STUDIE ZUR ANWENDBARKEIT SCHNELLER SPRACHSYNTHESE IN SPRACHDIALOGSYSTEMEN

Klaus-Peter Engelbrecht, Arne Denneler, Cliford Yangmia, Benjamin Weiss

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin
klaus-peter.engelbrecht@telekom.de

Abstract: Der Beitrag analysiert die Auswirkung von schneller Sprachsynthese auf die Usability von Sprachdialogsystemen. In einem Experiment mit 13 Testnutzern wurde dazu die Silbenrate eines Systems zur Fahrplanauskunft zwischen 3 und 10,5 Silben pro Sekunde (s/s) variiert. Das System wurde bei Sprechgeschwindigkeiten zwischen 5 und 7 s/s am besten bewertet wird. Bei diesen Silbenraten stellt sich auch ein optimales Verhältnis von Effizienz und kognitivem Aufwand ein.

1 Einführung

Sprachdialogsysteme können heute recht zuverlässig für Routine-Aufgaben, wie z.B. zur Fahrplanauskunft, eingesetzt werden. Allerdings sind diese Systeme im Vergleich zu menschlichen Gesprächspartnern deutlich ineffizienter. In einem empirischen Vergleich zwischen menschlichen und maschinellen Agenten identifizierten Ward et al. [1] mehrere Aspekte, hinsichtlich derer maschinelle Gesprächspartner den menschlichen unterlegen sind. Nach der Genauigkeit der Spracherkennung und der Flüssigkeit von Sprecherwechseln wurde die Sprechgeschwindigkeit der Systemausgabe genannt. Während zur Verbesserung der ersten beiden Aspekte momentan viel Forschung betrieben wird, wurde bisher nicht untersucht, ob die Effizienz von Systemausgaben sich durch eine Steigerung der Sprechgeschwindigkeit verbessern ließe. Daher nimmt sich dieser Beitrag dieser Fragestellung an.

In vorangegangenen Arbeiten wurde bereits festgestellt, dass Sprache bei deutlich höheren Geschwindigkeiten verstanden werden kann als sie in natürlicher Sprache auftreten [2, 3]. Während der Sprechgeschwindigkeit beim Menschen durch die Beschaffenheit der Artikulationsorgane Grenzen gesetzt sind, lässt sich mit Sprachgenerierungssystemen (Text-to-Speech; TTS) fast beliebig schnelle Sprache erzeugen. Blinde Menschen nutzen laut Moos und Trouvain [2] häufig Sprachsynthese, um den Inhalt von Texten zu erfassen, und sind aufgrund des damit verbundenen Trainingseffektes zu außergewöhnlichen Leistungen fähig. Sie können synthetisierte Sprache mit einer Geschwindigkeit von über 20 Silben pro Sekunde (s/s) noch gut verstehen [3]. Nach Trouvain [4] ist die Verständlichkeit von Formant-Synthese bei hohen Silbenraten (>7,5 s/s) besser als die von Diphon-Synthese, jedoch können andere Aspekte der Umsetzung als die Synthesetechnik einen starken Einfluss auf die Verständlichkeit haben [5]. Auch wurde bereits gezeigt, dass schnelle synthetisierte Sprache auch von nicht blinden Menschen gut verstanden werden kann [3], wobei die maximale Silbenrate hier deutlich niedriger liegt als bei geübten blinden Nutzern.

Während in den genannten Vorarbeiten die Verständlichkeit bei einseitiger Präsentation von Texten untersucht wurde, muss davon ausgegangen werden, dass im Dialog andere Anforderungen an Sprache und Sprachverstehen bestehen. Beispielsweise müssen in dialogischer Sprache bestimmte Informationen besonders genau verstanden werden (z.B. der Wortlaut von Kommandos, wenn diese durch das System vorgegeben werden). Zudem muss in Dialogen auch die pragmatische Bedeutung des Gesagten verstanden und in eine geeignete Antwort umgemünzt werden. Nebenbei muss aus linguistischen und para-linguistischen Eigenschaften des Gesagten erkennbar sein, wann ein Beitrag des Nutzers erwartet wird.

