UNTERSUCHUNG DER KOGNITIVEN BEANSPRUCHUNG DURCH SPRACHASSISTENZSYSTEME

Daniel Duran[1,2]*, Natalie Lewandowski*[2]
[1]*Albert-Ludwigs-Universität Freiburg,* [2]*Universität Stuttgart*
daniel.duran@germanistik.uni-freiburg.de

Kurzfassung: Wir verwenden den bekannten Lane Change Task um sprachliche Mensch-Maschine-Kommunikation unter quasi-alltäglichen Bedingungen zu testen. Wir präsentieren Ergebnisse einer Reihe von Experimenten mit dem Lane Change Task in einer einfachen PC-Implementierung mit am Computer angeschlossenem Lenkrad und Fußpedalen aus dem Gaming-Bereich. In Kombination mit verschiedenen Aufgaben zur Interaktion mit einem Sprachassistenzsystem (als scheinbaren Nebentätigkeiten) lässt sich so die kognitive Beanspruchung während der Sprachverarbeitung untersuchen. Wir diskutieren auch didaktische Aspekte im Einsatz der hier vorgestellten Methoden zur experimentellen Untersuchung der kognitiven Beanspruchung durch Sprachassistenzsysteme.

1 Einführung/Motivation

Sprachassistenzsysteme[1] (SAS) sind allgegenwärtig. Ihr Einsatz im Automobil, zum Beispiel, soll Fahrer entlasten indem manuelle Eingaben durch Sprachbefehle ersetzt werden, wobei die Hände am Lenkrad und die Augen auf die Straße gerichtet bleiben können. In diesem Zusammenhang rückt die Fahrerablenkung, auch durch sprachgesteuerte Assistenzsysteme im Automobil, immer stärker in den Fokus von Forschung, Entwicklung und Öffentlichkeit [1, 2]. Die mögliche Ablenkung durch verschiedene kommerzielle SAS im Automobil wurde z. B. von Cooper et al. [3] untersucht, die feststellen, dass schon einfache gesprochensprachliche Interaktionen die kognitive Beanspruchung der Fahrer signifikant erhöhen können. Der Ausgangspunkt für die hier vorgestellte Arbeit waren folgende Fragen: Wie hängen sprachliche Interaktion (Perzeption/Produktion) und allgemeine Kognition zusammen? Wie beeinflussen sich SAS und Kognition? Wie kann eine SAS-Evaluation in einer realistischeren Testumgebung aussehen?

Der vorliegende Beitrag befasst sich mit der gesprochensprachlichen Mensch-Maschine-Interaktion. Sein Titel lässt (mindestens) zwei Lesarten zu. In der einen steht die Erforschung der kognitiven Prozesse im Mittelpunkt, die an der menschlichen Sprachwahrnehmung und Produktion beteiligt sind. In der anderen steht die Erforschung der Auswirkungen von SAS auf den Benutzer im Mittelpunkt. Diese Ambiguität zwischen Kognitionswissenschaft, Psycholinguistik und Sprachtechnologie ist beabsichtigt. Zum einen bietet die Sprachtechnologie ein mächtiges Werkzeug zur systematischen Erforschung der menschlichen Sprachverarbeitung und zum anderen kann die Sprachtechnologie durch ein tiefgreifendes Verständnis der menschlichen Sprachverarbeitung verbessert werden [4]. Wir fassen Erkenntnisse aus einer Reihe von Experimenten zur kognitiven Beanspruchung durch sprachgesteuerte Assistenzsysteme zusammen und erörtern deren Einsatz als Werkzeug in Forschung, Entwicklung und auch der Hochschuldidaktik. Wir diskutieren methodische Aspekte im Studium der kognitiven Beanspruchung in der gesprochen-sprachlichen Mensch-Maschine-Kommunikation und der alltags- bzw. anwendungsnahen Evaluation von Dialogsystemen.

[1] In diesem Beitrag behandeln wir die Begriffe *Sprachassistenzsystem* und *Sprachdialogsystem* synonym.

