Sprachapplikationen: Qualität durch Standardisierung

Cornelia Hipp
Fraunhofer IAO
Nobelstraße 12
70569 Stuttgart
cornelia.hipp@iao.fraunhofer.de
www.iao.fraunhofer.de

Matthias Peissner
Fraunhofer IAO
Nobelstraße 12
70569 Stuttgart
matthias.peissner@iao.fraunhofer.de
www.iao.fraunhofer.de

Abstract

Die Branche der Sprachtechnologie entdeckt ein Defizit an standardisierter Qualitätssicherung von Sprachapplikationen. Unter der Leitung von Fraunhofer IAO und Initiative Voice Business wurde eine Qualitätsoffensive ins Leben gerufen, um gemeinsam mit führenden Experten der deutschsprachigen Fachszene Kriterien, Maße und Verfahren zur Messung und Optimierung der Qualität von Sprachapplikationen zu definieren. Konzepte und Methoden der Usability spielen dabei eine zentrale Rolle.

Keywords

Sprachinteraktion, Qualitätskriterien, Maß, Testverfahren, Voice User Interface

1.0 Einleitung

Im Rahmen eines Verbundprojekts unter der Leitung von Fraunhofer IAO und mind Business Consultants ist ein offener Qualitätsstandard für Sprachapplikationen erarbeitet worden. In einem interdisziplinären Team von 22 anerkannten Experten wurden Kriterien sowie Maße und Verfahren zur Erhebung und Optimierung der Qualität von Sprachapplikationen definiert und beschrieben. Die erarbeiteten Ergebnisse entsprechen damit einem breiten Konsens innerhalb der deutschsprachigen Fachszene. Aktuelle Erkenntnisse aus Wissenschaft und Forschung wurden dabei ebenso berücksichtigt wie die Anforderungen der Industrie nach Praxistauglichkeit und Wirtschaftlichkeit.

Sprachapplikationen haben ein hohes Potential, da sie nutzbar im mobilen Kontext sind, lediglich ein Telefon von überall als Eingabemedium ist erforderlich und theoretisch bietet Sprachinterkation eine schnelle und intuitive Nutzung, da mittels Sprache direkt gesagt werden kann, was das Anliegen des User ist. Weiterhin bietet die Sprachinteraktion für alte Menschen & visuell eingeschränkte Personengruppen, sowie in Nutzungskontexten, wenn keine Hände frei sind, ein großes Potential für die Nutzung, sprich wenn die Bedienung eines grafischen User Interfaces schwer fällt. Doch trotz des großen Potentials ist die Nutzung in Deutschland hinter ihren Erwartungen geblieben. Eine Studie des Fraunhofer IAO zeigt diesbezüglich auch die geringe Akzeptanz von Sprachapplikationen auf (Peissner, et al., 2006). Ein Ansatz um die Akzeptanz zu verbessern, ist die Qualität von Sprachapplikationen zu erhöhen. Doch anders als in Bereichen der Softwaretechnik oder Ingenieur-Fachdisziplinen, ist innerhalb der Sprachtechnologie noch keine strukturierte Qualitätssicherung verankert. Die Frage, welche Besonderheiten innerhalb eines Entwicklungsprozesses für Sprachapplikationen bestehen und wie dieser mit hoher Prozess- und Produktqualität gemeistert werden kann, fängt die Branche erst an zu überlegen. Es gibt zwar vereinzelt Literatur zu dieser Thematik (Möller (2005), Hempel (2008), Dybkiaer et al. (2007)), dahingegen startet die Branche der Sprachtechnologie nun hingegen gemeinschaftlich diese Thematik aufzugreifen und hat den Wunsch einen gemeinschaftlichen Konsens zu bilden. Damit ist die Bestrebung verknüpft eigene Sprachapplikationen von schlechten Applikationen abzugrenzen und nach außen mit einer Norm bewerten zu können. Weiterhin soll mit Hilfe der Qualitätskriterien eigene Sprachapplikationen im Vergleich zu anderen eingeordnet werden können.

2.0 Vorgehensweise

Um eine hohe Qualität, Realisierbarkeit und Validität der erarbeiteten Ergebnisse zu gewährleisten und um von Beginn an eine hohe Akzeptanz im Markt zu sichern, wurden im Rahmen eines 2-tägigen Workshops Experten aus einschlägigen deutschen Unternehmen beteiligt, die innerhalb der Sprachtechnologie tätig sind.

