Konzeption eines instrumentellen Evaluationsprotokolls für interaktive Sprachdienste im Kraftfahrzeug

Sebastian Möller[1], Klaus-Peter Engelbrecht[1,2], Stefan Hillmann[1], Florian Hinterleitner[1]

[1] *Quality and Usability Lab, TU Berlin; {sebastian.moeller|stefan.hillmnn|florian.hinterleitner}@telekom.de*

[2] *Heriot-Watt University, Edinburgh, UK; klaus-peter.engelbrecht@alumni.tu-berlin.de*

Einleitung

Trotz begrenzter Spracherkennungs-, Interaktions- und Sprachausgabeleistungen erhalten interaktive Sprachdienste zunehmend Einzug in das Kraftfahrzeug. Sollen diese in Zukunft komplexe Aufgaben unterstützen, wie sie heute immer häufiger auftreten (z. B. Auswahl eines Hotels nach unterschiedlichen Kriterien), müssen sie auch tatsächliche Dialoge mit Rückfragen durch das System oder den Nutzer unterstützen. Zudem erfordert die Durchführung der Aufgabe im Kontext einer potenziell gefährlichen Nebenaktivität (Autofahren) ein sorgfältiges Design, um die Konzentration des Fahrers nicht überzustrapazieren. Um eine für den Nutzer adäquate Qualität und Gebrauchstauglichkeit sicherzustellen bedarf es daher einer umfangreichen Evaluation. Leider sind die bisher verwendeten Verfahren [1] [2] [3, S. 71-76] personell sehr aufwändig und im Hinblick auf die Evaluationskriterien äquivalente instrumentelle Verfahren unbekannt, sodass eine detaillierte Evaluation häufig unterbleibt. Hier könnten automatisierte Verfahren helfen, welche einzelne Komponenten des interaktiven Systems oder ihr Zusammenspiel bei der Interaktion testen und Hinweise auf Schwachstellen liefern können.

Auf Basis von Erfahrungen in anderen Anwendungsdomänen soll im Vortrag ein Protokoll für eine instrumentelle Evaluation skizziert werden. Es umfasst eine Bestimmung der Performanz der Spracherkennung unter Anwendungsbedingungen, eine Bestimmung der Qualität der Sprachausgabe, sowie einen simulationsbasierten Test des Interaktionsverhaltens auf der pragmatischen Ebene, d.h. auf der Ebene der Bedeutung des im Dialog Gesagten. Bei der Spracherkennung kommt es auf die Aufzeichnung realistisch gestörter Sprachdaten an, welche idealerweise im fahrenden Kfz [1] oder alternativ mit einer akustischen Simulation geschehen kann. Für die Qualitätsbestimmung synthetisierter Sprache wurden instrumentelle Schätzer entwickelt, welche die Gesamtqualität sowie einzelne Qualitätsdimensionen zuverlässig vorhersagen können. Zur Analyse des Interaktionsverhaltens können gemischt regelbasiert-statistische oder rein statistische Simulationen verwendet werden, je nach verfügbarem Wissen über das System. Im Vortrag werden die derzeit bekannten Möglichkeiten vorgestellt und notwendiger Entwicklungsaufwand zur Anpassung an die Domäne im Kfz abgeschätzt.

Aufbau eines Sprachdialogsystems im Kfz

Wir gehen bei den folgenden Betrachtungen vom allgemeinsten Fall eines Sprachdienstes aus, der auf einem natürlichsprachlichen Dialogsystem basiert. Ein solches System verfügt über eine Spracherkennung, eine sprachverstehende Komponente, einen Dialogmanager, eine Antwortgenerierung und eine Sprachausgabe. Diese Komponenten müssen nicht unbedingt im Kfz integriert sein, sondern können auch auf einem entfernten Server implementiert sein.