2 Kognition, Sprachwahrnehmung und Sprachproduktion

Häufig untersuchte kognitive Funktionen im Zusammenhang mit Sprachwahrnehmung und Sprachproduktion sind das Arbeitsgedächtnis und die Aufmerksamkeit, die eng verzahnt sind. Studien zeigen z. B., dass Sprachwahrnehmung unter Störgeräuschen die kognitive Beanspruchung erhöht [5], dass sich kognitive Beanspruchung auf die Sprachproduktion [6] und die Sprachwahrnehmung [7] auswirkt, dass irrelevante Hintergrundgespräche den Zugriff auf das Arbeitsgedächtnis stören [8, 9], oder dass lexikalische Nachbarschaftsdichte sich bei Probanden mit niedrigeren Inhibitionswerten in der Worterkennung in Lärm auswirkt [10].

Ein Aspekt der Kognition, der in sicherheitsrelevanten Einsatzgebieten von Dialogsystemen eine große Rolle spielt, ist die *Ablenkung*. Dies ist, zum Beispiel, bei SAS im Automobil wichtig. Fahrer sollen durch diese Systeme entlastet und nicht abgelenkt werden, mit ihnen aber komplexe Aufgaben ausführen können. *Ablenkung* wird intuitiv als das Gegenteil von *Aufmerksamkeit* aufgefasst, ist aber ein nur wenig verstandenes Phänomen. Theoretische Modelle ihrer Mechanismen und Merkmale werden benötigt [11]. In der Aufmerksamkeitsforschung herrscht Einigkeit darüber, dass Aufmerksamkeitsprozesse als Filterprozesse anzusehen sind, die zu jedem Zeitpunkt aus allen uns erreichenden Informationen und Sinneseindrücken die für uns wichtigen weiterleiten und irrelevante hemmen. Unsere Aufnahmefähigkeit dabei ist generell begrenzt [12]. Dennoch gibt es zahlreiche Alltagssituationen in denen genau diese parallele Durchführung verschiedener Aufgaben mit unterschiedlichen Zielen nötig ist, obwohl echtes *Multitasking* für die meisten Menschen nicht möglich zu sein scheint [13]. Das Beispiel, auf das wir uns hier konzentrieren, ist die Interaktion mit SAS wie etwa beim Autofahren.

2.1 Messung kognitiver Beanspruchung, klassische Kognitionstests und Lane Change Task

Kognitive Beanspruchung lässt sich durch physiologische Merkmale wie Herzfrequenz, Hautleitwert, Blinzelfrequenz [14] oder mittels Pupillometrie [15, 16] ermitteln. Es wurden aber auch verschiedene Tests entwickelt um individuelle kognitive Merkmale zu erfassen. Dabei wird meist die Reaktionszeit (RT) und die Fehlerrate[2] als Maß bzw. Korrelat der kognitiven Beanspruchung angenommen. Ein Standardtest ist der *Simon Test* [17]. Dabei müssen Probanden auf einen (nonverbalen) Stimulus hin entweder eine linke oder eine rechte Taste drücken. Wird ein Stimulus auf einer Seite präsentiert, der eine Antwort auf der anderen Seite verlangt, steigen RT und Fehlerrate. Der *Stroop Test* ist ein Inhibitionstest in dessen klassischer Variante das Lesen eines Farbwortes unterdrückt werden muss, um die Aufmerksamkeit auf das Benennen der Wortfarbe bzw. Druckfarbe des eingeblendeten Wortes zu richten. Der *Detection Response Task* (DRT) wird eingesetzt um die Auswirkung kognitiver Beanspruchung auf die Aufmerksamkeit zur ermitteln [18, 19]. Dabei werden zu unvorhersehbaren Zeitpunkten Signale präsentiert, auf welche die Probanden reagieren müssen, während sie eine Aufgabe ausführen. Der OSPAN Test (*operation span*) ist eine Doppelaufgabe, welche eingesetzt wird um Exekutivfunktionen und Multitaskingfähigkeiten zu messen [20, 21]. Sie besteht aus zwei Teilen: Probanden müssen auf arithmetische Aufgaben mit *wahr* oder *falsch* antworten. Auf diese folgen Stimuluswörter, welche sie in der präsentierten Reihenfolge im Gedächtnis behalten und am Ende eines Blocks wiedergeben müssen. Die maximale Anzahl der korrekt wiedergegebenen Wörter wird als Maß für die Kapazität des Arbeitsgedächtnisses angesehen. Ein auditorischer OSPAN Test lässt sich in Kombination mit anderen Aufgaben in Situationen durchführen, die visuelle Aufmerksamkeit erfordern und visuelle Stimuluspräsentation nicht erlauben [22, 3]. Ein verbreite-