Zur Vorbereitung des Workshops reichten die beteiligten Experten Position Papers ein, in denen Qualitätskriterien, entsprechende Messgrößen sowie Evaluations- und Optimierungsverfahren beschrieben wurden. Weiterhin konnten die Experten Dokumente, die aus anderen Projektzusammenhängen oder Publikationen bereits zur Verfügung standen, einreichen, soweit sie für die behandelten Fragestellungen relevant schienen.

Die Experteneinreichungen wurden von Fraunhofer IAO ausgewertet, mit dem Ziel, die folgenden Expertenworkshops vorzubereiten und eine inhaltliche

Grundlage für die angestrebte Publikation zu schaffen. Die weitere Vorgehensweise zur Konsolidierung der Inhalte des Dokuments gestaltete sich wie folgt:

2.1 Qualitätskriterien

Fraunhofer IAO erarbeitete auf Grundlage der Experteneinreichungen einen Vorschlag, der während des Workshops im Plenum diskutiert und verabschiedet wurde.

2.2 Komponenten einer Sprachapplikation

Im Vorfeld des Workshops kristallisierte sich Diskussionsbedarf bezüglich einer Einigung zu den Applikationskomponenten heraus. Verschiedene technische Realisierungen wurden innerhalb einer speziell dafür gebildeten Arbeitsgruppe beleuchtet. Der resultierende Konsens war, dass die Zuordnung der Maße und Verfahren mit Hilfe einer funktionalen Beschreibung erfolgen soll, die unter 3.2 Komponenten einer Sprachapplikation beschrieben ist.

2.3 Themenbereiche

Sehr elementar wurden zu Beginn die Themenbereiche identifiziert, die relevant für die Qualitätssicherung waren – stets aus Sicht von Sprachapplikationen:

- Strategie & Business Logik: Ein wesentliches Erfolgskriterium einer Sprachapplikation ist deren wirtschaftlicher Ertrag. Dessen Messbarkeit sowie die Abschätzung von zukünftigen Investitionen und Einsparungspotenzialen sind wesentliche Gesichtspunkte dieses Themenschwerpunkts.
- Sprachtechnologie und Linguistik: Die Qualität der Sprachtechnologie ist ein essenzieller Aspekt in der Nutzung von Sprachapplikationen. Die zuverlässige Erkennung der gesprochenen Benutzereingaben und eine verständliche sowie angenehme Sprachausgabe sind zentrale Voraussetzungen für eine komfortable Interaktion.
- Dialogplattformen und Integration: Die Nutzung einer Sprachapplikation setzt deren Erreichbarkeit grundlegend voraus. Dafür ist die korrekte Integration der Sprachapplikation und der einzelnen Systemkomponenten in das Gesamtsystem erforderlich. Zusätzlich lassen Aspekte wie geringe Latenzzeiten oder die Nutzbarkeit der Sprachapplikationen auch bei Spitzenbelastungen auf eine sinnvolle technische Umsetzung schließen.
- Voice User Interface und Usability: Die erfolgreiche und angenehme Nutzung einer Sprachapplikation hängt insbesondere auch von der angemessenen Gestaltung der Benutzungsoberfläche, dem Voice User Interface, ab. Aspekte wie z.B. Wortwahl und Formulierungen der Systemausgaben oder das Management von Fehlern sind dabei relevant.

Diese Gliederung verdeutlicht den ganzheitlichen Ansatz, welcher innerhalb des Projektes genutzt wurde, allerdings mit den Prämissen, dass es sich um Telefoniersprachanwendungen handelt und somit nicht anwendbar auf multimodale Anwendungen ist. Eine Zuordnung zu einzelnen Applikationstypen wurde ebenfalls noch nicht vollzogen und als Option für weitere Schritte aufgenommen.