Das Eingabesignal des Nutzers wird über ein im Kfz eingebautes Mikrofon, ein Mikrofon-Array oder das Mikrofon eines anderen Endgerätes (bspw. eines Telefons) aufgenommen, vorverarbeitet, und dem Spracherkenner zugeführt. Dieser kann entweder als eine Einheit oder als verteiltes System (Distributed Speech Recognition) ausgeführt sein, wobei letzterer eine Merkmalsextraktion im Kfz durchführt, die Merkmalsvektoren über ein Netzwerk übertragen werden und die anschließende Mustererkennung auf einem entfernten Server stattfindet. Der erkannte Text wird anschließend semantisch interpretiert. Die mittels der sprachverstehenden Komponente extrahierten semantischen Informationen – abgelegt z.B. als Attribut-Wert-Paare – dienen als Eingabe zum Dialogmanager.

Der Dialogmanager ist für die Steuerung des Dialogablaufes zuständig. Er verteilt Initiative, bietet aufgabenbezogene Antworten sowie Metakommunikation (bspw. bei Missverständnissen, Korrekturen, Bestätigungen, Hilfeersuchen) an, und übernimmt typischerweise auch die Kommunikation mit dem hinter der Sprachschnittstelle stehenden Anwendungssystem (bspw. ein Auskunftssystem, Transaktionssystem, etc.). Der Dialogmanager dient auch der Speicherung des bisherigen Dialogablaufs und steuert je nach Dialogzustand das vom Spracherkenner verwendete Vokabular, Grammatik, etc. Er generiert anschließend die auszugebende Information, die entweder zunächst mittels einer Antwortgenerierung in Text umgesetzt und dann von einer Sprachsynthese ausgegeben wird, oder aber direkt (z.B. bei einer Concept-To-Speech-Synthese oder bei Canned-Speech) als Sprachsignal über den/die bordeigenen oder an einem Endgerät angeschlossenen Lautsprecher ausgegeben wird.

Die beschriebene Verarbeitung betrifft ein voll natürlichsprachlich arbeitendes System. Andere Systeme gestatten lediglich eine natürlichsprachliche Eingabe (Systemausgaben erfolgen bspw. über ein Display) oder akzeptieren lediglich Kommandowörter (keine sprachverstehende Komponente), oder sie umfassen lediglich eine Sprachausgabe mittels synthetischer oder vorher aufgezeichneter natürlicher Sprache. Wie widmen uns im Folgenden der Evaluation der Komponenten des kompletten Systems. Je nach Anwendungsfall kann die Evaluationsmethode reduziert und auf die zutreffenden Komponenten angepasst werden. Bei der Betrachtung steht zunächst die Evaluation einzelner Komponenten (Spracherkennung,

Sprachverstehen, Sprachausgabe) im Fokus, bevor auf die Evaluation kompletter Systeme eingegangen wird.

Evaluation von Spracherkennung und Sprachverstehen

Das Hauptproblem bei der automatischen Spracherkennung im Kfz besteht in der starken Störung der Eingangssprache. Trotz verschiedener Vorverarbeitung zur Geräuschunterdrückung, Echounterdrückung, etc. ist das eingehende Sprachsignal typischerweise gestört und führt zu einer reduzierten Erkennungsleistung. In der Vergangenheit wurde daher versucht, neben der Vorverarbeitung des Eingangssignals auch die akustischen Modelle des Spracherkenners an den Anwendungsfall anzupassen. Aufgrund der Vielzahl an Fahr- und damit verbunden Geräuschsituationen (Geschwindigkeit, Außengeräusche, Fahrbahnoberfläche, Fensteröffnung, parallel ausgegebene Musik, etc.) ist eine solche Adaption aber schwierig. Dadurch besteht die Gefahr, dass auch ein gut eingestellter Spracherkennung in bestimmten Situationen nur unzureichende Erkennungsleistungen liefert.