[2] Der Begriff *Fehlerrate* soll hier ganz allgemein die Anzahl der fehlerhaften Reaktionen bzw. das Gegenteil der Korrektheit bezeichnen.

tes Maß zur subjektiven Bestimmung der kognitiven Beanspruchung ist der *NASA Task Load Index* (NASA-TLX) [23, 24]. Er setzt sich aus 6 Dimensionen zusammen (mentale, physische und temporale Anforderungen, Frustration, Anstrengung und Performanz). Damit soll kognitive Beanspruchung bei konkreten Aufgaben ermittelt werden.

Der *Lane Change Task* (LCT) [25, 26] definiert eine standardisierte Testumgebung in der sich Fahrerablenkung durch diverse sekundäre Aufgaben gezielt im Labor untersuchen lässt. Der LCT ist eine Fahrsimulationsaufgabe, bei der Probanden einer dreispurigen Straße folgen. Bei einer Gesamtdauer von 3 Min. ist es möglich ca. 2 Min. Daten zu sammeln. Die primäre Aufgabe der Probanden ist auf visuelle Anweisung durch Schilder sofort auf eine bestimmte Spur zu wechseln und diese bis zum nächsten zu halten. Die Ablenkung durch Nebentätigkeiten wird durch Vergleich der Fahrleistung in verschiedenen Konditionen, mit und ohne Ablenkung, gemessen. Die Fahrleistung kann z. B. durch die mittlere Abweichung von der Ideallinie (mDev) oder durch die RT bis zur Initiierung des Spurwechsels gemessen werden [27, 28].

3 Sprachassistenzsysteme und ihre Evaluation

Die Ansprüche an SAS gehen heute weit über Verständlichkeit der Sprachsynthese und Genauigkeit der Spracherkennung hinaus. So spielen soziale Aspekte eine immer größere Rolle in Forschung und Entwicklung. Beňuš [29] z. B. stellt fest, dass durch die Ausnutzung des bei zwischenmenschlichen Dialogen beobachteten Konvergenzphänomens (d. h. der Angleichung der Gesprächspartner an einander [30]) die Mensch-Maschine-Kommunikation effektiver gestaltet werden kann. Durch eine soziale Komponente können Effekte wie Sympathie oder Dominanz sprachlich/phonetisch kodiert und zur Dialogsteuerung eingesetzt werden. Dadurch werden Benutzer emotional stärker eingebunden, was aber auch eine stärkere kognitive Beanspruchung und in der Praxis eine stärkere Ablenkung durch das System bedeuten könnte [31].

Der Begriff *Evaluation* wird in der Literatur unterschiedlich definiert und es werden unterschiedliche Metriken verwendet [32, 33, 34]. So können verschiedene Aspekte oder Komponenten eines Systems wie die Effizienz oder Dauer der Interaktion, der erfolgreiche Aufgabenabschluss, die Verständlichkeit (der Sprachsynthese) oder die Benutzerzufriedenheit evaluiert werden. Die Evaluation kann durch objektive Performanzmessungen erfolgen oder durch subjektive Maße mittels Fragebögen erfasst werden. In zahlreichen Projekten wurden Methoden zur Evaluation von SAS und ihrer Komponenten sowie zur Messung der Benutzerfreundlichkeit entwickelt, auch wenn sich nur wenige speziell auf diese Aspekte konzentrieren [35]. Als Beispiel sei nur PARADISE genannt, eine verbreitete Methode zur Erfassung der Benutzerfreundlichkeit auf Basis objektiv messbarer Parameter [36]. Die verlässliche Bestimmung der Benutzerfreundlichkeit sowie die statistische Auswertung der Daten stellen dabei besondere Herausforderungen dar [37]. Wizard-of-Oz (WoO) Szenarien kommen zum Einsatz, wenn SAS nicht zur Verfügung stehen oder über die zu untersuchende Funktionalität verfügen; oder auch als Referenz (*gold standard*) zum Vergleich mit einem System dienen [38].