2.4 Entwicklungsphasen

Um die erarbeiteten Maße und Verfahren dem Entwicklungsprozess zuzuordnen einigte man sich im Vorfeld auf einzelnen Phasen. Der gewonnene Konsens war, dass die Erstellung einer Sprachapplikationen sich in verschiedene Schritte unterteilen lässt, die folgende Phasen umfasst:

- Projektvorbereitung & Analyse
- Konzept & Design
- Implementierung
- Integration & Inbetriebnahme
- Betrieb

Diese einzelnen Phasen sind nicht streng voneinander getrennt, sondern werden in einem iterativen Prozess mehrfach (je nach Bedarf) durchlaufen. Der Grund hierfür ist die schwere Realisierbarkeit einer linearen Abfolge von Prozessschritten, da verschiedene Aspekte der einzelnen Phasen erst bei späteren Schritten aufgedeckt werden. Durch den iterativen Prozess können sukzessive Teilergebnisse verbessert und korrigiert werden. Während aller Prozessphasen werden zusätzlich kontinuierlich Tests durchgeführt, deren Ergebnisse direkt in die Entwicklung mit aufgenommen werden. Die Testung bereits in frühen Phasen verhindert ein zu spätes Aufdecken von Fehlern.

3.0 Erarbeitete Ergebnisse

Als zentrales Projektergebnis wurden die wichtigsten qualitätsrelevanten Maße und Verfahren zur Erhebung und Optimierung der Qualität von Sprachapplikationen identifiziert und beschrieben. In einer strukturierten Form wurden die beschriebenen Maße und Verfahren eingehend charakterisiert und bezüglich ihres Potenzials für den Praxiseinsatz und ihrer Wirtschaftlichkeit bewertet. Diese Übersicht soll dem Profi als Nachschlagewerk dienen und dem interessierten Laien einen umfassenden Überblick bieten.

Ein weiteres Ergebnis des beschriebenen Projekts ist die Verständigung auf eine gemeinsame Terminologie für zentrale Konzepte im Umfeld der Entwicklung und Qualität von Sprachapplikationen. Zur Förderung eines gemeinsamen Sprachgebrauchs und zur Steigerung der Transparenz auch für Laien und potenzielle Kunden wurde eine generische Systemarchitektur mit den Komponenten einer Sprachapplikation entwickelt. Darüber hinaus wurden die Pha-

sen eines Standardentwicklungsprozesses definiert.

Abb 1: Leitfaden Qualitätskriterien, Maße und Verfahren für Sprachapplikationen Qualitätskriterien

Die in diesem Abschnitt aufgeführten Kriterien zielen auf eine ganzheitliche Qualitätsbewertung von Sprachapplikationen. Sowohl wirtschaftliche Aspekte als auch die Qualität der Sprachtechnologie, die Integration einzelner Komponenten sowie die Gestaltung des Voice User Interfaces werden abgedeckt. Die Kriterien beziehen sich nicht auf einzelne Komponenten, sondern sind auf die Bewertung des Gesamtsystems ausgelegt.

Die einzelnen Qualitätskriterien stehen in engen Wechselbeziehungen. Daraus ergeben sich teilweise Zielkonflikte, welche die gleichzeitige Erfüllung aller resultierenden Anforderungen einschränken. So kann beispielsweise ein umfangreiches Funktionsangebot zu einer komplexen Navigationsstruktur führen und damit die schnelle Zielerreichung einschränken.

Darüber hinaus sind die einzelnen Qualitätskriterien je nach Zielsetzung und Applikationsausrichtung unterschiedlich zu priorisieren. Je nach dem, ob bestehende Dienste automatisiert werden, oder ob durch einen Mehrwertdienst neue Businesskomponenten eingeführt werden, treten beispielsweise unterschiedliche Kriterien mehr oder weniger stark in den Vordergrund. Ebenso können sich die Betreiberziele bezüglich der Gesprächsdauer unterscheiden je nach dem ob eine effiziente und schnelle Bearbeitung gewünscht wird oder ob das Geschäftsmodell einer Entertainment Anwendung auf längere Gesprächsdauer ausgerichtet ist.