Hierbei kann eine Simulation unterstützen, indem sie vorliegendes Trainings- und Testmaterial multipliziert mit einer großen Anzahl denkbarer Geräuschszenarien. Dabei muss berücksichtigt werden, dass die Stimme des Sprechers nicht unabhängig von der Geräuschsituation ist, sondern sich über den Lombard-Reflex bezüglich des Pegels, des Spektrums sowie der zeitlichen Struktur adaptiert [4]. Für eine realistische Simulation gestörten Sprachmaterials ist es also wichtig, neben einer adäquaten Simulation des Kfz-eigenen (Motor-, Roll- und Wind-) Geräusches sowie der Außengeräusche auch entweder eine Aufzeichnung oder eine Simulation von Lombard-Sprache zu verwenden. Eine gute Simulation (zum Zwecke des Trainings eines Spracherkenners) liegt unseres Kenntnisstandes nach noch nicht vor, sodass bislang mit Aufzeichnungen von Lombard-Sprache gearbeitet werden muss. Diese sind jedoch nur in sehr viel geringerem Umfang verfügbar als vergleichbare Korpora Nicht-Lombard-veränderter Sprache.

Die erzielbare Erkennungsrate ist insbesondere im Fall natürlichsprachlicher Systeme im Allgemeinen kein ausreichend guter Indikator für die Performanz im weiteren Dialogverlauf. Viele Wörter einer Phrase/eines Satzes sind nämlich nicht entscheidend für den Dialogverlauf, sodass die Rate korrekt erkannter und für den weiteren Dialogverlauf relevanter Konzepte (anstatt Wörter) normalerweise aussagekräftiger ist. Als Metriken hierfür eignen sich z.B. die Konzeptfehlerrate, aus Nutzersicht aber auch die innerhalb einer Äußerung korrekt neu überbrachten Konzepte o.ä. Eine Übersicht zu den in Frage kommenden Interaktionsparametern geben [5, S. 361-379], und darauf basierend [6, S. 421-425].

Evaluation der Sprachausgabe

Auch die Sprachausgabe ist aufgrund der Störgeräuschbelastung im Kfz kritisch. Sprachausgaben, die in ruhiger Umgebung eine gute Qualität und Verständlichkeit erzielen, können im fahrenden Kfz teilweise nur unzureichend verstanden und genutzt werden. Daher gehen Sprachausgabeproduzenten dazu über, die Verständlichkeit über eine Signalmanipulation oder durch ein speziell für den Anwendungsfall Kfz ausgewähltes Inventar zu steigern, manchmal auf Kosten der Natürlichkeit oder anderer perzeptiver Qualitätsdimensionen.

Die Qualität einer Sprachausgabe lässt sich valide und zuverlässig eigentlich nur mit menschlichen Testhörern unter Anwendungsbedingungen (im fahrenden Kfz) messen. Dieses Unterfangen ist allerdings offensichtlich sehr aufwändig, wenn es unter kontrollierten Bedingungen (z.B. auf einer gesperrten Teststrecke, u.U. mit vorgegebenen Fahrsituationen und Umweltbedingungen, in verschiedenen Kraftfahrzeugen, etc.) geschehen soll. Hier bietet sich wiederum die Simulation der Geräuschsituation in einem stehenden Kfz an, allerdings ist dann die Aufmerksamkeit nicht vergleichbar mit der beim Fahren. Dies kann durch einen Fahrsimulator verbessert werden, bei dem den Probanden parallel zur Beurteilungsaufgabe auch eine Fahraufgabe (bspw. ein Lane-Change-Task) gegeben wird.

Alternativ wurde an der TU Berlin und der Christian-Albrechts-Universität zu Kiel ein instrumenteller Schätzer zur Vorhersage der Gesamtqualität sowie einzelner Qualitätsdimensionen synthetisierter Sprache erstellt [7]. Hierzu wurden zunächst in Hörversuchen mittels multidimensionaler Analysen (Paarvergleich mit Multidimensionaler Skalierung [8] oder Semantisches Differenzial und Hauptkomponentenanalyse [9]) perzeptiv relevante Qualitätsdimensionen identifiziert. Ein Vergleich über mehrere Analysen [10] zeigte Unterschiede je nach Anwendungsfall (Sprachausgaben eines Dialogsystems vs. Hörbücher) und je nach verwendeten Synthesesystemen. Eine perzeptive Untersuchung der für Sprachdienste im Kfz relevanten Dimensionen steht noch aus, doch ist anzunehmen, dass hier vor allem Verständlichkeit und Annehmlichkeit der Stimme eine Rolle spielen könnten.

