Einfluss von Spracherkennung und Sprachsynthese auf die Qualität natürlichsprachlicher Dialogsysteme

Sebastian Möller, Janto Skowronek
Institut für Kommunikationsakustik (IKA), Ruhr-Universität Bochum; Email: moeller@ika.ruhr-uni-bochum.de

Einleitung

Bei der realen Anwendung von Sprachdialogsystemen kommt denjenigen Systemkomponenten, die sich direkt auf die akustische Realisierung gesprochener Sprache beziehen, eine wichtige Rolle zu. Einerseits hängt der Grad, zu dem sich der Benutzer vom System verstanden fühlt, zu einem großen Teil von der Erkennungsleistung des Systems ab. Diese wird wiederum durch die raumakustische Situation und mögliche weitere Übertragungskanäle (z.B. Telefon) beeinflusst. Andererseits kann die Sprachausgabe die wahrgenommene Gesamtqualität erheblich beeinträchtigen oder verbessern.

In diesem Beitrag soll der Einfluss der Spracherkennung und der Sprachausgabe (natürliche oder synthetisierte Sprache) auf die vom Benutzer erfahrene Gesamtqualität sowie auf einzelne Qualitätsdimensionen untersucht werden. Als Beispiel dient ein telefonbasiertes Dialogsystem für Restaurantauskünfte (Bochumer Restaurant-Informations-System, BoRIS), welches in einer Versuchsumgebung am IKA aufgebaut wurde. Mit ihm wurden Wizard-of-Oz-Experimente durchgeführt, in denen die Erkennungsleistung des Spracherkenners sowie unterschiedliche Sprachausgaben kontrolliert verändert wurden. Ausgangspunkt der Qualitätsuntersuchungen ist eine Taxonomie von Aspekten der Dienstequalität eines Sprachdialogdienstes, welche von Möller[1] vorgestellt wurde. Diese Taxonomie ermöglicht eine detaillierte Untersuchung des Einflusses der beiden Systemkomponenten auf einzelne Qualitätsdimensionen sowie auf die Gesamtqualität des Systems.

Qualität eines Sprachdialogdienstes

Die Qualität eines Telefon-Sprachdienstes hängt von einer Vielzahl von Einflussfaktoren des verwendeten Systems, des Benutzers, der (akustischen) Umgebung, des vom System angebotenen Dienstes, sowie des Benutzungskontexts ab. Qualität wird dabei als das "Ergebnis der Beurteilung der wahrgenommenen Beschaffenheit einer Einheit im Hinblick auf die erwünschte Beschaffenheit" definiert (Jekosch[2]). Sie ist somit eine perzeptive Größe und setzt einen Wahrnehmungs- und Beurteilungsvorgang voraus. Möller beschrieb eine Taxonomie der genannten Einflussfaktoren und setzte diese in Zusammenhang mit Aspekten der Dienstequalität wie der Kooperativität, der Qualität der Spracheingabe und -ausgabe, der Symmetrie der Interaktion, der Effizienz bezogen auf die Kommunikation, die Aufgabe und den Dienst, sowie globaler Qualitätsaspekte wie Zufriedenheit des Benutzers oder Akzeptanz.

Bei der Evaluierung von Sprachdialogdiensten werden i.A. zwei Arten von Informationen gesammelt. Zum einen können Benutzerurteile (meist in Form von Fragebögen) Aufschluss über die wahrgenommene Qualität liefern. Zum anderen werden während der Interaktion einzelne Parameter des Systems oder des Dialogs instrumentell gemessen (z.B. Dauer, Anzahl der *turns*, etc.). Nach einer anschließenden expertenbasierten Transkription und Annotation können sog. Interaktionsparameter bestimmt werden, welche Eigenschaften des Systems, des Benutzers und der Interaktion zwischen beiden beschreiben. Diese Beschreibung steht zwar im Zusammenhang mit der wahrgenommenen Qualität, jedoch sind bislang nur sehr unzureichende Ansätze zur Vorhersage der Qualität aus Interaktionsparametern bekannt.

Die Taxonomie der Qualitätsaspekte diente als Grundlage zur Erstellung eines dreiteiligen Fragebogens, welcher die verschiedenen Dimensionen der Qualität auf unterschiedlichen Ebenen der Taxonomie erfassen soll. Teil A des Fragebogens erfragt den Hintergrund und die Erwartungen der Versuchsperson vor der Interaktion mit dem System. Teil B wird nach jeder einzelnen Interaktion beantwortet und besteht aus 25 Fragen zum gerade geführten Gespräch. Teil C erfasst den Eindruck nach Ablauf aller Gespräche.

Das Versuchssystem BoRIS

Für die Untersuchungen wurde am IKA ein Dialogsystem erstellt, mit dessen Hilfe Informationen über Bochumer Restaurants per Telefon erfragt werden können. Suchkriterien sind die Art der Küche, der Stadtteil des Restaurants, der ungefähre Preis, sowie Wochentag und Uhrzeit (Öffnungszeiten des Restaurants).

