SPRACHSTEUERUNG EINES COMPUTERSPIELS - UNTERSUCHUNGEN ZUR LEISTUNGSFÄHIGKEIT UND ERGONOMIE

Daniel Sobe, Matthias Eichner

Technische Universität Dresden
daniel.sobe@epost.de, matthias.eichner@ias.et.tu-dresden.de

Abstract:

Die heutzutage verfügbaren Systeme zur Spracherkennung und Sprachsynthese befinden sich auf einem sehr hohen Entwicklungsstand. Ungeachtet dessen werden sie aber in der Praxis eher wenig eingesetzt. Begründet wird dies oft mit der zu hohen Fehlerkennungsrate der Spracherkennung resp. der geringen Qualität der synthetischen Sprache.

Eine am Institut für Akustik und Sprachkommunikation der Technischen Universität Dresden angefertigte Diplomarbeit mit dem Titel “Sprachsteuerung eines Computerspiels - Untersuchungen zur Leistungsfähigkeit und Ergonomie” sollte anhand eines konkreten Implementierungsversuches feststellen, inwieweit die Spracherkennung und Sprachsynthese auf dem heutigen Stand bereits einsetzbar sind.

Die Untersuchung orientierte sich dabei an den relevanten ergonomischen Eigenschaften, die die Bedienung des Computerspiels vor und nach dem Einbau der Sprachtechnologie aufweist. Eine möglichst hohe Nutzerfreundlichkeit der Sprachbedienung konnte dadurch erreicht werden, indem das zur Bedienung nötige Vokabular durch Simulation im voraus ermittelt wurde. Das fertige Gesamtsystem wurde durch mehrere Testpersonen evaluiert und mit der herkömmlichen Bedienung verglichen.

Die erhaltenen Ergebnisse zeigen eine hohe Akzeptanz der Sprachsynthese als zusätzliche Funktion zu auf dem Bildschirm ausgegebenen Text. Bei der Spracherkennung ist keine eindeutige Tendenz zu erkennen, die jeweilige Präferenz der Testperson hängt von der persönlichen Wichtung der ergonomischen und hedonistischen Eigenschaften der Bedienung ab. Auffällig ist, dass die Güte der Sprachtechnologie etwa gleich relevant für diese Entscheidung ist wie die Anpassung der Applikation an diese Technologie.

1 Einleitung

Ziel der hier vorgestellten Diplomarbeit[1] war, anhand einer konkreten Implementierung festzustellen welche Besonderheiten zu beachten und welche Schwierigkeiten zu meistern sind, wenn der heutige Stand von Sprachtechnologie in Applikationen eingesetzt werden soll. Dafür wurden die am Institut für Akustik und Sprachkommunikation vorhandenen Technologien verwendet, bestehend aus dem Dresdener Sprachsynthesesystem “Dress”[2] und dem “UASR-Kommandoworterkenner”. Diese Technologien stehen in einem Rahmenprogramm als so genannter “SpeechServer”[3] zur Verfügung. Somit konnten bei der Integration in das

Computerspiel die Vorteile einer generische API und des Client-/Server-Konzeptes ausgenutzt werden.

Als Untersuchungsobjekt wurde das Computerspiel “Krabat” verwendet. Dabei handelt es sich um ein von der Projektgruppe “RAPAKI” im Auftrag der Stiftung für das Sorbische Volk erstellte Abenteuerspiel (ein so genanntes “Adventure”). Die Bedienung dieses Spiels erfolgt mit der Maus, die Ausgabe der Informationen geschieht durch Grafik und Text. Während die Mausbedienung komplett durch eine Sprachbedienung ersetzt werden sollte, war der Einsatz der Sprachsynthese als Zusatz zu dem auf dem Bildschirm angezeigten Text geplant.

2 Ergonomische Betrachtungen

2.1 Grundlagen

Die Implementierung der Sprachtechnologie in das Computerspiel sollte unter Berücksichtigung ergonomischer Eigenschaften erfolgen. Die Ergonomie umfasst viele Teilgebiete, was eine Einschränkung erforderlich machte. Da es sich bei dem Untersuchungsobjekt um ein Computerspiel handelt, wurden ausschließlich softwareergonomische Aspekte betrachtet. Weitere Einschränkungen ergaben sich aus der Feststellung, dass für die Aufgabe nur solche Teilgebiete der Softwareergonomie relevant sind, die bei der Implementierung der Sprachtechnologie Veränderungen erfahren. Eine Zusammenfassung der verbleibenden Teilgebiete (ausgewählt aus [4]) stellt Tabelle 1 dar.