Anschließend werden die extrahierten Qualitätsdimensionen oder die Gesamtqualität aus dem Sprachsignal geschätzt. Hierfür wurden in der Vergangenheit zeitlich-energetische, spektrale oder prosodische Parameter mittels unterschiedlicher Algorithmen klassifiziert und zu einem Schätzwert für die Qualität einer Sprachprobe integriert. Für die wichtigsten Qualitätsdimensionen wurden dabei bisher ordentliche bis gute Schätzergebnisse erzielt [7]. Allerdings wurde dabei eine ruhige Abhörsituation angenommen, und den Probanden wurde meist keine parallele Aufgabe gestellt. Es bleibt zu zeigen, ob dieses Verfahren auch für die geräuschbehaftete Fahrsituation ausreichend aussagekräftige Ergebnisse liefert. Als Kriterium hierfür könnte z.B. herangezogen werden, ob ein Hörtest und eine instrumentelle Schätzung zwischen unterschiedlichen Synthesevarianten die gleiche Rangordnung abbilden können.

Evaluation kompletter Systeme

Die Evaluation der Spracheingabe und der Sprachausgabe sind nur begrenzt indikativ für die vom Nutzer bei der Interaktion erfahrene Gesamtqualität. Hierfür ist insbesondere das Nutzerverhalten wichtig, da es maßgeblich nicht nur Erkennungs- und Verstehensleistung, sondern auch

das Dialogverhalten des Systems bestimmt. Für eine valide und reliable Messung der Gesamtqualität sind also wieder normalerweise Test mit Probanden unumgänglich, welche wieder die o.a. Nachteile mit sich führen.

Zur Minimierung dieses Aufwandes wurden in der Vergangenheit verschiedene Ansätze untersucht, um das Nutzerverhalten im Umgang mit einem Sprachdialogsystem zu simulieren. Ziel war zum einen die Optimierung des Dialogverlaufes bei der Gestaltung dieser Systeme, zum anderen aber auch ein Vergleich zwischen unterschiedlichen Dialogstrategien, unterschiedlichen Nutzergruppen, etc. Zwei Simulationsansätze haben sich dafür in der Vergangenheit als nützlich erwiesen.

Zum einen kann der Dialogablauf eines typischen Systems als ein Zustandsautomat abgebildet werden – sozusagen ein Modell des Dialogverlaufes. Dieses Systemmodell umfasst genau genommen zwei Teile: Eine Beschreibung der Aufgabe, die das System erledigen soll/kann, sowie die Beschreibung des Interaktionsverlaufes zur Lösung dieser Aufgabe. Ein solches Systemmodell ist dann verfügbar, wenn das Sprachdialogsystem als sog. *Glass Box* vorliegt, also bzgl. seines „Innenlebens“ bekannt ist. Dieses Systemmodell wird nun mit einem Nutzermodell kombiniert, welches für jeden Zustandsübergang im Systemmodell eine Nutzereingabe generieren kann. Die im jeweiligen Zustand ausgewählte Nutzereingabe hängt von der Aufgaben-beschreibung [11] im Nutzermodell, der Sprachausgabe des Systems, dem bisherigen Dialogverlauf, sowie ggf. noch von weiteren Nutzerattributen (Sprachfähigkeiten, etc.) ab. Dabei wird i. Allg. mit Wahrscheinlichkeiten gearbeitet, welche verschiedene Varianten von Nutzerverhalten – und somit eine Vielzahl unterschiedlicher Dialogverläufe – mit vorher eingestellten Wahrscheinlichkeiten simulieren. Die Wahrscheinlichkeiten werden zuvor aus empirischen Experimenten extrahiert und ggf. über Regel weiter beeinflusst, bspw. Regeln, die die Auswirkung einer bestimmten Nutzereigenschaft auf die Nutzerantwort beschreiben. Eine solche Nutzersimulation wurde im Rahmen der MeMo-Werkbank implementiert und erfolgreich getestet [12].