4 Experimente

Wir konzentrieren uns hier auf die sprachliche Interaktion und gehen der Frage nach, welche Arten der Interaktion mit einem SAS die Benutzer so weit beanspruchen, dass sie von der Ausführung einer anderen Tätigkeit abgelenkt werden. Die Untersuchung dieser Fragestellung muss Dialogsysteme in einer natürlichen, alltagsnahen Situation testen, bei der die Aufmerksamkeit des Probanden nicht allein dem System und der Interaktion mit diesem gilt. Zur Veranschaulichung werden im folgenden eine Reihe von Experimenten vorgestellt, die am Institut für Maschinelle Sprachverarbeitung in studentischen Projekten durchgeführt wurden. Aufgrund

geringer Teilnehmerzahlen können die Ergebnisse der hier vorgestellten Experimente nicht verallgemeinert werden. Mit ihnen sollen aber verschiedene Ansätze und Möglichkeiten in der Untersuchung der kognitiven Beanspruchung durch SAS aufgezeigt werden. Der Versuchsaufbau sieht im Allgemeinen wie folgt aus: Als primäre Aufgabe verwenden wir eine Implementierung des LCT [39] für Windows-PC, mit dem eine alltagsnahe Testumgebung im Labor praktisch simuliert wird. Verschiedene sekundäre Aufgaben wie das Verfassen von Emails, das Erstellen von Kalendereinträgen, das Hören von Audiobooks oder Musik können so untersucht werden. Dabei wurde immer auch eine *Baseline* erfasst, bei der die Probanden keine Nebentätigkeit ausführen und keinen Ablenkungen ausgesetzt waren. Um die Fahrsimulation realistischer zu gestalten wurde ein *Hama Thunder V5 Racing Wheel* Lenkrad mit Fußpedalen verwendet. Für die SAS wurden Smartphones (*Apple*, *Samsung*) oder Tablets (*Sony*, *Huawei*) verwendet und rechts neben dem Lenkrad auf dem Tisch oder einem freistehenden Stativ positioniert. Für Videoaufnahmen wurde eine *Cisco flip video* Digitalkamera verwendet und auf das Gesicht der Probanden gerichtet. Die kognitive Beanspruchung kann somit objektiv erfasst werden durch: die LCT-Daten, die Erfassung der Blickrichtung und der Blinzelfrequenz anhand der Videoaufnahmen, die Dauer und Korrektheit der sekundären Aufgaben. Die subjektive Wahrnehmung der Schwierigkeit der jeweiligen Aufgaben wurde mit dem NASA-TLX abgefragt. Die Probanden erhielten für ihre Teilnahme an den Experimenten keine Aufwandsentschädigung.

4.1 Experimente und Ergebnisse

Experiment 1. Mit 4 Probanden (2 w[3], im Schnitt 23,5 Jahre alt mit 6 Jahren Fahrerfahrung) wurden aktive und passive Ablenkungen während der Interaktion mit einem SAS untersucht. Dabei wurde der Frage nachgegangen welche alltäglichen Tätigkeiten in einer solchen Situation am stärksten ablenken. Der LCT wurde 16 mal absolviert. Während dessen mussten verschiedene Aufgaben mit Google Now auf einem LG G3 Smartphone durchgeführt werden sowie jeweils eine weitere Aufgabe/Ablenkung. Die sprachgesteuerten Aufgaben waren: Email diktieren, Navigation und Anrufe tätigen. Die Ablenkungen waren: grelles Licht und Lärm (passiv) sowie Suchen eines Gegenstands in einer Tasche und Beantwortung von Fragen durch einen „Beifahrer“ (aktiv). Die Tasche stellte dabei die größte Ablenkung und das Licht die geringste dar. Zusätzlicher Lärm beeinträchtigt die Spracherkennung was zu Fehlern in der Interaktion mit dem SAS führt. Diese Systemfehler führten zu sehr starker Beeinträchtigung beim LCT.

