# Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten

Sebastian Möller

*Deutsche Telekom Laboratories, TU Berlin, 10587 Berlin, Deutschland, Email: sebastian.moeller@telekom.de*

## Einleitung

Ein wichtiger Aspekt der Qualität sprachlicher Kommunikation ist ihre Effizienz, d.h. das Verhältnis von eingesetzten Ressourcen zum Grad der Erreichung des kommunikativen Zieles (vgl. z.B. Definitionen von ISO und ETSI in [1]). Effizienz spielt insbes. bei der Interaktion mit Sprachdialogdiensten eine große Rolle, bei denen das Ziel der Interaktion durch die Möglichkeiten des Systems vorgegeben ist, bspw. bei Auskunfts- und Reservierungssystemen oder beim Telefon-Banking. Zwei Arten von Effizienz sind hierbei zu unterscheiden [2]: Die Effizienz der sprachlichen Interaktion selbst (*communication efficiency*), d.h. bezogen auf den Verlauf des Dialogs zwischen Benutzer und System, und die Effizienz bei der Lösung der übergeordneten Aufgabe (*task efficiency*), d.h. bezogen auf das Ziel, das durch die Benutzung des Systems erreicht werden soll.

Im Rahmen des EU-geförderten IST-Projektes INSPIRE [3] wurde die Effizienz der Interaktion mit einem Smart-Home-System untersucht. Dazu wurden Benutzer zu verschiedenen Qualitätsaspekten befragt und es wurden Parameter extrahiert, die den Verlauf der Interaktion zwischen Benutzer und System quantitativ erfassen. Im Folgenden werden die Zusammenhänge zwischen den Benutzerurteilen und den Parametern analysiert und es wird versucht, Benutzerurteile aus den Parametern vorherzusagen.

## Dialogsystem und Versuchsdurchführung

Das INSPIRE-System gestattet die Bedienung verschiedener Hausgeräte (Leuchten, Rollos, TV, Videorekorder, Anrufbeantworter, etc.) über gesprochene Sprache. Es umfasst eine Signalvorverarbeitung, eine Spracherkennung, sprachverstehende Komponenten, einen Dialogmanager, eine Geräteschnittstelle sowie eine Sprachausgabe aus Bausteinen vorher aufgezeichneter Sprache. Für den beschriebenen Versuch stand ein Prototyp zur Verfügung, bei dem der Spracherkenner durch eine Online-Transkription ersetzt wurde.

24 Versuchspersonen (19-29 Jahre) führten jeweils 3 Dialoge mit dem System. Die Dialoge folgten Szenarien, die jeweils 9-11 unterschiedliche Bedienaufgaben umfassten. Nach jedem Dialog beurteilten die Versuchspersonen jeweils 37 Aussagen zu unterschiedlichen Qualitätsaspekten auf einem Fragebogen. Die Interaktionen wurden aufgezeichnet und durch einen Experten annotiert. Aus den Annotationen wurden 53 Parameter pro Dialog berechnet, die die Leistungen der Systemkomponenten und das Verhalten von Benutzer und System quantitativ beschreiben. Details zur Versuchsdurchführung, zu den Fragebögen und den Parametern finden sich in [4].

## Analyse der Ergebnisse

Die Urteile der Versuchspersonen wurden zunächst einer Hauptkomponentenanalyse mit Varimax-Rotation unterzogen. Dabei ergaben sich 8 Faktoren, die interpretierbar waren und 72,6% der Varianz abdeckten [4]. Die Faktoren wurden wie folgt benannt:

- C1: Akzeptanz (*acceptability*)
- C2: Kognitive Belastung (*cognitive demand*)
- C3: Aufgaben-Effizienz (*task efficiency*)
- C4: Fehleranfälligkeit des Systems (*system errors*)
- C5: Einfache Benutzbarkeit (*ease of use*)
- C6: Kooperativität (*cooperativity*)
- C7: Natürlichkeit der Systemstimme bzw. Symmetrie des Dialogs (*naturalness of system voice* bzw. *dialogue symmetrie*)
- C8: Schnelligkeit (*speed*)

Offensichtlich hängen die Faktoren C3 und C8 mit der Effizienz zusammen. C3 zeigt Korrelationen $|r| \geq 0{,}6$ mit den Aussagen 4.3 („Das System reagierte nicht immer wie erwartet“, $r = -0{,}76$), 2.2 („Die vom System gelieferten Informationen waren klar und deutlich“, $r = 0{,}65$) und 2.1 („Das System tat nicht immer das, was ich wollte“, $r = -0{,}64$), C8 mit der Aussage 4.1 („Das System reagierte zu langsam“, $r = -0{,}78$), außerdem etwas schwächer mit der Aussage 5.4 („Das Gespräch war zu lang“, $r = -0{,}53$).