Während der gemischt regelbasiert-probabilistische Ansatz immer dann sinnvoll verwendet werden kann, wenn das „Innenleben“ des Sprachdialogsystems bekannt ist, so lässt sich als Alternative ein komplett automatischer Lernansatz verwenden, wenn das Sprachdialogsystem als *Black Box* vorliegt, also sein „Innenleben“ unbekannt ist. Dazu wird ein einfaches Dialogsystem als Modell eines Nutzers auf ein vorliegendes Sprachdialogsystem angepasst, sodass es in bestimmten Dialogsituationen eine adäquate Spracheingabe machen kann. „Adäquat“ heißt in diesem Zusammenhang, dass das zu testende Sprachdialogsystem in einen neuen Systemzustand wechselt, welcher sich einem vorher definierten Zielzustand annähert [13]. Die vom Nutzer-simulator generierten Sprachäußerungen werden dabei in der gleichen Domäne (aber nicht unbedingt mit dem gleichen System) aufgezeichnet, und eine geeignete Auswahl wird dem Nutzersimulator über einfache Sprachmodelle (bspw. *n*-Gramme) antrainiert. Flexibel arbeitet ein solches System, wenn zur Sprachausgabe auch synthetische Sprache generiert werden kann, welche vom Sprachdialogsystem dann mit ähnlicher Erkennungsleistung aufgenommen wird wie natürliche Sprache.

Zur Simulation von Erkennungsfehlern lässt sich in einem solchen Dialogsimulator die Sprachausgabe auf vielfältige Art manipulieren. Auch auf inhaltlicher Ebene können Verständnisprobleme simuliert werden. In Kombination mit dem zuvor angesprochenen Lombard-Sprachsimulator ließen sich durch einen solchen Ansatz in Zukunft eine Vielzahl von Dialogabläufen und damit einhergehenden Verständnisproblemen sowie geeignete Korrektur-mechanismen testen.

Aus den simulierten Dialogverläufen lassen sich Abschätzungen für die vom Nutzer erfahrene Gesamtqualität sowie für verschiedene Aspekte der Qualität und Gebrauchstauglichkeit des Sprachdialogsystems treffen. Hierbei kann der Dialogverlauf z.B. parametriert werden, und aus den Parameterwerten mittels eines linearen Regressionsmodells Qualitätsschätzwerte berechnet werden. Die Parameter können sowohl die (beobachtbare) „Oberflächenstruktur“ des simulierten Dialogs als auch interne Leistungsparameter des Sprachdialogsystems umfassen, sofern sie verfügbar sind. Ein hierfür in der Vergangenheit oftmals eingesetztes Modell ist das PARADISE-Modell [14], welches Nutzerzufriedenheit als Linearkombination von Aufgabenerfolg und „Dialog-kosten“ (Dauer, Erkennungsfehler, Korrekturen, etc.) schätzt.

Diskussion und Ausblick

Die beschriebenen Ansätze sollen Möglichkeiten für eine automatisierte Schätzung von Leistung und Qualität bei natürlichsprachlichen Dialogsystemen aufzeigen. Anzumerken ist aber, dass (mit Ausnahme der Geräuschsimulation für die Spracherkennung) kaum einer der beschriebenen Ansätze bereits im Anwendungsfall Kfz verwendet wurde [3, S. 70]. Bei der instrumentellen Schätzung der Ausgabequalität müssen hierzu zunächst die relevanten Qualitätsdimensionen auditiv bestimmt werden, um anschließend Schätzer dafür zu bauen, welche die relevanten Dimensionen zuverlässig abbilden.