Das System arbeitet entweder mit einem kommerziellen Spracherkenner oder mit einer Erkennersimulation, welche kontrolliert einstellbare Erkennungsraten erzeugen kann. Diese Simulation basiert auf einer Realzeit-Transkription des Versuchsleiters, auf der kontrollierte "Erkennungsfehler" nach vorher bestimmten Vertauschungsmustern erzeugt werden. Die Dialogsteuerung erfolgt über einen Zustandsgraphen, welcher mit Hilfe des CSLU-Toolkit implementiert wurde und der wahlweise über unterschiedliche Bestätigungsstrategien verfügt. Die Sprachausgabe erfolgt entweder über vorher aufgezeichnete natürlichsprachliche Ansagen (männliche oder weibliche Stimme) oder über ein TTS-System, oder es werden Kombinationen beider Verfahren (jeweils für die festen Systemmeldungen und für die variablen Restaurantinformationen) verwendet. Hierzu wurde die am IKA erstellte Verkettungssynthese, bestehend aus der symbolischen Vorverarbeitung SyRUB[3] und dem Synthetisator IKAphon[4], benutzt. Die einzelnen Module können gezielt zu bestimmten Systemkonfigurationen mit verschiedenen Eigenschaften (Erkennngsrate, Art der Bestätigung und der Sprachausgabe) kombiniert werden.

Benutzer kommunizieren mit dem System über eine simulierte Telefonverbindung in einer Testumgebung. Am hier beschriebenen Experiment nahmen 40 Versuchspersonen (11 w, 29 m) im Alter zwischen 23 und 51 Jahren teil. Sie hatten größtenteils keine Erfahrung im Umgang mit Sprachdialogsystemen, jedoch Kenntnisse der Bochumer Umgebung. Die Versuchspersonen wurden für ihre Tätigkeit bezahlt.

Vor Beginn des Versuchs füllten die VPen zunächst Teil A des Fragebogens aus. Dann wurde ihnen die Aufgabe gestellt, anhand von 5 Szenarien Restaurants nach unterschiedlichen Kriterien auszusuchen und das von BoRIS ermittelte Ergebnis (Name des Restaurants) niederzuschreiben. Nach jeder Interaktion wurde Teil B des Fragebogens ausgefüllt. Abschließend beurteilten sie in Teil C ihren Eindruck nach allen 5 Dialogen. Während der Interaktionen

wurde eine Log-Datei angelegt, die nach anschließender Transkription und Annotation durch einen Experten zur Bestimmung von 47 Interaktionsparametern führte. Diese Parameter beziehen sich auf die Spracheingabe (Wort- und Konzeptfehlerrate etc.), die Kooperativität des Systems (*contextual appropriateness*, Anzahl der System- und Benutzerfragen, Korrektheit der Systemantworten, etc.), die Meta-Kommunikation (Hilfeanforderungen, *barge-ins*, etc.), den Dialogverlauf (Länge, Antwortverzögerung, etc.) und die Erfüllung der Aufgabe (*task success* bzw. Kappa-Koeffizient).

Einfluss der Spracherkennung

Die (simulierte) Erkennungsrate zeigt - neben der offensichtlichen Beeinflussung der erkennungs- und verstehensbezogenen Parameter - einen signifikanten Einfluss auf den Anteil der Korrekturäußerungen, die Kooperativität der Systemäußerungen, sowie auf den *task success* (Kappa-Koeffizient). Perzeptiv schlägt sich eine Verschlechterung der Erkennungsrate vor allem in den Urteilen auf die Fragen B5 ("Wie fühlten Sie sich vom System verstanden?"), B9 ("Ihrer Einschätzung nach verarbeitete das System Ihre Angaben richtig.") und B11 ("Wie häufig machte das System Fehler?") nieder, siehe Abb. 1. Offensichtlich sind die VPen in der Lage, die Quelle der Interaktionsprobleme recht gut zu lokalisieren. Im Bereich höherer Erkennungsraten tritt jedoch eine Sättigung der Urteile ab einer Schwelle bei ca. 80% auf. Demnach fließen auch andere Aspekte in die Bewertung der drei Fragen ein.

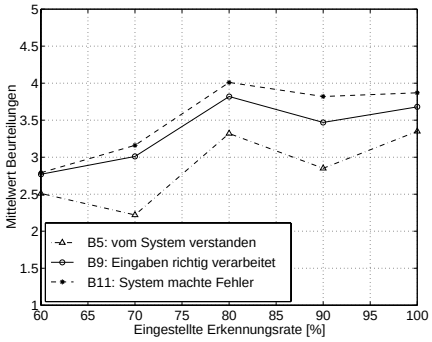

Abbildung 1: Einfluss der Erkennungsrate auf die Beurteilungen der Fragen B5, B9 und B11.