Nachfolgend werden die identifizierten Qualitätskriterien aufgeführt:

- Angemessener Funktionsumfang und Inhaltsangebot:
  »Eine Sprachapplikation ist gut, wenn die Sprachapplikation durch ein attraktives und vollständiges Funktionsangebot einen Mehrwert für die Kunden schafft.«
- Einwandfreie Funktionsfähigkeit und Leistungsfähigkeit:
  »Eine Sprachapplikation ist gut, wenn ein sicheres, performantes und fehlerfreies Funktionieren des Systems gewährleistet ist – auch bei zu erwartenden Belastungsspitzen.«
- Administrierbarkeit und effizienter Betrieb:
  »Eine Sprachapplikation ist gut, wenn technische Aufwände nach Inbetriebnahme des Systems minimal gehalten werden können.«
- Erweiterbarkeit und Skalierbarkeit:
  »Eine Sprachapplikation ist gut, wenn die Systemarchitektur zukünftige Erweiterungen und Veränderungen leicht ermöglicht.«
- Wirtschaftlichkeit:
  »Eine Sprachapplikation ist gut, wenn der Betrieb der Sprachapplikation wirtschaftlich rentabel ist.«
- Zuverlässige Erkennung der Benutzeräußerungen:
  »Eine Sprachapplikation ist gut, wenn die Spracherkennung einen angemessenen Umfang zu erwartender Benutzeräußerungen zuverlässig erkennt.«
- Effektives Fehlermanagement:
  »Eine Sprachapplikation ist gut, wenn Erkennungsfehler und Bedienfehler keinen großen Schaden anrichten.«
- Effektive und flexible Dialogabläufe:
  »Eine Sprachapplikation ist gut, wenn die Navigationsstruktur die Benutzer unterstützt, schnell und sicher ihr Ziel zu erreichen.«
- Verständliche und zielführende Systemausgaben:
  »Eine Sprachapplikation ist gut, wenn die akustischen Systemausgaben den Benutzer bei der Orientierung und der Formulierung zielführender Äußerungen unterstützen.«
- Anmutung und emotionale Adressierung:
  »Eine Sprachapplikation ist gut, wenn eine positive und angemessene Einstellung des Benutzers gegenüber der Sprachapplikation, ihrer Nutzung und ihres Betreibers erzielt wird.«

3.1 Komponenten einer Sprachapplikation

Die Realisierung einer Sprachapplikation stellt das Zusammenspiel verschiedener Teilkomponenten dar, die für unterschiedliche Funktionsbereiche innerhalb der Sprachapplikation verantwortlich sind. Diese Komponenten wurden innerhalb dieser Qualitätsoffensive nach dem etablierten Software-Architekturmuster »Model-View-Control« nach funktionalen Kriterien aufgeteilt. Dabei wurde bei Auflistung der Maße und Verfahren angegeben, welchen logischen Bereichen des Model-View-Control Entwurfsmusters sie zugeordnet sind. Es erfolgte keine Zuordnung zu einzelnen technischen Komponenten, da diese teilweise mit der konkreten Implementierung und der verwendeten technischen Plattform eng zusammenhängen und deshalb nicht immer zugeordnet werden können.

Zusätzlich wurde das Model-View-Control-Konzept für die Bedürfnisse von Sprachapplikationen erweitert um die Komponente »Access«. Diese stellt die Telefonie-Funktionalität dar, welche für die Verbindungs-Erstellung von einer Sprachapplikation mit Hilfe des Telefonnetzwerkes notwendig ist.

3.2 Maße

Es wurden gemeinsam 32 Maße erarbeitet, welche den in 2.3 genannten Themenschwerpunkten »Strategie und Business Logik«, »Sprachtechnologie und Linguistik«, »Dialogplattformen und Integration« und »Voice User Interface und Usability« zugeordnet wurden. Dabei konnten einzelne Maße auch verschiedenen Themenbereichen zugeordnet werden. Weiterhin wurden bei den Maßen folgende Beschreibungsmerkmale erfasst:

- Synonyme
- Kurzbeschreibung
- Auf welche Qualitätskriterien Bezug genommen wird
- Zuordnung zu Applikationskomponenten
- Aktueller Praxiseinsatz
- Potenzial für Praxiseinsatz
- Verfahren, um das Maß zu erheben
- Bewertung der Wirtschaftlichkeit

Auszugsweise wird anhand von einem Beispiel diese Beschreibungsstruktur verdeutlicht:

3.2.1 Effektives Fehlermanagement

- Synonyme:
  Keine
- Kurzbeschreibung:
  Dieses Maß bewertet die Fehlerrobustheit des Sprachdialogs. Trotz auftretender Erkennungsfehler kann das Ziel sicher und mit nur geringem zusätzlichem Zeitaufwand erreicht werden.
  Zur Erhebung des Maßes werden die Effizienzeinbußen berechnet, die sich ergeben, wenn sich ein zusätzlicher Fehler in der Sprachapplikation befindet. Diese Effizienzeinbußen werden für alle Fehler berechnet und anschließend aufsummiert. Letztlich wird der Durchschnitt von allen unterschiedlichen Effizienzeinbußen berechnet, indem die gewonnene Summe durch die Anzahl der Erkennungsfehler dividiert wird.
- Auf welche Qualitätskriterien Bezug genommen wird:
  Fehlermanagement, Effektive & flexible Dialogabläufe
- Zuordnung zu Applikationskomponenten:
  View
- Relevanz Themenschwerpunkte:
  Strategie und Business Logik, Voice User Interface und Usability
- Aktueller Praxis-einsatz:
  gelegentlich, selten
- Potenzial für Praxiseinsatz:
  mittel
- Verfahren, um das Maß zu erheben:
  Logfile-Analyse, unter dem Vorbehalt der verlässlichen Erhebung der Task Completion. Die erforderlichen Zahlen lassen sich (in kleinerem Maßstab) auch bereits aus Usability Tests und aus Friendly User Tests gewinnen. Weitere relevante Verfahren sind das Rapid Prototyping und der Wizard of Oz Test.
- Bewertung der Wirtschaftlichkeit:
  Note 3

3.3 Verfahren

Im Rahmen des Workshops wurden von den Experten 23 Verfahren erarbeitet, mit Hilfe deren die Maße erhoben werden können, welche im vorangegangenen Unterkapitel 3.3 dargestellt sind. Bei der Definition der Verfahren wurde darauf geachtet, alle Entwicklungsphasen (vgl. Unterkapitel 2.4) einer Sprachapplikationen mit einzuschließen und auch alle Themenbereiche (vgl. Unterpunkt 2.3). Die Beschreibungsmerkmale der Verfahren wurden im Vergleich zu denen der Maße um die folgenden Punkte erweitert:

- Welche Maße können erfasst bzw. optimiert werden?
- Zuordnung zu Entwicklungsphasen
- Reifegrad des Verfahrens
- Anforderungen des Verfahrens

Auszugsweise wird analog zu den Maßen nachfolgend eines der insgesamt 23 Verfahren aufgeführt:

3.3.1 Friendly User Test

- Synonyme:
  keine
- Kurzbeschreibung:
  Potentielle Nutzer testen das System vor Betrieb und agieren dabei kooperativ. Dieser Test erbringt erste Nutzungsdaten und Feedback vor der Live Schaltung. Hier können gravierende Fehler im Betrieb rechtzeitig erkannt und behoben werden.
- Welche Maße können erfasst bzw. optimiert werden:
  Die gesamte Applikation und alle Funktionalitäten und Qualitäten werden getestet, also auch die Anbindung an Backendsysteme und beispielsweise Transaktionen. Gegebenenfalls können auch Teilbereiche bereits in einem frühen Stadium getestet werden. Zusätzlich können das Effizienzmaß der Bedienung und das Maß für effektives Fehlermanagement gemessen werden.
- Auf welche Qualitätskriterien Bezug genommen wird:
  Funktionalität und Inhaltsangebot, Funktions- und Leistungsfähigkeit, Administrierbarkeit und Betrieb, Zuverlässige Erkennung, Fehlermanagement, Effektive & flexible Dialogabläufe, Verständlichkeit & Zielführung, Anmutung und Emotion
- Zuordnung zu Appliakationskomponente:
  Betroffen sind die Spracherkennung, die semantische Analyse, die Ausgabegenerierung und der Content & Daten; somit sind Model, View und Control betroffen
- Relevanz Themenschwerpunkte:
  Strategie und Business Logik, Voice User Interface und Usabili-