Bei der Simulation von Dialogverläufen sollten insbesondere die durch die Situation im Kfz bedingten Dialogphänomene (niedrigere Erkennungsleistung und dadurch mehr Missverständnisse bei Lombard-Sprache, Vertauschung von phonemisch ähnlichen Wörtern in geräuschbehafteter Umgebung, begrenzte Aufmerksamkeit durch Konzentration auf die Fahraufgabe [15]) betrachtet werden. Insbesondere Für die Fahrsicherheit muss sichergestellt werden, dass die Interaktion mit den Diensten im Fahrzeug den Fahrer nicht von der Hauptaufgabe des Fahrens ablenkt (seine begrenzten kognitiven Ressourcen [16] unangemessen verwendet werden [17]). Sollte für die genannten Phänomene eine geeignete Simulation gelingen ergeben sich daraus allerdings sehr große Potenziale: So könnte das Dialogsystem während der Interaktion diese Probleme abfangen und adäquate Lösungsstrategien entwickeln lernen. Zum Training und zum Test solchermaßen adaptiver und „kognitiver“ Systeme ist eine Simulation von unschätzbarem Wert, liefert sie doch

umfangreiches Datenmaterial auf sehr effiziente Art und Weise.

Literatur

[1] L. Hassel und E. Hagen, „Evaluation of a Dialogue System in an Automotive Environment Assesment," in *Proc. Int. Driving Sym. on Human Factors in Driver*, Aspen, USA, 2005.

[2] W. Minker, U. Haiber, P. Heisterkamp and S. Scheible, "Teh SENECA spoken language dialogue system," *Speech Communication,* Bd. 43, S. 89-102, 2004.

[3] S. W. Hamerich, Sparchbedienung im Automobil, Berlin, Heidelberg: Springer, 2009.

[4] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication,* Bd. 20, Nr. 1-2, S. 13-22, 1996.

[5] S. Möller, Quality of Telephone-Based Spoken Dialogue Systems, Boston: Kluwer Academic Publishers, 2005.

[6] P. L. Mateo Navarro, S. Hillmann, S. Möller, D. Sevilla Ruiz und G. Martínez Pérez, „Run-time model based framework for automatic evaluation of multimodal interfaces," *Journal on Multimodal User Interfaces,* Bd. 8, Nr. 4, S. 399-427, 2014.

[7] C. Norrenbrock, F. Hinterleitner, U. Heute and S. Möller, "Quality Prediction of Synthesized Speech based on Perceptual Quality Dimensions," *Speech Communication,* S. 17-35, 2015.

[8] F. Hinterleitner, C. Norrenbrock and S. Möller, "What Makes this Voice Sound so Bad? A Multidiemnsional Analysis of State-of-the-Art Text-to-Speech Systems," in *Proc. of the 2012 IEEE Workshop on Spoken Language Technology*, 2012.

[9] F. Hinterleitner, S. Möller, N. C. und U. Heute, „Perceptual Quality Dimensions of Text-to-Speech Systems," in *Interspeech*, 2012.

[10] F. Hinterleitner, C. Norrenbrock und S. Möller, „Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech.," in *Proc. of 8th ISCA Speech Synthesis Workshop*, 2013.

[11] S. Hillmann and K.-P. Engelbrecht, "Modelling Goal Modifications in User Simulation," in *Proc. of Future and Emerging Trends in Language Technology*, Sevilla, ES, 2015.

[12] K.-P. Engelbrecht, Estimating Spoken Dialog System Quality with User Models, 2011, S. 125.

[13] T. Scheffler, R. Roller and N. Reithinger, "SpeechEval - Evaluating Spoken Dialog Systems by User Simulation," in *Proc. 6th IJCAI Workshop*, Pasadena, CA, 2009.

[14] M. Walker, D. Litman, C. Kamm and A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in *Proc. of ACL 1997*, 1997.

[15] D. D. Salvucci, „Distract-R: Rapid prototyping and evaluation of in-vehicle interfaces," in *Proc. CHI 2005*, 2005.

[16] C. D. Wickens, "Multiple resources and performance prediction," *Theor. Issues in Ergon. Sci.,* Bd. 3, Nr. 2, S. 159-177, 2002.

[17] J. Niemann, *Designing Speech Output for In-car Infotainment Applications Based on a Cognitive Model of Attention Allocation,* Berlin, 2013.