Neue ITU-T-Empfehlungen zur Evaluierung telefonbasierter Sprachdialogdienste

Sebastian Möller

Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7
D-10587 Berlin, Germany
sebastian.moeller@telekom.de

Abstract: In diesem Beitrag wird ein neuer Satz von ITU-T-Empfehlungen vorgestellt, welcher sich mit der Evaluierung telefonbasierter Sprachdialogdienste befasst. Die Empfehlungen wurden erst kürzlich verabschiedet oder befinden sich noch im Planungsstadium. Sie wurden im Rahmen des INSPIRE-Projektes bereits zur Evaluierung eines Smart-Home-Systems verwendet. Ausgewählte Ergebnisse dieser Evaluierung werden vorgestellt und im Hinblick auf die empfohlene Methodologie diskutiert.

1 Einführung

Sprachdialogsysteme werden in zunehmendem Maße über das Telefonnetz zugänglich gemacht. Sie erlauben es, neue Dienste anzubieten (z.B. Abfrage von Börsenkursen, Fußballergebnissen oder Telefon-Banking) oder bestehende, auf zwischenmenschlicher Kommunikation beruhende Dienste effizienter zu gestalten (z.B. Firmenportale oder Fahrplanauskunft). Je mehr Systeme und damit Dienste verfügbar werden, desto stärker steigt der Bedarf an einer fundierten Evaluierung. Da Systementwickler und Netz- bzw. Dienstbetreiber meist nicht identisch sind, müssen sich beide Seiten auf allgemein akzeptierte Evaluierungsmethoden einigen. Eine nach diesen Methoden durchgeführte Evaluierung sollte sowohl dem Systemhersteller Hinweise auf mögliche Schwachstellen seines Systems liefern, als auch dem Dienstbetreiber ein realistisches Bild der Qualität zeichnen, welche seine Kunden bei der Benutzung des Dienstes erfahren.

Zur Definition geeigneter Evaluierungsmethoden hat der *Telecommunication Standardization Sector* der *International Telecommunication Union* (ITU-T) kürzlich eine neue Arbeitsgruppe ins Leben gerufen, welche Empfehlungen zur Evaluierung telefonbasierter Sprachdienste erstellen soll [ITU05a]. Die Arbeit baut auf zwei bestehenden Empfehlungen auf, die in den vergangenen Jahren verabschiedet wurden und Methoden zur subjektiven Beurteilung der Sprachausgabekomponente [ITU94] bzw. des Gesamtsystems [ITU03] beschreiben. Diese Methoden sollen nun in vier Arbeitspunkten erweitert werden:

(1) Definition eines Satzes von Interaktionsparametern zur quantitativen Erfassung der qualitätsrelevanten Eigenschaften der Interaktion. Dieser Arbeitspunkt umfasst zunächst eine umfangreiche Sammlung bekannter Parameter in einem sog. *Supplement*, und darauf aufbauend die Definition eines reduzierten Satzes qualitäts-

relevanter Parameter in einer neuen Empfehlung.

(2) Definition einer neuen Empfehlung zur Beschreibung des Einflusses des Übertragungskanals auf die Leistung sprachtechnologischer Komponenten, insbes. auf die Leistung des Spracherkenners und die Qualität der Sprachsynthese.

(3) Definition von Modellen zur Vorhersage der Qualität telefonbasierter Sprachdienste auf der Basis messbarer Parameter oder Signale.

(4) Überprüfung und u.U. Erweiterung bzw. Modifikation der Empfehlungen zur subjektiven Qualitätsbeurteilung [ITU03,ITU94], insbes. in Bezug auf die Benutzbarkeit (*usability*) der entsprechenden Dienste.

Es sei angemerkt, dass Arbeitspunkt 2 speziell auf telefonbasierte Dienste ausgerichtet ist, während alle anderen Arbeiten für eine größere Klasse von Diensten relevant sind. Neben den erwähnten Empfehlungen zur subjektiven Beurteilung [ITU94,ITU03] liegt bereits eine Liste von Parametern mit den entsprechenden Definitionen vor, welche die Grundlage des *Supplement* bilden soll [ITU05]. Allerdings ist die Erfahrung sowohl mit den subjektiven Methoden als auch mit den Parametern bislang sehr begrenzt.