ty; Sprachtechnologie und Linguistik
- Zuordnung zu Entwicklungsphasen:
  Projektvorbereitung & Analyse, Konzept & Design sowie Implementierung
  Kommentar dazu ist, dass dieses Verfahren in unterschiedlichen Projektphasen einsetzbar ist, in der Regel aber vor dem Live Betrieb eingesetzt wird.
- Reifegrad des Verfahrens:
  hoch
- Aktueller Praxiseinsatz:
  immer, Standard
- Potenzial für Praxiseinsatz:
  hoch
- Anforderungen des Verfahrens:
  Es muss ein funktionsfähiger Prototyp vorhanden sein. Tests können auf unterschiedlichen Ebenen in unterschiedlichen Stadien durchgeführt werden: auf reiner Spracherkennungsebene (Vokabular), auf Dialogebene, auf Anwendungsebene (inklusive Transaktionen und Interaktion mit Fremdsystemen).
- Bewertung der Wirtschaftlichkeit:
  Note 1-2: Allumfassend; Ergebnisse sind wichtig, bevor System abgenommen werden kann.

4.0 Danksagung

Das Verbundprojekt unter der Leitung Fraunhofer IAO und mind Business Consultants ist mit freundlicher Unterstützung von Semantic Edge, Sikom, Sympalog, Telenet und VMA entstanden. Zum erfolgreichen Gelingen möchten wir uns bei den mitwirkenden Experten bedanken: Ludovica De Sio (IBM Entwicklung GmbH), Dr. Christian Dugast (zum Zeitpunkt selbstständig), Dr. Carsten Günther (zum Zeitpunkt IBM Entwicklung GmbH), Mark Gutmann (STRATECO GmbH & Co. KG), Dr. Jürgen Haas (Sympalog Voice Solutions GmbH), Tom Houwing (Voiceandvision B.V.), Markus Kesting (Telenet GmbH Kommunikationssysteme), Dietmar Kneidl (Sikom Software GmbH), Jörn Kreutel (SemanticEdge GmbH), Dr. Guntbert Markefka (T-Mobile Deutschland GmbH, VMA), Dr. Marion Mast (zum Zeitpunkt IBM Entwicklung GmbH), Jürgen Mehring (Sparda Bank Hamburg eG), Prof. Dr. Sebastian Möller (Deutsche Telekom AG Laboratories, TU Berlin), Frank Oberle (T-Systems Enterprise Services GmbH), Dr. Lupo Pape (SemanticEdge GmbH), Andreas Schaub (Unisys Deutschland GmbH, VMA), Bernhard Steimel (mind Business Consultants), Paul Hubert Vossen (Voice & Visual Design), Dr. Frank Wanning (HFN Medien GmbH, VMA)

5.0 Ausblick

Um bestehende Forschungslücken weiterhin zu schließen und den Nutzwert des ersten Projekts für die Praxis weiter zu erhöhen, haben sich die Projektbeteiligten entschieden, ihr Vorhaben 2008 weiter voranzubringen und haben sich bereits auf die thematischen Schwerpunkte der zweiten Forschungsphase geeinigt. Im Mittelpunkt steht dabei die eingehende Betrachtung unterschiedlicher Applikationstypen und die Erarbeitung eines entsprechenden Merkmalkatalogs. Anschließend soll erarbeitet werden, welche der in der ersten Forschungsphase identifizierten Methoden sich zur Qualitätsoptimierung für die verschiedenen Applikationstypen am besten eignen.

Wie bereits im ersten Jahr ist das Ziel zahlreiche Experten zu gewinnen und dadurch unter Beteiligung eines möglichst breit aufgestellten Expertenkreises die Akzeptanz in der Praxis zu erhöhen.

6.0 Literaturverzeichnis

Dybkjaer, L.; Hemsen, H.; Minker, W. (2007). Evaluation of Text and Speech Systems. Dordrecht: Springer.

Hempel, T (Ed.). (2008). Usability of Speech Dialog Systems. Berlin: Springer-Verlag.

Möller, S. (2005). Quality of Telephone-Based Spoken Dialogue Systems. New York: Springer Sciene + Business Media.

Peissner, M.; Sell, D.; Steimel, B. (2006). Akzeptanz von Sprachapplikationen. Stuttgart: Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO.

Peissner, M.; Hipp, C.; Steimel, B. (2007). Qualitätskriterien, Maße und Verfahren für Sprachapplikationen. Stuttgart: Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO.

http://www.hci.iao.fraunhofer.de/de/projekte/qualitatskritrien_masse_und_verfahren_fur_ sprachapplikationen/index.html

Interaktion von Morgen