Experiment 2. Mit 6 Probanden (2 w, 23–28 Jahre, alle mit Führerschein) wurde der Einfluss verteilter Aufmerksamkeit [40] auf das Fahrverhalten untersucht. Neben dem LCT wurde Google Now auf einem Sony Tablet verwendet um Emails zu verfassen und Kalendereinträge zu erstellen. Als zusätzliche Ablenkungen wurden Radio und Baulärm abgespielt. Alle Probanden absolvierten außerdem den *Simon Task* in einer separaten Sitzung. Entgegen der Erwartung wurde keine signifikante Korrelation zwischen den Ergebnissen des Simon Task und den einzelnen Aufgaben festgestellt. Das Verfassen einer Email war stärker ablenkend als die Terminfestlegung. Auch hier waren systembedingte Fehler von Google Now stark ablenkend.

Experiment 3. Mit 10 Probanden (mit Fahrerfahrung) wurde der Einfluss von Speech-to-Text Aufgaben auf kognitive Ablenkung untersucht. Nach ersten Tests mit Google Now, Siri und Cortana viel die Wahl auf ein WoO Design. Die sekundären Aufgaben waren: Navigation, Musik, Anrufe, Kalender, Nachrichten und der OSPAN. Subjektiv wie objektiv stellte der OSPAN die größte kognitive Belastung dar. Subjektiv wurde die Anruf-Aufgabe nach der Musik als am wenigsten belastend wahrgenommen, stellte objektiv gemessen (LCT) aber die stärkste Ablenkung nach dem OSPAN dar.

Experiment 4. Mit 10 Probanden (4 w, Studenten, 6,9 Jahre Fahrerfahrung) wurde kogniti-

[3] w = weiblich

ve Ablenkung während dem Fahren mit folgenden Aufgaben mit einem SAS untersucht: Audiobook, Email, Navigation. Außerdem wurde ein OSPAN durchgeführt. Probleme mit dem SAS führten hier teilweise zu größerer Belastung durch die Email-Aufgabe als durch den OSPAN.

Experiment 5. Mit 6 Probanden (3 w, 20–28 Jahre) wurde hier ebenfalls die Auswirkungen von Multitasking beim Autofahren auf die Fahrleistung untersucht. Es wurden ein iPhone (6. Generation) und ein Samsung Galaxy S7 mit Halterung verwendet. Sekundäre Aufgaben waren: Beifahrergespräch, Radiohören, Hörbuch, Interaktion mit Siri und Google. Die stärkste Belastung verursachte die Interaktion mit Siri, gefolgt von Google, die passiven sprachlichen Höraufgaben die geringste und das Beifahrergespräch lag im Mittelfeld.

Experiment 6. Stilz [41] untersuchte eine eigene Implementierung des LCT mit der Unity Spieleengine[4]. Im Vergleich zur Referenzimplementierung [39] integriert Stilz [41] Fahrsimulation und Analyse in einem Programm. Außerdem wurde die Integration des Fahrzeugfolgen-Paradigma [42] getestet (wobei Probanden ein vorausfahrendes Fahrzeug in einem konstanten Abstand bei variirender Geschwindigkeit folgen müssen). In einem Test mussten 8 Probanden (3 w, 22–26 Jahre, 1 ohne Führerschein) zunächst den LCT und die Fahrzeugfolgeaufgabe mit Tastatursteuerung bedienen. Anschließend wurden beide Aufgaben mit Lenkrad und Fußpedalen wiederholt. Subjektive Bewertungen wurden per Fragebögen abgefragt. Bei Tastatursteuerung zeigte in beiden Aufgaben eine (nicht signifikante) geringere mDev und Standardabweichung als mit Lenkradsteuerung. Fahrerfahrung der Probanden zeigte keinen signifikanten Einfluss. Den Probanden gefiel die visuell minimalistische Implementierung ohne ablenkende Details. Der Schwierigkeitsgrad und die Dauer der Tests wurden positiv bewertet.