Die Steigerung der Sprechgeschwindigkeit führt also nicht unbedingt zu einer Steigerung der Effizienz des Dialogsystems, da möglicherweise mehr Nutzerfehler auftreten, wenn die Verständlichkeit der Prompts Einbußen erfährt, und da insgesamt weniger Zeit zur Verarbeitung der Systemausgabe zur Verfügung steht. Daher wurden in dieser Studie neben dem Einfluss des Sprechtempos auf die gesamte Dialogdauer auch Effekte auf Fehlerraten untersucht. Weiterhin sind neben der Effizienz auch andere für die Usability des Systems relevante Qualitätsaspekte von Interesse, um ein Gesamtbild von der Qualität eines Sprachdialogsystems mit schneller Sprachsynthese zu erhalten.

Neben der Erhöhung der Sprechgeschwindigkeit wäre eine Verbesserung der Effizienz von Systemausgaben prinzipiell auch durch eine Verkürzung des gesprochenen Textes zu erreichen. Ward et al. [1] merken jedoch an, dass Dialogbeiträge eines automatischen Dialogsystems in der Regel ausführlicher sind als Dialogbeiträge von Menschen, da zum einen durch die Systemausgabe übermittelt werden soll, welche Art von Antworten das System verstehen kann, und zum anderen der gleiche Prompt für unterschiedliche Nutzer tauglich sein soll, während sich Menschen flexibler auf den Interaktionspartner einstellen können.

2 Experiment

Um den Einfluss des Sprechtempos des Systems auf dessen Usability zu untersuchen, wurden in einem Experiment Versuchspersonen gebeten, Interaktionen mit einem Beispielsystem durchzuführen und zu bewerten. Dabei wurde die Silbenrate der Systemausgabe systematisch variiert.

2.1 Ermittlung der Silbenraten für das Experiment

Die unterschiedlichen Sprechtempi des Systems sollten für das Experiment so gewählt werden, dass die Ermittlung einer optimalen Silbenrate möglich ist, d.h., sie sollten zwischen einem als langsam empfundenen Tempo und dem maximalen noch verständlichen Tempo liegen. Die normale Sprechgeschwindigkeit von Menschen in einer Unterhaltung oder beim Vorlesen liegt zwischen 4 und 8 s/s [2]. Daher wurde als Minimum eine Silbenrate von 3 s/s gewählt. Die maximale Silbenrate, bei der die Verständlichkeit von synthetisierter Sprache bereits leichte Einbußen erfährt, liegt bei nicht blinden Menschen bei etwa 10,5 s/s [4]. In einem informellen Vorversuch bestätigte sich, dass bei dieser Silbenrate mit der verwendeten Sprachsynthese bereits leichte Einbußen bei der Verständlichkeit auftreten. Die Silbenrate wurde im Experiment also zwischen 3 und 10,5 s/s variiert.

Um die Ergebnisse bewerten zu können, wurden zudem typische Silbenraten für Menschen beim Dialog mit Sprachdialogsystemen, sowie die Silbenrate des Systems selbst ermittelt. Nach Dioubina [6] sprechen Menschen im Dialog mit Dialogsystemen im Mittel mit einer Silbenrate von etwa 5-6 s/s. Zur Ermittlung der typischen Sprechgeschwindigkeit eines Sprachdialogsystems wurde ein bestehendes System der Deutschen Bahn auf die Geschwindigkeit der dort verwendeten Sprachsynthese hin untersucht. Das System ist als automatisches Fahrplanauskunftssystem unter der Telefonnummer 0800 150 70 90 zu erreichen. Hierzu wurden drei Dialoge mit dem System aufgezeichnet und die Sprechraten dreier unterschiedlicher Sprachausgaben des Systems ermittelt. Anschließend wurde der Mittelwert aller gemessenen Sprechraten des Systems errechnet. Im Ergebnis liegt die Rate des Auskunftssystems bei durchschnittlich 3,5 Silben pro Sekunde. Auffällig ist, dass diese Sprechrate bei zwischenmenschlichen Dialogen bereits als langsam empfunden wird. Sie liegt jedoch innerhalb der zuvor beschriebenen Grenzen, die für das Experiment festgelegt wurden und bestätigt diese somit.