Einfluss der Sprachausgabe

Bei der Sprachausgabe sind nur zwei Parameter signifikant betroffen: Die Dauer der Systemäußerungen (bedingt durch die niedrigere Sprechgeschwindigkeit der Synthese) und die Verzögerung der Benutzeräußerungen (*cognitive demand*). Demgegenüber verschlechtern sich eine Reihe von VP-Urteilen relativ drastisch. So sind neben den offensichtlichen Fragen nach der Höranstrengung (B6: "Sie mussten sich sehr konzentrieren um zu verstehen, was das System von Ihnen verlangte") und der Verständlichkeit (B7: "Wie gut war das System akustisch verstehbar?") auch die Klarheit der Auskünfte betroffen (B3: "Die Auskünfte waren klar...unklar."), die Natürlichkeit (B12: "Das System reagierte wie ein Mensch."; B18: "Sie empfanden das Gespräch als natürlich...unnatürlich."; B22: "Die Stimme des Systems war natürlich...unnatürlich."), die Freundlichkeit (B16: "Das System reagierte freundlich...unfreundlich."), Glattheit (B21: "Das Gespräch verlief glatt...holprig."), Annehmlichkeit (B24: "Sie empfanden das Gespräch als angenehm...unangenehm."), Stress (B25: "Sie fühlten sich während des Gespräches entspannt...gestresst.") sowie der Gesamteindruck. Die Verschlechterung (vgl. Abb. 2 und 3) fällt deutlich stärker als bei der Spracherkennung aus.

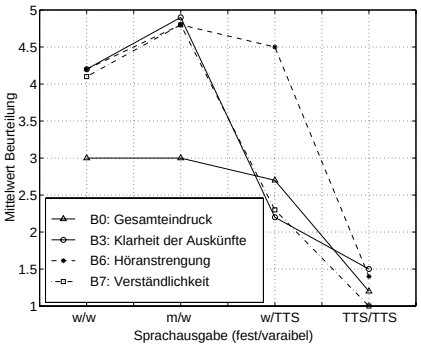

Abbildung 2: Einfluss der Sprachausgabe auf die Beurteilungen der Fragen B0, B3, B6 und B7.

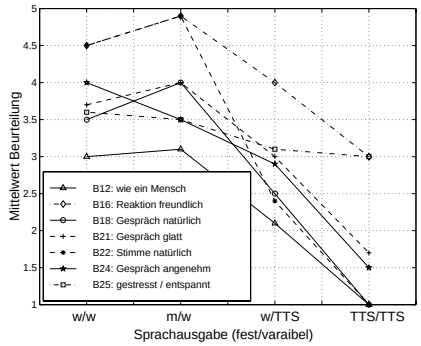

Abbildung 3: Einfluss der Sprachausgabe auf die Beurteilungen der Fragen B12, B16, B18, B21, B22, B24 und B25.

Diskussion und Ausblick

Sowohl auf der Eingabe- als auch auf der Ausgabeseite lassen sich signifikante Einflüsse auf die Qualität feststellen. Während die Erkennungsrate hauptsächlich Interaktionsparameter der Kategorien Spracheingabequalität, Kooperativität, Effizienz der Kommunikation und *task success* beeinträchtigt, so ist die Auswirkung der Sprachsynthese perzeptiv sehr viel deutlicher auszumachen. Dabei werden Qualitätsdimensionen auf unterschiedlichen Ebenen berührt, bis hin zum Gesamteindruck vom System. Die Sprachausgabe wirkt offensichtlich wie eine Art Visitenkarte des Systems und beeinflusst dessen Gesamt-Erscheinungsbild und "Persönlichkeit". Benutzer sind hier viel weniger in der Lage, die perzeptiven Einflüsse einzugrenzen.

Die Untersuchungen zum Einfluss der Systemkonfigurationen werden fortgeführt, wobei sich die Taxonomie der Dienstequalität als ein wichtiges Hilfsmittel zur Testgestaltung und Interpretation der Ergebnisse bewährt hat. Durch die gleichzeitige kontrollierte Sammlung von Interaktionsparametern und subjektiven Qualitätsurteilen lassen sich neue Ansätze zur Vorhersage der Qualität entwickeln. Untersuchungen hierzu werden augenblicklich durchgeführt.

[1] Möller, Sebastian. A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems. Proc. 3rd SIGdial Workshop, USA-Philadelphia PA, 2002, 142-153.

[2] Jekosch, Ute. Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitationsschrift, Uni/GH, D-Essen, 2000.

[3] Böhm, Arnd. Maschinelle Sprachausgabe deutschen und englischen Textes. Shaker Verlag, D-Aachen, 1993.

[4] Köster, Stefanie. Modellierung von Sprechweisen für widrige Kommunikationsbedingungen mit Anwendung auf die Sprachsynthese. Shaker Verlag, D-Aachen, 2003.