Es wurden Korrelationen zwischen den Faktoren (C3, C8) bzw. den Einzelaussagen (2.1, 2.2, 4.1, 4.3, 5.4) und den extrahierten Parametern bestimmt (Spearman $\rho$). Signifikante Korrelationen ($p < 0{,}01$) ergeben sich zwischen folgenden Faktoren/Urteilen und Parametern:

- C3: Anzahl der Fragen des Systems ($\rho = -0{,}34$), Anzahl der Korrekturäußerungen des Benutzers ($\rho = -0{,}32$)
- C8: Parameter, die die Erkennungsleistung beschreiben (Wortfehlerrate etc.), $|\rho| = 0{,}36 \ldots 0{,}39$
- 2.1: Anzahl der inkorrekt geparsten Benutzeräußerungen ($\rho = -0{,}34$)
- 2.2: Anzahl der Äußerungen ($\rho = -0{,}32$), Anzahl der Fragen des Systems ($\rho = -0{,}34$), Anzahl der teilweise geparsten Benutzeräußerungen ($\rho = -0{,}32$)
- 4.1: Dauer der Benutzeräußerungen ($\rho = 0{,}36$)
- 4.3: Anzahl Cancel-Versuche des Benutzers ($\rho = -0{,}33$)
- 5.4: Parameter, die die Erkennungsleistung beschreiben (Wortfehlerrate etc.), $|\rho| = +/-0{,}35 \ldots 0{,}39$

Die beobachten Korrelationen sind recht begrenzt: Maximal werden Werte um 0,4 erreicht. Insbesondere ist der Zusammenhang zwischen den Urteilen zu Aussage 5.4 („Das Gespräch war zu lang“) und der Dialogdauer in s ($\rho = 0{,}09$)

sehr schwach. Gleiches gilt für die Beurteilung der Aussage 2.1 („Das System tat nicht immer das, was ich wollte“) und dem von einem Experten gemäß den Szenarien annotierten Task-Success-Parameter ($\rho$ = -0,15). D.h. die externe Erfassung von *communication efficiency* (z.B. durch Messung der Dauer des Dialogs) bzw. von *task efficiency* (z.B. durch Vergleich mit den Vorgaben des Szenarios) spiegelt nicht die Wahrnehmung des Benutzers wider. Stattdessen scheint die wahrgenommene Effizienz mit dem Fortschritt oder der „Glattheit“ des Dialogs zusammen zu hängen: Beeinflussende Größen sind z.B. Korrekturversuche des Benutzers, Erkennungsfehler und Parsing-Fehler.

Zur Messung von *communication efficiency* werden in der Literatur häufig die Dauer, Länge bzw. Anzahl der Äußerungen von Benutzer und System verwendet. Einige dieser Parameter korrelieren signifikant (Spearman $\rho$) mit den folgenden Benutzerurteilen bzw. Faktoren:

- Dauer der Systemäußerungen: Zufriedenheit des Benutzers (0,40), einfache Bedienbarkeit (0,34), Komfort (0,51) und lohnenswerte Benutzung (0,35); außerdem mit C1 (Akzeptanz, 0,38), C4 (Fehleranfälligkeit des Systems, 0,32), C5 (einfache Benutzbarkeit, 0,36) und C6 (Kooperativität, 0,32)
- Dauer der Benutzeräußerungen: C7 (Symmetrie, 0,33)
- Anzahl der Systemäußerungen: Information war klar (-0,32), System transparent (-0,34), keine Höranstrengung (-0,38), Annehmlichkeit (-0,33), einfache Bedienung (-0,37), einfach zu lernen (-0,45), Komfort (-0,32), System hilfreich (-0,44); C1 (Akzeptanz, -0,35), C2 (kognitive Belastung, -0,30), C5 (einfache Benutzbarkeit, -0,49) und C6 (Kooperativität, -0,36)
- Wörter pro Systemäußerung: Information vollständig (0,39), Bedienung einfach (0,35); C6 (Kooperativität, 0,37)
- Anzahl der Wörter des Systems: Transparenz (-0,31), keine Höranstrengung (-0,39), einfach zu lernen (-0,38), System hilfreich (-0,33); C5 (einfache Benutzbarkeit, -0,43)
- Wörter pro Benutzeräußerung: Transparenz (0,34), Komfort (0,35); C5 (einfache Benutzbarkeit, 0,42)