Teilgebiet	Thema	Beispiele
Dialogtechniken	Informationsdarstellung	visuelle Darstellung akustische Darstellung
	Interaktionsformen	deskriptive Form deiktische Form
Dialogparadigmen	Kommandos	formale Sprache natürliche Sprache Symbolsprache
	direkte Manipulation	
Wahrnehmungspsychologie	visuelle Wahrnehmung akustische Wahrnehmung	Gestaltungsgesetze Sprachausgabe
Kognitionspsychologie	Gedächtnis	Erkennen Erinnern
	Verstehen	
	Handeln	explorativer Vorgehen

Tabelle 1: Ausgewählte Themen der Software-Ergonomie

Die Evaluierung softwareergonomischer Kriterien ist sehr stark abhängig von den Anforderungen, die an das System, den Benutzer und die Anwendung gestellt werden. Beispielsweise unterscheiden sich die Anforderungen an eine Applikation, die demselben Einsatzzweck dienen soll, bei der Implementierung zwischen einem PC und einem Mobiltelefon aufgrund der unterschiedlichen Voraussetzungen. Unterschiedliche Anforderungen an den Benutzer kann man verdeutlichen, indem man die jeweiligen Altersgruppen und die signifikant unterschiedlichen Vorkenntnisse bei der Benutzung von Software bedenkt. Eine typische Anforderung an das jeweilige System ist eine ausreichend geringe Antwort- oder zumindest Reaktionszeit, was zumindest bei Geräten für den mobilen Einsatz auch heute noch relevant ist.

Allgemeine Aussagen zur Ergonomie sind also, wie eben dargelegt, nur sehr schwer möglich. Aus diesem Grund haben die in der DIN EN ISO 9241 mit dem Titel “Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten” festgehalteten Anforderungen auch eher Vorschlagscharakter. Um spezielle Aussagen über die Ergonomie treffen zu können dürfen sich die Anforderungen an den Benutzer, die Anwendung und das System nicht signifikant ändern.

2.2 Anwendung auf das Computerspiel

Um Aussagen über die ergonomischen Eigenschaften beim Dialog des Benutzers mit dem Computerspiel treffen zu können, werden folgende Voraussetzungen für die Anforderungen an System, Anwendung und Benutzer geschaffen:

1. Das System ist ein PC mit heutzutage üblicher Rechenleistung[1]. Angeschlossen sind eine Computermaus, ein Mikrofon (Headset) und Lautsprecher.
2. Die Anwendung ist das bereits genannte Computerspiel.
3. Die Benutzer (Testpersonen) haben unterschiedliche Voraussetzungen beim Umgang mit Computern und Sprachtechnologie. Gemeinsamkeiten bestehen im Interesse für die Sprachtechnologie und das Computerspiel.

In diesem Fall kann man die relevanten ergonomischen Kriterien für die Bedienung des Computerspiels wie in Tabelle 2 dargestellt zusammenfassen.

trifft eher auf Sprache zu	trifft eher auf Maus zu
natürlich	künstlich
unkontrollierbar	kontrollierbar
einfach	kompliziert
fehleranfällig	fehlerfrei
deskriptiv	deiktisch
Erinnern	Erkennen
flexibel	beschränkt
mehrdeutig	eindeutig
langsam	zügig
ungewohnt	vertraut

Tabelle 2: Zuordnung ergonomischer Kriterien zu Eingabearten

Dazu ist Folgendes anzumerken:

- Bei der Zuordnung des Attributes “einfach” zur Spracheingabe darf der Lernaufwand beim Umgang mit dem Mikrofon und der Sprechweise bei der Spracheingabe nicht unterschätzt werden, was in [5] bereits festgestellt wurde. Ebenfalls ist zu erwarten, dass die Mausbedienung Ungeübte vor Schwierigkeiten stellen wird.
- Bisher kaum beachtet wurde die mögliche Zurückhaltung oder Scham eines Benutzers mit einer Maschine zu sprechen. Dies kann man z.B. bei Anrufbeantwortern beobachten, die bei einem Teil der Bevölkerung immer noch auf Ablehnung stoßen.
- In [6] wird die These aufgestellt, dass kontrollierbare und zügige Interaktionsformen, wenn sie etabliert sind, durchaus als natürlich empfunden werden, obwohl eher eine Anpassung des Menschen an die Interaktionsform vorliegt.