In diesem Beitrag sollen daher die Methoden vorgestellt und einer ersten unabhängigen Überprüfung unterzogen werden. Hierzu wurde im Rahmen des EU-geförderten IST-Projektes INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces; IST-2001-32746) ein Experiment durchgeführt, mit dem die Qualität des dort erstellten Sprachdialogsystems zur Steuerung von Hausgeräten erfasst wurde. Dieses Experiment soll nun im Hinblick auf die Evaluierungsmethoden untersucht werden. In Abschnitt 2 werden zunächst der Begriff der Qualität eines Sprachdialogdienstes und dessen Einflussfaktoren, wie sie in [Mö05] und [ITU03] definiert wurden, vorgestellt. Methoden zur subjektiven Beurteilung der Qualität sowie zur Messung von Interaktionsparametern werden in Abschnitt 3 beschrieben. Abschnitt 4 diskutiert die Ergebnisse des INSPIRE-Experimentes. Eine Zusammenfassung und ein Ausblick auf zukünftige Schritte bei der Standardisierung folgen in Abschnitt 5.

2 Qualität und ihre Einflussfaktoren

Die Qualität eines telefonbasierten Sprachdialogdienstes hängt von einer Vielzahl von Einflussfaktoren ab. Diese werden bestimmt durch das verwendete System, den Benutzer, die (akustische) Umgebung, den vom System angebotenen Dienst, sowie den Kontext der Benutzung [Mö05]. Qualität wird dabei als das „Ergebnis der Beurteilung der wahrgenommenen Beschaffenheit einer Einheit im Hinblick auf die erwünschte Beschaffenheit“ definiert [Je00]. Sie ist somit eine perzeptive Größe und setzt einen Wahrnehmungs- und Beurteilungsvorgang voraus.

In [Mö02] wurde eine Taxonomie der genannten Einflussfaktoren entwickelt, welche diese in Zusammenhang setzt mit verschiedenen Aspekten der Dienstequalität wie der Kooperativität, der Qualität der Spracheingabe und -ausgabe, der Symmetrie der Interaktion, der Effizienz bezogen auf die Kommunikation, die zu lösende Aufgabe und den Dienst insgesamt, sowie globalen Qualitätsaspekten wie der Zufriedenheit des Benutzers oder der Akzeptanz. Diese Taxonomie wird in [ITU03] zur Definition von Fragen eines

Beurteilungsbogens vorgeschlagen. Sie kann darüber hinaus auch zur Festlegung von Parametern zur Quantifizierung einzelner Qualitätsaspekte verwendet werden. Beide Anwendungen werden in [Mö05] am Beispiel eines telefonbasierten Dialogsystems für Restaurantauskünfte illustriert.

3 Evaluierungsmethoden

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 die Interaktionen aufgezeichnet und aus den Aufzeichnungen – nach einer Transkription und Annotation durch einen menschlichen Experten – sog. Interaktionsparameter bestimmt. Diese Interaktionsparameter beschreiben Eigenschaften des Systems, des Benutzers und der Interaktion zwischen beiden. Die Beschreibung steht zwar im Zusammenhang mit der wahrgenommenen Qualität, ist jedoch keine direkte Qualitätsmessung; sie gibt dem Entwickler aber wertvolle Hinweise auf Schwachstellen des Systems.

Zur Erfassung von Benutzerurteilen (d.h. direkten Qualitätsmesswerten) schlägt [ITU03] einen dreiteiligen Fragebogen vor, 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 einzelnen Test-Interaktionen mit dem System beantwortet und besteht aus Fragen zum gerade geführten Gespräch. Teil C erfasst den Gesamteindruck des Benutzers vom Dienst nach Ablauf aller Gespräche.

Die nach Transkription und Annotation durch einen menschlichen Experten bestimmten Interaktionsparameter sollten eine quantitative Beschreibung des Verlaufes des Dialogs, der Fähigkeit des Systems zu Meta-Kommunikation (Bestätigungen, Korrekturen, etc.), der Leistung der Spracheingabe, der Kooperativität des Systems, sowie seiner Fähigkeit, die gestellten Aufgaben zu lösen, liefern [ITU05]. Zu einigen Qualitätsaspekten (bspw. der Qualität der Sprachausgabe) sind allerdings noch keine Parameter bekannt. Es sei darauf hingewiesen, dass viele der Interaktionsparameter nicht unabhängig voneinander sind; der sehr umfangreiche Satz von Parametern sollte daher nach einer Übergangsphase reduziert werden, um nur noch informationstragende Parameter zu umfassen, welche tatsächlich relevant für die Bestimmung der Qualität sind.