5 Diskussion und Ausblick

Die Ergebnisse deuten darauf hin, dass interaktive, kreative sprachliche Aufgaben, wie das Verfassen einer Email, stärker ablenken als strukturierte Aufgaben, wie die Bedienung des Navigationssystems, oder passive Aufgaben wie das Anhören eines Audiobooks. Sie lenken auch stärker ab als eine Unterhaltung mit einem Beifahrer. Der LCT als vordergründig primäre Aufgabe wird von Probanden überwiegend als interessanter Test mit angemessener Schwierigkeit und Testdauer wahrgenommen. Die Motivation der Probanden durch eine ansprechende Testumgebung ist wichtig um schnelle Ermüdung und den Abbruch der Experimente durch die Probanden (vor allem bei mehreren Sitzungen) zu vermeiden. Eine alltagsnahe Testumgebung erhöht die Validität der erhobenen Daten. Die Erfahrung mit den Experimenten hat gezeigt, dass der LCT eine effektive aber auch einfach umzusetzende und kostengünstige Methode bietet gesprochensprachliche Mensch-Maschine-Kommunikation unter kontrollierten, aber dennoch realitätsnahen Bedingungen zu erforschen. Die Methodik ist flexibel und ermöglicht es unterschiedliche Aspekte zur untersuchen. Dadurch ist sie auch in der Hochschuldidaktik gut einzusetzen.

Der vorliegende Beitrag basiert auf der Annahme, dass zum einen die Sprachtechnologie ein mächtiges Werkzeug zur systematischen Erforschung der menschlichen Sprachverarbeitung bietet und dass zum anderen die Sprachtechnologie nur durch ein tiefgreifendes Verständnis der menschlichen Sprachverarbeitung ein dem Menschen vergleichbares Level erreichen kann. Der LCT bietet zusammen mit anderen kognitiven Tests wie z. B. einem auditorischen OSPAN oder der DRT, mit physiologischen Maßen wie der Blinzelfrequenz und subjektiven Maßen wie dem NASA-TLX eine einfache und kostengünstige aber dennoch effektive Testbatterie zur interdisziplinären Untersuchung der kognitiven Beanspruchung durch Sprachassistenzsysteme.

[4] http://unity3d.com

6 Danksagung

Diese Arbeit wurde zum Teil von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des SFB 732 finanziert. Die Experimente wurden von Studenten der Master- und Bachelorstudiengänge am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart 2015–2017 unter Anleitung von Prof. Grzegorz Dogil und den beiden Autoren DD und NL durchgeführt.

Dieser Beitrag verdankt seine Existenz im wesentlichen Prof. Grzegorz Dogil († 2017), der noch vor Fertigstellung des Manuskripts von uns gegangen ist. Seiner Erfahrung, Weisheit und Inspiration verdanken wir nicht nur die hier präsentierte Arbeit.

Literatur

[1] MANN, S.: *User Concepts for In-Car Speech Dialogue Systems and their Integration into a Multimodal Human-Machine Interface*. Doctoral Dissertation, Universität Stuttgart, 2010. doi:10.18419/opus-2680.

[2] 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.

[3] COOPER, J. M., H. INGEBRETSEN, und D. L. STRAYER: *Mental workload of common voice-based vehicle interactions across six different vehicle systems*. Tech. Rep., University of Utah & AAA Foundation for Traffic Safety, 2014.

[4] HERING, K. P.: *Situationsabhängiges Verfahren zur standardisierten Messung der kognitiven Beanspruchung im Straßenverkehr: Literaturübersicht und empirische Felduntersuchung*. Inaugural-Dissertation, Philosophische Fakultät, Universität zu Köln, 1999.

[5] ZEKVELD, A. A. und S. E. KRAMER: *Cognitive processing load across a wide range of listening conditions: Insights from pupillometry: Processing load across a wide range of listening conditions*. *Psychophysiology*, 51(3), S. 277–284, 2014.