[1] Verwendet wurde ein System mit AMD Athlon XP 2500+ Prozesor

Des Weiteren müssen die hedonistischen Qualitäten beachtet werden, die besonders bei einem Computerspiel relevant sein können. Eine Auswahl dieser Eigenschaften wurde aus [7] übernommen und ist in Tabelle 3 zusammengestellt. Es ist zu erwarten, dass die Testpersonen bei der Evaluierung des Gesamtsystems sowohl ergonomische als auch hedonistische Eigenschaften berücksichtigen.

positiv	negativ
interessant	langweilig
kostbar	billig
spannend	fad
exklusiv	durchschnittlich
eindrucksvoll	unauffällig
originell	gewöhnlich
innovativ	konservativ

Tabelle 3: Auswahl hedonistischer Eigenschaften

3 Sprachtechnologie

3.1 Spracherkennung

Bei dem verwendeten Spracherkenner handelt es sich um einen Kommandoworterkenner. Vor dem Einsatz dieses Erkenners muss das Vokabular eingegeben werden. Beim Erkennungsvorgang wird derjenige Vokabulareintrag der eingegebenen Äußerung zugewiesen, welcher dieser am ähnlichsten ist. Anhand von Konfidenzmaßen wird eine Aussage über die Qualität dieser Erkennung gefällt. Eine Übersicht über den verwendeten Erkenner ist in Abbildung 1 dargestellt. Im Rahmen der Diplomarbeit wurden die Eigenschaften des Erkenners bezüglich Geschwindigkeit und Nutzerfreundlichkeit verbessert.

3.2 Sprachsynthese

Die Ausgabe von Text auf dem Bildschirm wurde quasi gleichzeitig mit Hilfe des Dresdner Sprachsynthesesystems “Dress” akustisch dargeboten. Da neben der bereits vorhandenen männlichen deutschen Stimme eine weitere benötigt wurde, wurde das System so erweitert, dass die Inventare des MBROLA-Projektes[8], speziell die Stimme “de6”, verwendet werden konnten. Um die Gleichzeitigkeit von Anzeige und akustischer Darbietung des Textes zu ermöglichen, wurde jeder Teilsatz einzeln synthetisiert und die Sprachausgabe nach Fertigstellung des ersten Teilsatzes gestartet. Durch Ausnutzung der so genannten “full-duplex”-Funktionalität war es möglich, die Sprachausgabe bei Äußerung eines Kommandos abzubrechen (auch als “barge-in” bezeichnet).

4 Untersuchungsobjekt

Nicht jede Applikation eignet sich automatisch für eine Sprachbedienung. Die Auswahl dieses Computerspiels hängt vor allem damit zusammen, dass eine Sprachbedinung im Vergleich zu einer äquivalenten Bedienung mit der Computermaus langsamer ist. Des Weiteren ist die Sprachbedienung fehleranfälliger, was die Geschwindigkeit bei der Bedienung u.U. weiter herabsetzt. Eine Anforderung an die Applikation ist also, möglichst invariant gegenüber der Geschwindigkeit der Bedienung zu sein. Damit scheiden alle Computerspiele aus, bei denen die Reaktionszeit des Benutzers eine Rolle spielt.