4 Experimentelle Überprüfung

Die in [ITU03] beschriebene Methode zur Durchführung von Beurteilungsexperimenten sowie ein großer Teil der in [ITU05] und [Mö05] vorgestellten Interaktionsparameter wurden bei der Evaluierung des INSPIRE-Sprachdialogsystems verwendet. Das INSPIRE-System kann entweder über einen Telefonkanal betrieben werden (bspw. zur Steuerung von Hausgeräten während der Abwesenheit) oder direkt aus der häuslichen Umgebung (in der auch visuelles Feedback über einen Bildschirm und durch die bedienten Geräte gegeben wird). Letzteres Szenario wurde für die Evaluierung gewählt; die hier vorgestellten Ergebnisse beziehen sich also nicht auf ein telefonbasiertes System und unterstreichen die Gültigkeit des Ansatzes für eine größere Klasse von sprachbasierten und multimodalen Dialogsystemen.

Das Experiment wurde als Wizard-of-Oz-Versuch durchgeführt, bei dem die Spracherkennung und die Bedienung der Geräte vom Versuchsleiter übernommen wurde. 24 Versuchspersonen interagierten mit jeweils einer von drei unterschiedlichen Systemvarianten und gaben anschließend je 37 Urteile ab, die 7 Kategorien zuzuordnen waren (Gesamteindruck, Erreichen der gewünschten Ziele, Verständigung mit dem System, Verhalten des Systems, Gespräch, persönliche Wirkung, Benutzbarkeit des Systems). Zusätzlich wurden in einem Eingangsbogen Hintergrundinformationen abgefragt und in einem Abschlussbogen eine Beurteilung des Gesamtsystems gegeben. Die Gespräche wurden aufgezeichnet, transkribiert und annotiert, und daraus wurden 53 Interaktionsparameter bestimmt.

Die Einzelbeurteilungen nach jedem Dialog wurden einer Hauptkomponentenanalyse mit Varimax-Rotation und Kaiser-Normierung unterzogen. Dabei ergaben sich 8 Dimensionen, mit denen 72,8% der Varianz in den Urteilen erklärt werden konnte. Die Dimensionen erfassen (in der Reihenfolge der erklärten Varianz) die Akzeptanz, die kognitive Belastung des Benutzers, die Lösbarkeit der Aufgabe mit Hilfe des Systems, die Fehleranfälligkeit des Systems, die Benutzbarkeit des Systems, seine Kooperativität, die Natürlichkeit der Stimme und die Symmetrie der Interaktion (eine Dimension!), sowie die Geschwindigkeit der Interaktion. Offenbar decken diese Dimensionen eine große Anzahl von Qualitätsaspekten ab; der Fragebogen scheint also gut zur Erfassung einer Vielzahl qualitätsrelevanter Eigenschaften geeignet zu sein.

Die Zuverlässigkeit der Skalen zur Ermittlung dieser Dimensionen wurde anhand von Cronbach's α ermittelt. Dabei zeigten sich Werte von $\alpha \geq 0{,}7$ für alle Dimensionen außer Kooperativität, und $\alpha \geq 0{,}8$ für die ersten vier Dimensionen. Diese Werte weisen auf eine zufrieden stellende Reliabilität der verwendeten Beurteilungsskalen hin.

Die Korrelation zwischen Interaktionsparametern und Beurteilungen durch die Versuchspersonen war enttäuschend gering. Im höchsten Falle erreichte sie Werte um 0,6 (Spearman ρ). Vor allem die erkennungsbezogenen Parameter scheinen mit den Benutzerurteilen zu korrelieren; dies ist umso erstaunlicher, da die Erkennungsrate (bedingt durch den Einsatz des Wizards) nahezu perfekt war (gemessene Rate: 97,2%). Der Gesamteindruck korreliert nur mäßig mit der Dauer der Systemäußerungen (0,40), der Interpretationsrate (0,36…0,39) sowie der Erkennungsrate (0,39…0,42). Diese Korrelation reicht aber bei weitem nicht aus, um die direkten Beurteilungen von Versuchspersonen durch extern (instrumentell oder durch einen menschlichen Experten) bestimmte Interaktionsparameter zu ersetzen.

