assistenten*. In R. BÖCK, I. SIEGERT, und A. WENDEMUTH (Hrsg.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, S. 227–234. TUDpress, Dresden, 2020.

[4] DEPPERMANN, A., S. KLEINER, und R. KNÖBL: *'Standard usage': Towards a realistic conception of spoken standard German*. In P. AUER, J. C. REINA, und G. KAUFMANN (Hrsg.), *Language variation – European perspectives IV: selected papers from the Sixth International Conference on Language Variation in Europe (ICLaVE 6), Freiburg, June 2011*, S. 83–116. John Benjamins Publishing Company, Amsterdam ; Philadelphia, 2013.

[5] ELSPASS, S. und S. KLEINER: *6. Forschungsergebnisse zur arealen Variation im Standarddeutschen*. In J. HERRGEN und J. E. SCHMIDT (Hrsg.), *Sprache und Raum - ein Internationales Handbuch der Sprachvariation: Deutsch*, Bd. 4, S. 159–184. De Gruyter Mouton, Boston, MA, 2020.

[6] NEUBARTH, F., M. PUCHER, und C. KRANZLER: *Modeling austrian dialect varieties for TTS*. In *INTERSPEECH 2008*, S. 1877–1880. 2008.

[7] MÄURER, D. K.: *Start-up will Sprachassistenten Schweizerdeutsch beibringen*. Online Artikel / Reportage, 2020-02-05. URL `https://www.br.de/nachrichten/netzwelt/start-up-will-sprachassistenten-schweizerdeutsch-beibringen,RpdkfrD`. Zuletzt aufgerufen am 20.01.2021.

[8] RAVEH, E., I. STEINER, I. SIEGERT, I. GESSINGER, und B. MÖBIUS: *Comparing phonetic changes in computer-directed and human-directed speech*. In *Electronic Speech Signal Processing (ESSV)*. Dresden, Germany, 2019.

[9] RAVEH, E., I. SIEGERT, I. STEINER, I. GESSINGER, und B. MÖBIUS: *Three's a crowd? effects of a second human on vocal accommodation with a voice assistant*. In *Interspeech*, S. 4005–4009. Graz, Austria, 2019.

[10] MEYER, S., M. JUNGHEIM, und M. PTOK: *Kindgerichtete Sprache*. *HNO*, 59(11), S. 1129–1134, 2011.

[11] JEANNIN, S., C. GILBERT, M. AMY, und G. LEBOUCHER: *Pet-directed speech draws adult dogs' attention more efficiently than adult-directed speech*. *Scientific Reports*, 7(1), 2017.

[12] SIEGERT, I. und J. KRÜGER: *How do we speak with ALEXA*. *Kognitive Systeme*, 2018(1), 2018.

[13] NILSSON, J.: *Dialect accommodation in interaction: Explaining dialect change and stability*. *Language & Communication*, 41, S. 6–16, 2015.

[14] BURIN, L. und N. BALLIER: *Accommodation in learner corpora: A case study in phonetic convergence*. *Anglophonia*, 24, 2017. doi:10.4000/anglophonia.1127.

[15] GILES, H.: *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities across Contexts*. Cambridge University Press, 2016.

[16] WARCHHOLD, S. und D. DURAN: *Perception of Synthetic Voices in Human-Agent Interaction*. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, S. 224–226. ACM, Virtual Event USA, 2020. doi:10.1145/3406499.3418756.

[17] BOERSMA, P.: *Praat: doing phonetics by computer*. *Glot International*, 5(9/10), S. 341–345, 2001.

[18] CHOMSKY, N.: *Syntactic Structures*. Mouton, 1957.

[19] GOLDSTEIN, M.: *Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener*. *Speech Communication*, 16(3), S. 225–244, 1995.

[20] BENOÎT, C., M. GRICE, und V. HAZAN: *The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences*. *Speech Communication*, 18(4), S. 381–392, 1996.

[21] DELOGU, C., S. CONTE, und C. SEMENTINA: *Cognitive factors in the evaluation of synthetic speech*. *Speech Communication*, 24(2), S. 153–168, 1998.

[22] PISONI, D. P.: *Perception of Synthetic Speech*. In J. P. H. VAN SANTEN, R. W. SPROAT, J. P. OLIVE, und J. HIRSCHBERG (Hrsg.), *Progress in Speech Synthesis*, S. 541–560. Springer-Verlag, New York, 1997.

[23] PEIRCE, J., J. R. GRAY, S. SIMPSON, M. MACASKILL, R. HÖCHENBERGER, H. SOGO, E. KASTMA, und J. K. LINDELØV: *PsychoPy2: Experiments in behavior made easy*. *Behavior Research Methods*, 51(1), S. 195–203, 2019.

[24] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[25] BATES, D., M. MÄCHLER, B. BOLKER, und S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), S. 1–48, 2015. Paket Version 1.1-23.

[26] KUZNETSOVA, A., P. B. BROCKHOFF, und R. H. B. CHRISTENSEN: *lmerTest package: Tests in linear mixed effects models*. *Journal of Statistical Software*, 82(13), S. 1–26, 2017.

[27] WOODS, K. J. P., M. H. SIEGEL, J. TRAER, und J. H. MCDERMOTT: *Headphone screening to facilitate web-based auditory experiments*. *Attention, Perception & Psychophysics*, 79(7), S. 2064–2072, 2017.

[28] BEŇUŠ, V., M. TRNKA, E. KURIC, L. MARTÁK, A. GRAVANO, J. HIRSCHBERG, und R. LEVITAN: *Prosodic entrainment and trust in human-computer interaction*. In *9th International Conference on Speech Prosody*, S. 220–224. ISCA, Poznan, Poland, 2018.

[29] STRAYER, D. L., J. TURRILL, J. M. COOPER, J. R. COLEMAN, N. MEDEIROS-WARD, und F. BIONDI: *Assessing cognitive distraction in the automobile*. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(8), S. 1300–1324, 2015.

[30] DURAN, D. und N. LEWANDOWSKI: *Untersuchung der kognitiven Beanspruchung durch Sprachassistenzsysteme*. In A. BERTON, U. HAIBER, und W. MINKER (Hrsg.), *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, S. 159–166. TUDpress, Dresden, 2018.

[31] WILKE, M., D. DURAN, und N. LEWANDOWSKI: *Wanted! – Assessing effects of distraction on working memory in speech perception*. In E. PUSTKA, M. A. PÖCHTRAGER, A. N. LENZ, J. FANTA-JENDE, J. HORVATH, L. JANSEN, J. KAMERHUBER, N. KLINGLER, H. LEYKUM, und J. RENNISON (Hrsg.), *Akten der Konferenz „Phonetik und Phonologie im deutschsprachigen Raum (P&P14)“*, S. 99–102. Institut für Romanistik, Universität Wien & Institut für Schallforschung, Österreichische Akademie der Wissenschaften, Wien, 2020. doi:10.25365/phaidra.159.

[32] TROUVAIN, J. und B. MÖBIUS: *Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache*. In A. BERTON, U. HAIBER, und W. MINKER (Hrsg.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, S. 334–341. TUDpress, Dresden, 2018.

[33] BETZ, S., S. ZARRIESS, É. SZÉKELY, und P. WAGNER: *The Greennn Tree – lengthening position influences uncertainty perception*. In *Proc. Interspeech*, S. 3990–3994. 2019.

[34] MOORE, R. K.: *Is Spoken Language All-or-Nothing? Implications for future Speech-Based Human-Machine Interaction*. In K. JOKINEN und G. WILCOCK (Hrsg.), *Dialogues with Social Robots*, S. 281–291. Springer Singapore, 2017.