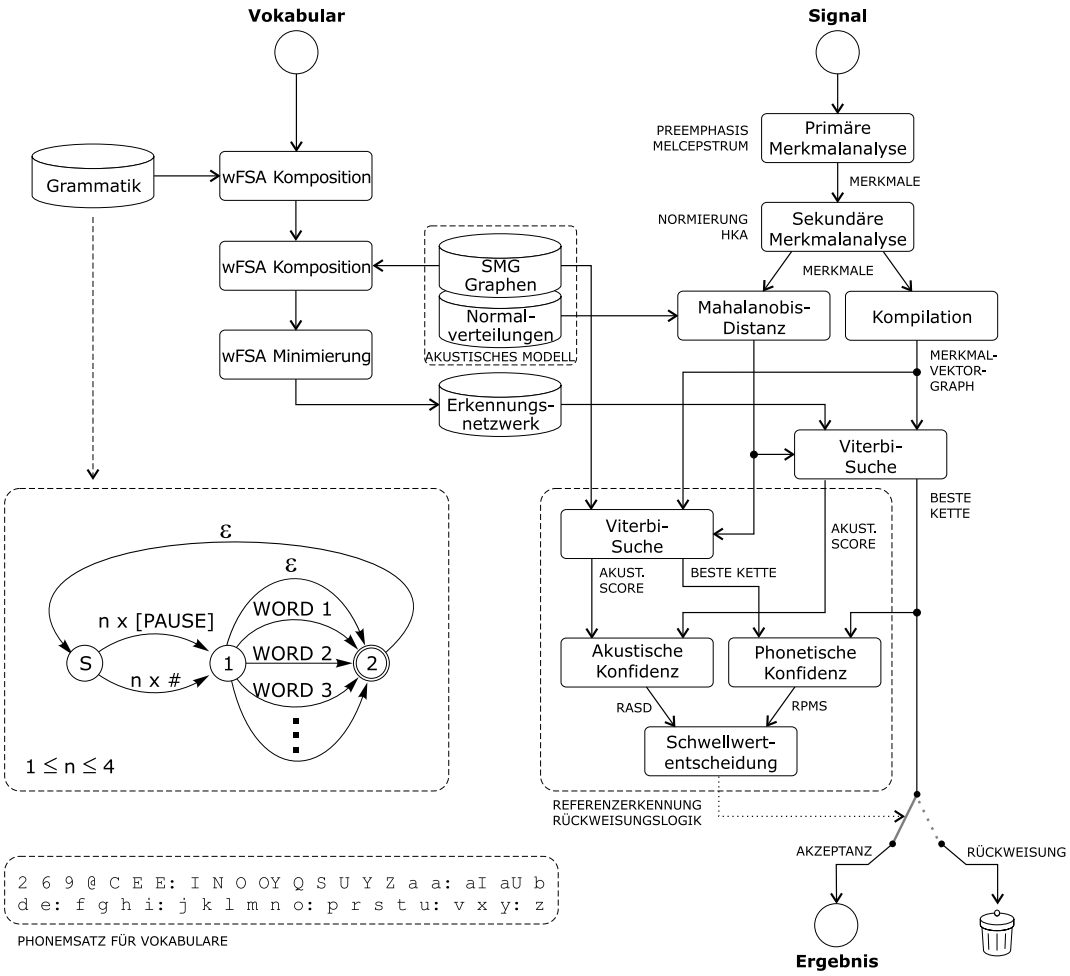

Abbildung 1: Aufbau des verwendeten Erkenners

Ein Abenteuerspiel ist jedoch weitgehend unabhängig von der Geschwindigkeit der Bedienung. Außerdem begünstigen die folgenden Eigenschaften den Einsatz von Spracherkennung in diesem Computerspiel:

- Falsch ausgeführte Aktionen führen nicht in ausweglose Situationen oder zum Abbruch des Spiels. Somit kann eine Fehlerkennung bei der Spracheingabe eher toleriert werden.
- Der Umfang verschiedenartiger Aktionen ist klein. Abgesehen von den Möglichkeiten das Spiel selbst zu bedienen (z.B. Laden und Speichern von Spielständen) stehen dem Benutzer nur wenige Aktionen der Spielfigur zur Auswahl. Dieser Umstand lässt vermuten, dass das zu implementierende Vokabular für die Spracherkennung nicht unverhältnismäßig groß sein wird.
- Kritische Aktionen (z.B. Beenden des Spiels) sind mit einer Sicherheitsabfrage versehen, es ist also sehr unwahrscheinlich, dass im Spielverlauf unbeabsichtigte fatale Fehlentscheidungen auftreten können.