2.2 Design

Das Intervall zwischen der minimalen und maximalen Silbenrate wurde in 5 Stufen unterteilt, die Silbenrate betrug also in jedem Durchlauf entweder 3, 5, 7, 9 oder 10,5 s/s. Für die Durchführung des Experiments wurde ein Design mit Messwiederholung gewählt, d.h., jede Versuchsperson führte einen Dialog mit jeder der fünf Silbenraten durch. In jedem Durchgang erledigte die Versuchsperson eine andere Aufgabe, wobei darauf geachtet wurde, dass alle Aufgaben etwa gleich komplex waren. Um Reihenfolgeeffekte möglichst über den gesamten Datenkörper zu streuen, wurden die fünf Faktorstufen und die Zuteilung der fünf Interaktionsaufgaben nach dem Schema eines griechisch-lateinischen Quadrates angeordnet.

Nach jedem Durchlauf bewertete die Versuchsperson das Systems auf einer gekürzten Version des SASSI-Fragebogens [7]. Um der Ermüdung der Versuchspersonen vorzubeugen, wurden aus jeder Skala des Fragebogens zwei möglichst repräsentative Items ausgewählt. Dabei handelte es sich in der Regel um die beiden Items mit der höchsten Ladung auf den jeweiligen Faktor. In zwei Fällen wurde das am höchsten ladende Item jedoch nicht genutzt, da die Formulierung weniger klar war als beim am dritthöchsten ladenden Item. Das SASSI-Item „Die Interaktion mit dem System ist effizient“ wurde wegen seiner besonderen Bedeutung für die Fragestellung zusätzlich erhoben. Außerdem beantworteten die Versuchspersonen zwei Fragen zum Inhalt der Auskunft des Systems. Dadurch sollte sichergestellt werden, dass der Dialog tatsächlich verstanden wurde.

2.3 System und Aufgaben

Für alle Aufgaben wurde dasselbe Sprachdialogsystem verwendet. Dabei handelte es sich um ein Fahrplanauskunftssystem nach dem Vorbild des o.g. Systems der Deutschen Bahn. Um den Implementierungsaufwand zu minimieren, wurde das Experiment nach dem „Wizard-of-Oz“-Paradigma durchgeführt, d.h., einer der Experimentatoren ersetzte die Spracherkennungs- und Sprachverstehens-Komponente, indem er die vom Nutzer genannten semantischen Konzepte direkt in den Dialogmanager eingab. Dementsprechend traten in den Interaktionen keine Spracherkennungsfehler auf, jedoch wurde ein „no-match“-Event ausgelöst, wenn die Nutzereingabe nicht mit den im aktuellen Dialogzustand zur Verfügung stehenden Konzepte beschrieben werden konnte. Für die Sprachausgabe wurde die Mary-TTS [8] verwendet, wobei wegen der (nach dem Urteil der Autoren) vergleichsweise schlechten Qualität der Formant-Synthesen von Mary eine Diphon-Synthese gewählt wurde.

Der Dialog begann jeweils mit der Abfrage der Eckdaten für die gesuchte Verbindung. Im Anschluss stellte das System verschiedene Varianten der Verbindung bereit und erlaubte dem Nutzer, eine Verbindung auszuwählen, Details dazu zu erfahren und eine Fahrkarte zu bestellen. Für die Auswahl einer Verbindung wurden in den Aufgabenbeschreibungen unterschiedliche Kriterien vorgegeben (z. B. kürzeste Verbindung oder Verbindung mit den wenigsten Umstiegen).

Bei den Aufgaben wurde darauf geachtet, dass die Nutzereingaben möglichst lang sind, so dass Silbenraten ermittelt werden können. Dazu wurden für Start- und Zielbahnhöfe möglichst lange Namen gewählt. Zudem war es Teil der Aufgabenstellung, nach Abschluss des Dialogs jeweils 2 Fragen zur Auskunft des Systems (z. B. Dauer der Reise insgesamt) zu beantworten. Um zu vermeiden, dass die Nutzer einen unnatürlichen Aufwand damit haben, Informationen in der Aufgabenbeschreibung zu suchen, wurde diese Information auf dem Aufgabenblatt in Stichpunkten zusammengefasst, wie in folgendem Beispiel:

> *Bitte suchen Sie eine Verbindung von Eisenhüttenstadt nach Brandenburg an der Havel mit Abfahrt morgen Vormittag. Wählen Sie bei der Verbindungsauswahl die Option mit der kürzesten Reisedauer und bestellen Sie anschließend zwei Fahrkarten.*

Start- und Zielbahnhof: Von Eisenhüttenstadt nach Brandenburg
Zeit: Abfahrt morgen Vormittag.
Verbindungsoption: kürzere Reisedauer.
Anzahl Fahrkarten: 2

2.4 Probanden

Dreizehn Probanden wurden aus dem Bekanntenkreis der Experimentatoren rekrutiert. Unter den Teilnehmern waren 5 Frauen und 8 Männer. Das Durchschnittsalter betrug 24,5 Jahre. Neun Teilnehmer waren Studenten, drei waren berufstätig und eine Versuchsperson befand sich in einer Berufsausbildung. Unter den Teilnehmern waren fünf Personen, die bereits Erfahrung mit einem Dialogsystem oder einem sprachgesteuerten persönlichen Assistenten hatten. Vier dieser fünf Personen nutzten derartige Systeme jedoch nur selten, eine Person zumindest wöchentlich. Alle Versuchspersonen verfügten nach eigenem Kenntnisstand über eine normal ausgeprägte Hörfähigkeit. Einer der 13 Probanden war kein deutscher Muttersprachler.

3 Daten

3.1 Performanz

Die Effizienz des Systems wurde über die Ausführungszeit (*DD* für *Dialog duration*) operationalisiert. Um den Einfluss der Silbenrate auf den Dialogverlauf quantifizieren zu können, wurden weitere Interaktions-Parameter gemessen:

- *#Turns*: Anzahl Dialogschritte beider Interaktionspartner
- *#Repetition-requests*: Anzahl von Kommandos, die letzte Systemausgabe zu wiederholen
- *#No-matches*: Anzahl von Nutzeräußerungen, die im aktuellen Zustand des Systems nicht verarbeitet werden können
- *#Abortions*: Anzahl Aufgabenabbrüche durch den Nutzer
- *#Barge-ins*: Anzahl der Unterbrechungen der Systemausgabe durch den Nutzer

Außerdem wurde die Silbenrate der Nutzeräußerungen geschätzt, indem für jeden Dialog die Rate für die erste Nutzeräußerung im Dialog ermittelt wurde. Ein Adaptionseffekt konnte bereits zum Zeitpunkt der ersten Nutzeräußerung erwartet werden, da dieser ein relativ langer Willkommens-Prompt durch das System voran ging.

3.2 Fragebogen

Der SASSI-Fragebogen umfasst die Skalen *System Response Accuracy*, *Likeability*, *Cognitive Demand*, *Annoyance*, *Habitability* und *Speed*. Für jede Skala wurden 2 repräsentative Fragen ausgewählt. Bei der Datenauswertung zeigten sich jedoch niedrige Korrelationen zwischen den zu einer Skala gehörenden Fragebogen-Items (r=[0,04; 0,69]).

Daher wurde, anstatt die Mittelwerte der beiden Items jeder Skala zu bilden, zunächst eine Faktorenanalyse mit Oblimin-Rotation zur Identifikation der dem Fragebogen unterliegenden Bewertungsdimensionen durchgeführt. Faktoren mit Eigenwerten größer 1 wurden weiter ausgewertet. Die Oblimin-Methode wurde für die Rotation gewählt, da von Korrelation der Bewertungs-Dimensionen ausgegangen werden kann. Um plausible und konsistente Faktoren zu erhalten, wurden nach dem ersten Versuch drei Items von der Faktorenanalyse ausgeschlossen, die im Vergleich zu den anderen eher die Einstellung der Nutzer zum System

als deren Wahrnehmung des Systems repräsentieren. Diese Items wurden im Folgenden einzeln ausgewertet. Aus den verbleibenden Items wurden 3 Faktoren gebildet:

- *Genauigkeit der Systemantwort*
- *Kognitive Beanspruchung*
- *Tempo*

Die Faktoren sind also inhaltlich mit den SASSI-Skalen vergleichbar, jedoch fehlen die Skalen *Annoyance*, *Likability* und *Habitability*. Der Aspekt *Likability* wird jedoch durch zwei der drei verbleibenden Items abgedeckt. Da die Reliabilität der gebildeten Skalen (Cronbach's Alpha) dürftig ausfiel, wurden die Faktor-Scores für die Auswertung verwendet.