Obwohl die direkten Korrelationen gering sind, besteht die Möglichkeit, dass sich Gesamtqualität oder einzelne Qualitätsaspekte aus einer Kombination mehrerer Interaktionsparameter schätzen lassen. Diese Idee wird mit dem PARADISE-Modell verfolgt, welches versucht, einen Indikator der Gesamtqualität (Mittelwert über mehrere Beurteilungen) aus einer Linearkombination mehrerer Interaktionsparameter vorherzusagen [Wa97]. Aus den hier erhobenen Daten wurden Modelle mit unterschiedlichen Ziel- und Prädiktorvariablen mittels einer multivariaten linearen Regression berechnet. Auch dabei zeigte sich, dass subjektive Beurteilungen von Qualität nicht durch Parameter zu ersetzen sind: Nur maximal 48% der Varianz in den Beurteilungen ließ sich mit den linearen

Modellen abdecken. Daher werden auch in Zukunft beide Arten von Informationen – direkte Qualitätsbeurteilungen durch Versuchspersonen und Quantifizierung der Interaktionen durch Interaktionsparameter – zur vollständigen Evaluierung von Sprachdialogdiensten benötigt.

5 Schlussbemerkungen

Es wurde ein Überblick über Methoden zur Evaluierung telefonbasierter Sprachdienste gegeben, die in jüngster Zeit von der ITU-T empfohlen wurden oder als Empfehlung vorgesehen sind. Die Empfehlung umfasst derzeit eine Methode zur auditiven Bestimmung der Qualität synthetisierter Sprache [ITU94], eine Methode zur subjektiven Beurteilung kompletter Dialogsysteme [ITU03], sowie einen Vorschlag zur Bestimmung von Interaktionsparametern [ITU05]. Eine erste Anwendung im Rahmen des INSPIRE-Projektes zeigte, dass subjektive Beurteilungen der Qualität und instrumentelle oder expertenbasierte Messung von Interaktionsparametern unabhängige Informationen liefern; diese Verfahren sind also komplementär.

Die Liste von Interaktionsparametern bedarf einer weiteren Vervollständigung, Überprüfung und Kürzung, damit letztlich nur qualitätsrelevante Parameter zur Evaluierung empfohlen werden. Solche Parameter könnten dann auch zur Vorhersage von Qualität geeignet sein. Ob dabei lineare Modellansätze ausreichen muss in weiteren Untersuchungen geklärt werden.

Die vorgestellten Arbeiten wurden am Institut für Kommunikationsakustik (IKA) der Ruhr-Universität Bochum durchgeführt. Der Autor dankt der EU für die Förderung der experimentellen Arbeiten im Rahmen des IST-Projektes INSPIRE (www.knowledge-speech.gr/inspire-project/index.html) sowie Jan Krebber, Noha El Mehelmi, Jörn Opretzka und Rosa Pegam für die Hilfe bei der Durchführung des Experimentes und die Annotierung der Daten.

Literaturverzeichnis

[ITU94] ITU-T Rec. P.85: A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union, Geneva, 1994.

[ITU03] ITU-T Rec. P.851: Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, International Telecommunication Union, Geneva, 2003.

[ITU05] ITU-T Del. Contr. D.030: Proposal for Parameters Describing the Performance of Speech Technology Devices, Fed. Rep. Germany (Author: S. Möller), International Telecommunication Union, ITU-T SG12 Meeting, Jan. 18-27, Geneva, 2005.

[ITU05a] http://www.itu.int/ITU-T/studygroups/com12/q12roadmap/index.html.

[Je00] Jekosch, U.: Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung, Habilitationsschrift, Universität/GH Essen, 2000.

[Mö02] Möller, S.: A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems. In: Proc. 3rd SIGdial Workshop, Philadelphia PA, 142-153.

[Mö05] Möller, S.: Quality of Telephone-Based Spoken Dialogue Systems, Springer, New York NY, 2005.

[Wa97] Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In: Proc. ACL/EACL 35th Ann. Meeting, Madrid, 1997; 271-280.