Bei der Mausbedienung ist ein komplexes Dialogmodell vorhanden, da mit lediglich 2 Maustasten nicht alle möglichen Aktionen in einer Dialogebene durchgeführt werden können. Bei der Implementation der Sprachbedienung kann dieses Modell jedoch aufgrund der hohen Flexibilität der Sprache stark vereinfacht werden. Im konkreten Fall bedeutete dies eine Reduktion der Dialogebenen von 7 auf 3. Um die fehlende Kontrollierbarkeit der Sprachbedienung zu kompensieren, wird der Zustand des Erkenners sowie die letzte erkannte Äußerung stets im Fenstertitel dargestellt.

5 Vokabularermittlung

Um die Erwartungskomformität der Sprachbedienung so hoch wie möglich zu gestalten, kann das Vokabular nicht analytisch ermittelt werden. Deshalb wurde eine Simulation mit mehreren Personen durchgeführt. Dabei konnte festgestellt werden, dass sich das von den Probanden verwendete Vokabular ähnelte. Während die Benennungen für Objekte und Personen stark differierte, war der Satzbau weitestgehend gleich. Die Benutzer formulierten die Kommandos als Imperativsätze in direkter Rede. Somit bot sich die Verwendung syntaktischer Schablonen an.

Ein besonders auffälliges Missverständnis bei der Simulation war die häufige Verwendung des Singulars für Objekte, die in der Mehrzahl vorkamen. Die korrekte Vorgehensweise, das Objekt in der Mehrzahl zu bezeichnen oder es eindeutig zu beschreiben, war den Testpersonen offensichtlich nicht bewusst. Vielmehr waren diese der Überzeugung, dass die Aussage vollständig war, obwohl sie diese nicht sprachlich (sondern durch Gesten oder Blicke) vollständig geäußert hatten. Des Weiteren wurde die Richtungssteuerung der Spielfigur aus verschiedenen Perspektiven aus vorgenommen. Hier musste eine Perspektive festgelegt werden, da sich die einzelnen Varianten untereinander ausschließen.

Es wurde versucht, viele unterschiedliche Bezeichnungen für Objekte in die Schablonen aufzunehmen, da beobachtet werden konnte, dass sich die Bezeichnungen von Person zu Person stark unterschieden. Um so schwieriger das Objekt auf dem Bildschirm zu erkennen war, desto mehr (und abwegigere) Begriffe wurden genannt. Das fertige Vokabular enthielt je nach Komplexität der Spielszene zwischen 700 und 2400 Äußerungen.