[6] SCHULLER, B., S. STEIDL, A. BATLINER, J. EPPS, F. EYBEN, F. RINGEVAL, E. MARCHI, und Y. ZHANG: *The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load*. In *Proc. Interspeech*, S. 427–431. ISCA Archive, 2014.

[7] MATTYS, S. L. und L. WIGET: *Effects of cognitive load on speech recognition*. *Journal of Memory and Language*, 65(2), S. 145–160, 2011.

[8] COLLE, H. A. und A. WELSH: *Acoustic masking in primary memory*. *Journal of Verbal Learning and Verbal Behavior*, 15(1), S. 17–31, 1976.

[9] FARLEY, L. A., I. NEATH, D. W. ALLBRITTON, und A. M. SURPRENANT: *Irrelevant speech effects and sequence learning*. *Memory & Cognition*, 35(1), S. 156–165, 2007.

[10] TALER, V., G. P. AARON, L. G. STEINMETZ, und D. B. PISONI: *Lexical neighborhood density effects on spoken word recognition and production in healthy aging*. *The Journals of Gerontology Series B: Psych. Sciences and Soc. Sciences*, 65B(5), S. 551–560, 2010.

[11] REGAN, M. A., J. D. LEE, und K. L. YOUNG (Hrsg.): *Driver distraction: theory, effects, and mitigation*. CRC Press/Taylor & Francis Group, Boca Raton, 2009.

[12] KAHNEMAN, D.: *Attention and effort*. Prentice-Hall series in experimental psychology. Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[13] SANBONMATSU, D. M., D. L. STRAYER, N. MEDEIROS-WARD, und J. M. WATSON: *Who multi-tasks and why? Multi-tasking ability, perceived multi-tasking ability, impulsivity, and sensation seeking*. *PLoS ONE*, 8(1), S. e54402, 2013.

[14] STERN, J. A., L. C. WALRATH, und R. GOLDSTEIN: *The endogenous eyeblink. Psychophysiology*, 21(1), S. 22–33, 1984.

[15] KAHNEMAN, D. und J. BEATTY: *Pupil diameter and load on memory*. *Science*, 154(3756), S. 1583–1585, 1966.

[16] SCHWALM, M.: *Pupillometrie als Methode zur Erfassung mentaler Beanspruchungen im automotiven Kontext*. Dissertation, Universität des Saarlandes, Saarbrücken, 2009. doi:10.22028/D291-23297.

[17] LU, C.-H. und R. W. PROCTOR: *The influence of irrelevant location information on performance: A review of the simon and spatial stroop effects*. *Psychonomic Bulletin & Review*, 2(2), S. 174–207, 1995.

[18] RANNEY, T. A., G. H. S. BALDWIN, L. A. SMITH, E. N. MAZZAE, und R. S. PIERCE: *Detection response task (DRT) evaluation for driver distraction measurement application*. Final Report DOT HS 812 077, National Highway Traffic Safety Administration, 2014.

[19] CONTI-KUFNER, A. S.: *Measuring cognitive task load: An evaluation of the Detection Response Task and its implications for driver distraction assessment*. Dissertation, Technische Universität München, 2017.

[20] TURNER, M. L. und R. W. ENGLE: *Is working memory capacity task dependent? Journal of Memory and Language*, 28, S. 127–154, 1989.

[21] UNSWORTH, N., R. P. HEITZ, J. C. SCHROCK, und R. W. ENGLE: *An automated version of the operation span task*. *Behavior Research Methods*, 37(3), S. 498–505, 2005.

[22] WATSON, J. M. und D. L. STRAYER: *Supertaskers: Profiles in extraordinary multitasking ability*. *Psychonomic Bulletin & Review*, 17(4), S. 479–485, 2010.

[23] HART, S. G. und L. E. STAVELAND: *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*. In P. A. HANCOCK und N. MESHKATI (Hrsg.), *Human Mental Workload*. North Holland Press, Amsterdam, 1988.