Wiederum finden sich keine signifikanten Korrelationen für die Gesamtdauer des Dialogs und für den Experten-annotierten Task-Success-Parameter, weder mit einzelnen Aussagen noch mit den extrahierten Faktoren. Offensichtlich sind diese Parameter nicht entscheidend für die wahrgenommene Effizienz, und sie korrelieren auch nicht mit anderen Qualitätsaspekten. Die Dauer der Systemäußerungen beeinflusst Akzeptanz, Benutzbarkeit und Kooperativität positiv, ihre Anzahl jedoch negativ.

## Modellierung

Trotz der geringen Korrelationen soll nun versucht werden, die mit der Effizienz zusammenhängenden Faktoren und Benutzerurteile aus Parametern mittels einer multivariaten linearen Regression vorherzusagen. Dazu wurden Regressionsmodelle berechnet, bei denen die Zielvariablen entweder Einzelurteile (bei einzelnen Aussagen) oder Mittelwerte über alle Urteile waren, die ≥ +/-0,4 mit den extrahierten Faktoren korrelierten. Die Eingangsparameter wurden Z-transformiert und schrittweise in die Gleichung eingeschlossen (Ersetzung fehlender Werte durch den Mittelwert, keine Konstante in der Regressionsgleichung). Die damit erzielten Varianzabdeckungen (korrigiertes $R^2_{corr}$) sind in Tab. 1 dargestellt.

**Tabelle 1**: Regressionsmodelle

| Zielvariable | | $R^2_{corr}$ |
|---|---|---|
| 2.1 | System tat nicht immer das, was ich wollte | 0,097 |
| 2.2 | Informationen waren klar und deutlich | 0,062 |
| 4.1 | System reagierte zu langsam | 0,211 |
| 4.3 | System reagierte nicht immer wie erwartet | 0,111 |
| 5.4 | Gespräch war zu lang | 0,111 |
| C1 | Akzeptanz | 0,453 |
| C2 | Kognitive Belastung | 0,424 |
| C3 | Aufgaben-Effizienz | 0,221 |
| C4 | Fehleranfälligkeit des Systems | 0,342 |
| C5 | Einfache Benutzbarkeit | 0,440 |
| C6 | Kooperativität | 0,290 |
| C7 | Natürlichkeit der Stimme bzw. Symmetrie | 0,316 |
| C8 | Schnelligkeit | 0,141 |

Die Ergebnisse bestätigen, dass sich die Effizienz-bezogenen Benutzerurteile bzw. Faktoren (grau unterlegt) nur schlecht aus den Parametern vorhersagen lassen. Demgegenüber ist die Vorhersagekraft für andere Faktoren deutlich besser.

## Diskussion

Die beschriebenen Analysen zeigen nur geringe Korrelationen zwischen gemessenen Parametern und Benutzerurteilen zur Effizienz. Dies könnte zwei Ursachen haben: (1) Die bislang betrachteten Urteile und/oder Parameter erfassen nicht wirklich „Effizienz“; oder (2) der Zusammenhang zwischen Benutzerurteilen und Parametern ist komplexer, als bislang vielfach angenommen wird. Wahrgenommene Effizienz scheint sich im Gesamt-Fortschritt des Dialogs widerzuspiegeln, und hängt nicht direkt mit der Dauer des Dialogs oder der Dauer und Anzahl der Äußerungen zusammen. Letztere haben aber durchaus einen Einfluss auf Benutzbarkeit, Kooperativität und Akzeptanz.

Der Versuch wurde am Institut für Kommunikationsakustik der Ruhr-Universität Bochum durchgeführt. Besonderer Dank gilt Jan Krebber und Rosa Pegam für die Unterstützung bei der Durchführung und die Annotierung der Daten.

## Literatur

[1] ETSI Technical Report ETR 095. Human Factors (HF); Guide for Usability Evaluation of Telecommunication Systems and Services. ETSI, Sophia Antipolis, 1993

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

[3] IST-Projekt INSPIRE. URL: http://www.knowledge-speech.gr/inspire-project/index.html

[4] Möller, S., Smeele, P., Boland, H., Krebber, J. Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study. Computer Speech and Language, zur Veröffentlichung angenommen.