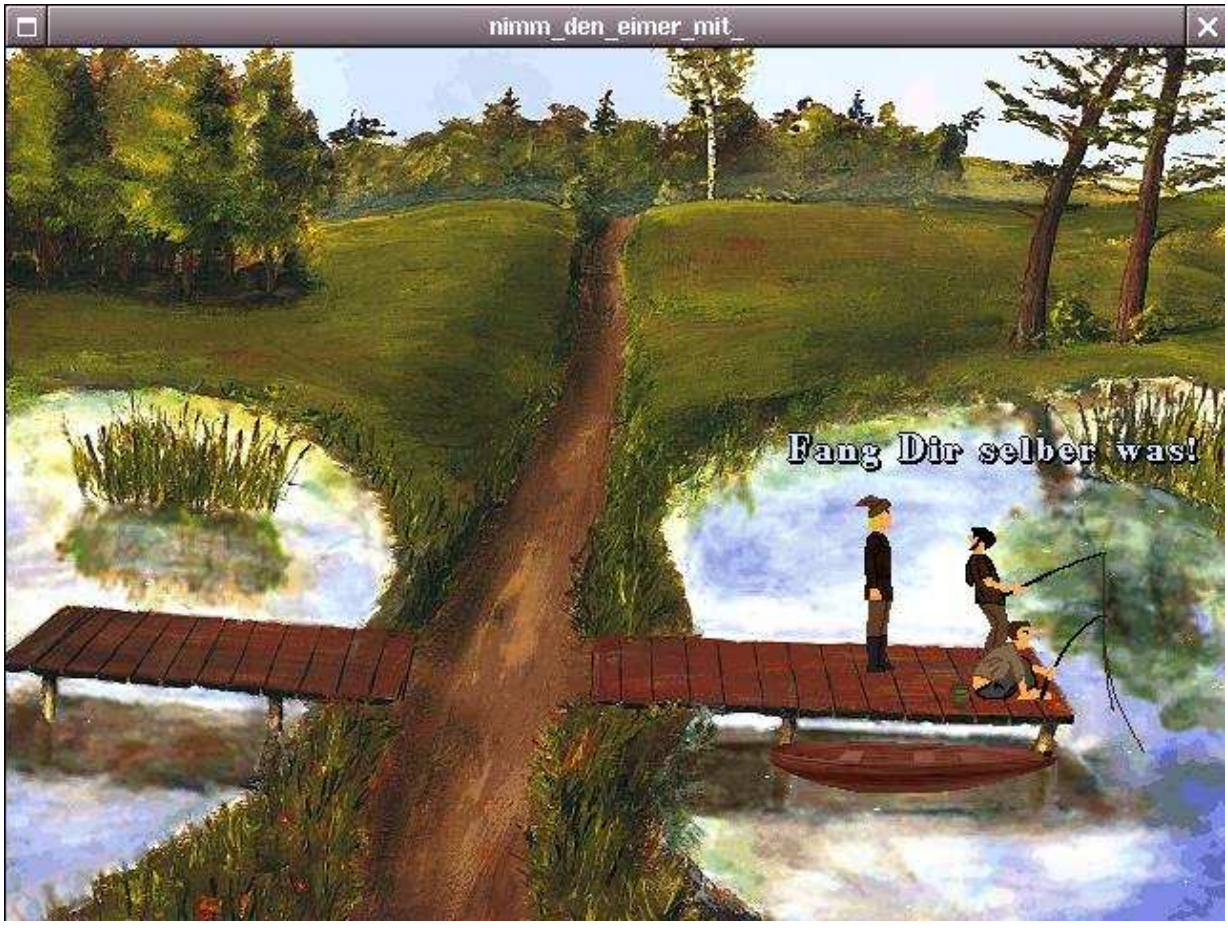

Abbildung 2: Beispielszene aus dem Computerspiel

6 Evaluation

Die Evaluation wurde mit 12 Personen durchgeführt, die unterschiedliche Vorkenntnisse im Umgang mit Computern, Sprachtechnologie und dem Genre des Computerspiels hatten. Während der Evaluation konnte jederzeit zwischen der herkömmlichen und der neuen Bedienung gewählt werden. Die Tester waren angewiesen, Spracherkennung und -synthese ausgiebig im Vergleich zu den herkömmlichen Varianten zu testen und ihre persönlichen Präferenzen zu ermitteln. In einem Auswertebogen sollten des Weiteren die Gründe für die jeweiligen Entscheidungen angegeben werden.

Für die Sprachbedienung an Stelle der Mausbedienung entschieden sich 7 von 12 Probanden. Diese Entscheidung war hauptsächlich von hedonistischen Eigenschaften der Sprachbedienung, wie größere Einbezogenheit oder Bequemlichkeit, geprägt. Diejenigen Testpersonen, die den Umgang mit der Maus noch nicht gewohnt waren, empfanden die Sprachbedienung sogar als leichter. Die Personen, die sich für die Mausbedienung entschieden haben, haben deren ergonomische Vorzüge hervorgehoben (Schnelligkeit, Gewöhnung) und die ergonomischen Schwächen der Sprachbedienung bemängelt (Fehlerkennungen, unzureichender Umfang des Vokabulars).

Bei der Sprachsynthese als zusätzliche Ausgabe zur Darstellung des Textes haben sich 10 von 12 Probanden dafür entschieden, wobei sowohl hedonistische als auch ergonomische Eigenschaften diese Entscheidung beeinflusst haben. Kritisiert wurden vor allem die prosodischen Eigenschaften der Sprachsynthese.