4 Ergebnisse

Die Auswertung des Experiments erfolgte zunächst anhand von Grafiken, die Mittelwerte der gemessenen Parameter für die unterschiedlichen Silbenraten gegenüberstellen. Zusätzlich wurden Unterschiede zwischen den Silbenraten auf Signifikanz geprüft. Dabei wurde soweit möglich das Verfahren für intervallskalierte Daten (ANOVA mit Messwiederholung) verwendet. Fiel der Mauchly-Test auf Sphärizität signifikant aus (d.h., konnte Sphärizität nicht angenommen werden), wurde der p-Wert über das Greenhaus-Geisser-Verfahren berechnet, um eine progressive Entscheidung zu vermeiden.

4.1 Einfluss von Silbenrate auf Performanz

Erwartungsgemäß fällt die Ausführungszeit bei steigender Silbenrate des Systems stark ab, wobei sie zwischen der langsamsten und schnellsten Bedingung etwa halbiert wird (s. Abb. 1). Zwischen den Bedingungen bestehen signifikante Unterschiede ($F(1,537)=68,971$; $p<0,01$). Hingegen liegt kein signifikanter Einfluss auf die Anzahl der benötigten Dialogschritte vor ($F(1,787)=0,624$; $p=0,53$), d.h., auch bei schnellem Sprechtempo des Systems konnten die Nutzer die Aufgaben ohne Umwege erledigen. Dementsprechend wurde kein nennenswerter Einfluss auf *#No-matches*, *#Abortions* oder *#Barge-ins* gefunden (Abb. 2). Die Anzahl der Kommandos zur Wiederholung des letzten Prompts (*#Repetition-requests*) steigt mit steigender Silbenrate leicht an, insbesondere wenn die TTS mehr als neun Silben pro Sekunnde produziert. Da es sich um Häufigkeiten mit sehr niedrigen Werten handelt, wurde zur Überprüfung der Signifikanz der Unterschiede der nicht-parametrische Friedman-Test gewählt. Dieser fiel signifikant aus ($Chi^2(4)=15,227$; $p<0,01$).

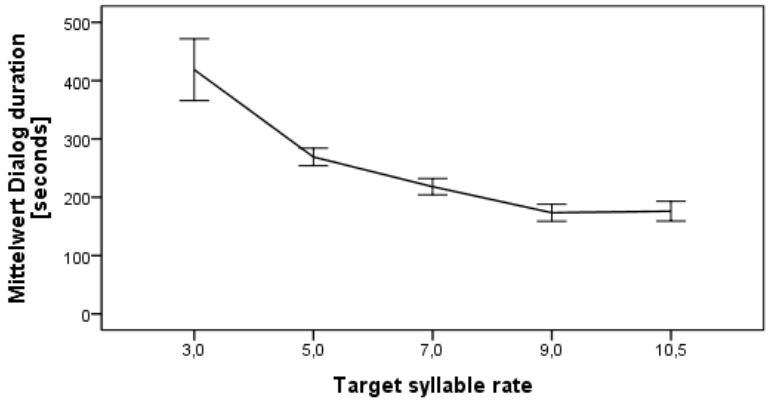

Abbildung 1 - Einfluss der Silbenrate auf die mittlere Ausführungszeit (Fehlerbalken bedeuten 95%-Konfidenzintervalle)