[24] HART, S. G.: *NASA-task load index (NASA-TLX); 20 years later*. In *Proc. of the Human Factors and Ergonomics Society 50th Annual Meeting*, S. 904–908. Santa Monica, 2006.

[25] MATTES, S.: *The lane-change-task as a tool for driver distraction evaluation*. In H. STRASSER, K. KLUTH, H. RAUSCH, und H. BUBB (Hrsg.), *Quality of Work and Products in Enterprises of the Future / Qualität von Arbeit und Produkt in Unternehmen der Zukunft*, S. 57–60. Ergonomia Verlag, Stuttgart, 2003.

[26] MATTES, S. und A. HALLÉN: *Surrogate distraction measurement techniques: The lane change test*. In *Driver distraction: theory, effects, and mitigation*, S. 107–122. CRC Press/Taylor & Francis Group, Boca Raton, 2009.

[27] HARBLUK, J. L., P. C. BURNS, M. LOCHNER, und P. L. TRBOVICH: *Using the lane-change test (LCT) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces*. In *Proc. of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, S. 16–22. Public Policy Center, University of Iowa, 2007.

[28] YOUNG, K. L., M. G. LENNÉ, und A. R. WILLIAMSON: *Sensitivity of the lane change test as a measure of in-vehicle system demand. Applied Ergonomics*, 42(4), S. 611–618, 2011.

[29] BEŇUŠ, Š.: *Social aspects of entrainment in spoken interaction. Cognitive Computation*, 6(4), S. 802–813, 2014.

[30] SCHWEITZER, A., N. LEWANDOWSKI, und D. DURAN: *Attention, please! Expanding the GECO database*. In *Proc. of ICPhS 18*. Glasgow, UK, 2015. Paper number 620.

[31] PÊCHER, C., C. LEMERCIER, und J.-M. CELLIER: *Emotions drive attention: Effects on driver's behaviour. Safety Science*, 47(9), S. 1254–1259, 2009.

[32] GIBBON, D., I. MERTINS, und R. K. MOORE (Hrsg.): *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, Boston, MA, 2000.

[33] SCHIEL, F.: *Evaluation of multimodal dialogue systems*. In W. WAHLSTER (Hrsg.), *SmartKom: Foundations of Multimodal Dialogue Systems*, S. 617–643. Springer, 2006.

[34] MÖLLER, S., K.-P. ENGELBRECHT, F. KRETZSCHMAR, S. SCHMIDT, und B. WEISS: *Position paper: Towards standardized metrics and tools for spoken and multimodal dialog system evaluation*. In *NAACL-HLT Workshop SDCTD*, S. 5–6. ACL, 2012.

[35] DYBKJÆR, L., N. O. BERNSEN, und W. MINKER: *Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication*, 43(1–2), S. 33–54, 2004.

[36] WALKER, M. A., D. J. LITMAN, C. A. KAMM, und A. ABELLA: *PARADISE: A framework for evaluating spoken dialogue agents*. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, S. 271–280. ACL, Madrid, Spain, 1997.

[37] HAJDINJAK, M. und F. MIHELIČ: *The PARADISE evaluation framework: Issues and findings. Computational Linguistics*, 32(2), S. 263–272, 2006.

[38] PAEK, T.: *Toward evaluation that leads to best practices: Reconciling dialog evaluation in research and industry*. In *Proc. of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, S. 40–47. ACL, Rochester, NY, 2007.

[39] *Lane-Change-Test Driving Simulation according to ISO 26022*. 2011. URL `http://isotc.iso.org/livelink/livelink?func=ll&objId=11560806`. Software.

[40] TREISMAN, A. M. und G. GELADE: *A feature-integration theory of attention. Cognitive Psychology*, 12(1), S. 97–136, 1980.

[41] STILZ, S.: *Umsetzung des Lane Change Task in Unity 3D*. Bacheloarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2017. Examiner: G. Dogil, A. Schweitzer, supervisor: D. Duran.

[42] BROOKHUIS, K., D. D. WAARD, und B. MULDER: *Measuring driving performance by car-following in traffic. Ergonomics*, 37(3), S. 427–434, 1994.