7 Schlussfolgerungen

Anhand der erhaltenen Präferenzen der Testpersonen lassen sich einige Aussagen treffen. Bemerkenswert ist die Tatsache, dass die Erkennungsrate bei der Spracherkennung nicht als das einzigste Kriterium bei der Entscheidung für oder gegen die Sprachbedienung angegeben wurde. Bei Personen, die bei der Evaluation oft mit Fehlerkennungen umgehen mussten, war dies zwar der Hauptgrund für die Ablehnung. Wenn die Erkennungsrate jedoch ein für den jeweiligen Tester akzeptables Minimum überschritten hat, werden andere Kriterien zu der Entscheidung für oder gegen Spracherkennung hinzugezogen.

Unabhängig von der Akzeptanz der Sprachbedienung wurde die Unzulänglichkeit des implementierten Vokabulars kritisiert. Daraus kann man einerseits schlussfolgern, dass es sehr schwierig ist, ein Vokabular zu entwerfen, welches den Erwartungen einer großen Personengruppe entspricht. Diese Anforderung konnte trotz vorhergehender Simulation nicht erfüllt werden. Außerdem wird dadurch deutlich, wie wichtig die Anpassung der Applikation an die neue Form der Bedienung ist. Die Flexibilität der Sprachbedienung sollte dabei so gut wie möglich ausgeschöpft werden, um deren ergonomischen Schwächen zu kompensieren.

Die durch die Sprachsynthese zusätzlich angebotene akustische Ausgabe der auf dem Bildschirm erscheinenden Texte wurde überwiegend akzeptiert. Die Mehrzahl der Testpersonen hat sich offensichtlich an die im Vergleich zu aufgezeichneter Sprache geringere Qualität der Sprachausgabe gewöhnt. Deshalb kann von einem Mehrwert dieser Funktion ausgegangen werden.

Man kann also feststellen, dass im vorliegenden Einsatzfall Sprachtechnologie durchaus einsetzbar ist. Diese Aussage kann zwar nicht problemlos auf andere Einsatzszenarien übertragen werden. Trotzdem kann man feststellen, dass die Verwendung von Sprachtechnologie in Applikationen sowohl von der Qualität der Sprachtechnologie als auch von der Anpassung der Applikation an diese neue Technologie abhängt. Aufgrund der verschiedenen ergonomischen und hedonistischen Qualitäten von Sprach- und Mausbedienung

ist nicht zu erwarten, dass sich eine gegenüber der anderen durchsetzen wird.

Literatur

[1] Sobe, D.: “Sprachsteuerung eines Computerspiels - Untersuchungen zur Leistungsfähigkeit und Ergonomie”, Technische Universität Dresden, Fakultät Elektrotechnik und Informationstechnik, Institut für Akustik und Sprachkommunikation, Diplomarbeit, 2004

[2] Hoffmann, R.: A multilingual text-to-speech system, The Phonetician 80 (1999/II), pp. 5 - 10, 1999

[3] Eichner, M.; Kühne, M.; Werner, S.; Wolff, M.: Sprachtechnologien in der Lernumgebung eines internet-basierten Studienganges, Tagungsband zur 14. Konferenz “Elektronische Sprachsignalverarbeitung” vom 24. bis 26.9.2003 in Karlsruhe, S.370 - 377

[4] Herczeg, M.: Software - Ergonomie Grundlagen der Mensch-Computer-Interaktion, Bonn; Paris; Reading, Massachusattes [u.a.]: Addison-Wesley, 1994

[5] Helbig, J.; Schindler, B.: Spracheingabe bei der technischen Inspektion von Kraftfahrzeugen, Tagungsband zur 13. Konferenz “Elektronische Sprachsignalverarbeitung” vom 25. bis 27.9.2002 in Dresden, S.88 - 95

[6] Klemmert, H.; Stock, C.; Marzi, R.: Erhöht gesprochene Interaktion die Nutzungsfreundlichkeit von Software? Ergebnisse einer empirischen Studie, Tagungsband zur 13. Konferenz “Elektronische Sprachsignalverarbeitung” vom 25. bis 27.9.2002 in Dresden, S.392 - 398

[7] Hassenzahl, M.; Platz, A.; Burmester, M.; Lehner, K.: Hedonic *and* Ergonomic Quality Aspects Determine a Software’s Appeal, Proceedings of the CHI 2000, S. 201 - 208

[8] The MBROLA Project
http://tcts.fpms.ac.be/synthesis/mbrola/
Stand: 18.06.2004