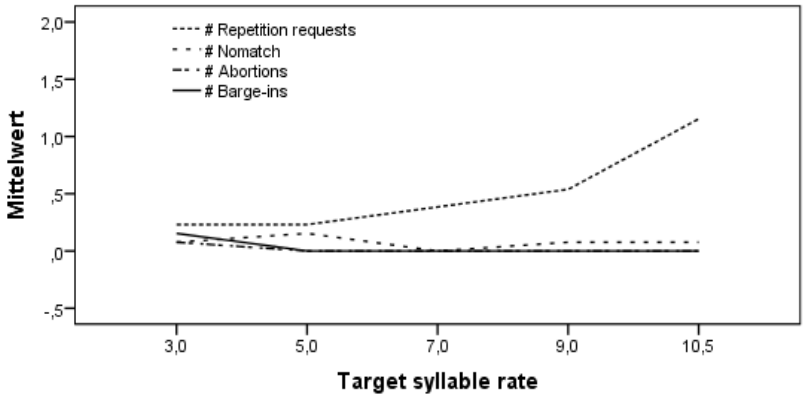

Abbildung 2 - Einfluss der Silbenrate auf verschiedene Interaktionsparameter

4.2 Einfluss von Silbenrate auf Beurteilung

Wie Abb. 3 zu entnehmen ist, steigt die Bewertung des Tempos der Interaktion mit der Silbenrate des Systems ($F(2,394)=36,128$; $p<0,01$). Andererseits steigt jedoch auch die kognitive Beanspruchung, d.h., die Bewertung des Systems wird auf diesem Faktor schlechter ($F(4)=7,157$, $p<0,01$). Obwohl die Performanz-Daten allenfalls auf geringfügige Unterschiede in den Dialogverläufen bei unterschiedlichen Silbenraten hindeuten, wurde die Genauigkeit der Systemantwort bei mittleren Silbenraten tendenziell am besten bewertet. Die Varianzanalyse zeigt hier einen Effekt auf dem 5%-Signifikanzniveau ($F(4)=2,991$, $p=0,028<0.05$), der Effekt ist also gegenüber den anderen Ergebnissen mit einer höheren Irrtumswahrscheinlichkeit behaftet.

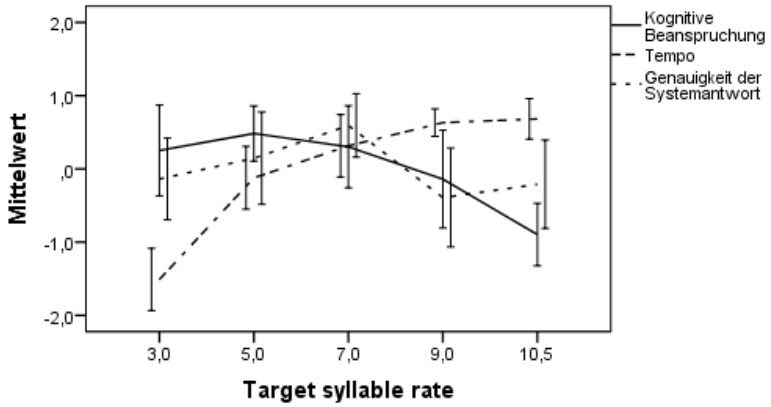

Abbildung 3 - Einfluss der Silbenrate auf die Wahrnehmung der Eigenschaften des Dialogs (Bewertungsdimensionen aus dem Fragebogen; Fehlerbalken bedeuten 95%-Konfidenzintervalle)

Die Einstellung der Nutzer zum System ist erwartungsgemäß bei mittleren Silbenraten am besten. Dies steht auch mit den Performanz-Daten bei den unterschiedlichen Silbenraten in Einklang Das Item „Das System ist nützlich“ zeigte die höchste mittlere Bewertung für Silbenraten zwischen 5 und 7 s/s ($F(2,376)=6,711$, $p<0,01$). Das gleiche gilt für das Item „Das System ist angenehm“ ($F(4)=14,869$; $p<0,01$). Interessanterweise wird auch die Effizienz des Systems für mittlere Silbenraten am besten bewertet. Auffällig ist hier außerdem, dass die Bewertung zur langsamsten Silbenrate von 3 s/s besonders stark abfällt, während zwischen 5 und 7 s/s kein Unterschied in der wahrgenommenen Effizienz beobachtet wurde.

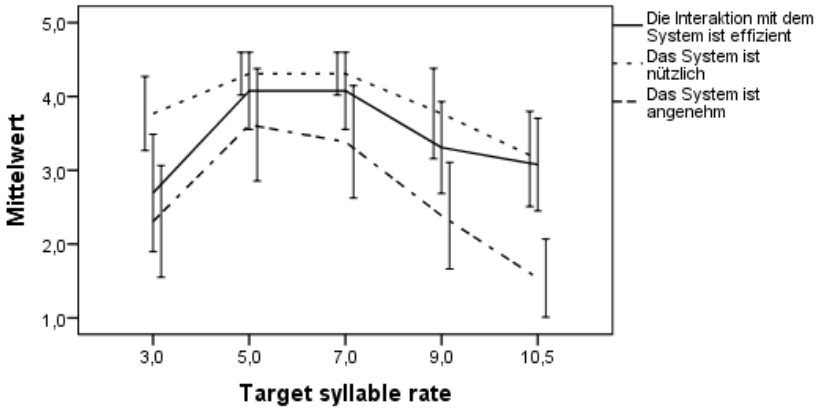

Abbildung 4 - Einfluss der Silbenrate auf die Bewertung des Systems

4.3 Einfluss von Silbenrate auf Timing

Die Silbenrate der Nutzer steigt bis zu einem Sprechtempo des Systems von 7 s/s leicht an und bleibt dann konstant ($F(4)=5{,}153$, $p<0{,}01$. Es ist also ein Adaptionseffekt zu beobachten. Im Mittel liegen die Silbenraten der Nutzer im Einklang mit [6] bei 5 s/s.

Die geplante Analyse der Nutzerbewertungen der Reaktionszeiten des Systems konnte nicht sinnvoll durchgeführt werden, da das entsprechende Fragebogen-Item („Das System reagiert zu langsam“) anscheinend anders interpretiert wurde als gedacht: Die Aussage erhielt mit steigender Sprechgeschwindigkeit des Systems zunehmend weniger Zustimmung (erwartet wurde, dass Reaktionszeiten bei schnelleren Sprechtempi als länger wahrgenommen werden).

5 Diskussion

Zusammenfassend lässt sich sagen, dass Silbenraten zwischen 5 und 7 s/s am besten bewertet wurden. Ein Optimum zwischen minimaler und maximal verständlicher Silbenrate war erwartet worden, da sich hier ein Kompromiss zwischen der Länge des Dialogs und der kognitiven Beanspruchung der Nutzer einstellt.

Eine Silbenrate von 7 s/s entspricht einer Verdopplung der Silbenrate des Auskunftssystems der Deutschen Bahn, das für diese Studie als Referenz diente. Die Ausführungszeit wurde für das hier verwendete System mit dieser Silbenrate in etwa halbiert. Bei einer weiteren Erhöhung des Sprechtempos ist der Abfall der Ausführungszeit weniger steil, d.h., eine weitere Erhöhung der Silbenrate ist auch im Hinblick auf die Performanz des Gesamtsystems kaum noch sinnvoll. Auf den Dialogverlauf im Sinne der hier gemessenen Interaktionsparameter hatte eine Erhöhung der Silbenrate auf 7 s/s keinen nennenswerten Einfluss: lediglich die Anzahl der Kommandos zur Wiederholung des vorangehenden System-Prompts zeigt bei dieser Silbenrate bereits einen leichten Anstieg.

Dies spiegelte sich auch in der Wahrnehmung der Interaktion durch die Nutzer wider. Während das wahrgenommene Tempo der Interaktion bis 7 s/s stark ansteigt, wurden für die kognitive Beanspruchung und die Genauigkeit der Systemantwort bei dieser Silbenrate noch keine Einbußen gegenüber den langsameren Sprechtempi beobachtet. Bei höheren Geschwindigkeiten verbessert sich die Wahrnehmung des Tempos der Interaktion nur geringfügig, was jedoch auf Kosten einer schnellen Erhöhung des kognitiven Aufwands geht.

Dass der Effekt auf die Beurteilung der Effizienz des Systems hier eher der Einstellung der Nutzer zum System als der Tempo-Wahrnehmung folgt, könnte daran liegen, dass die Nutzer auch die kognitive Belastung als Kriterium für Effizienz sehen. Möglicherweise handelt es sich aber hierbei um einen Halo-Effekt, d.h., die Gesamtbeurteilung färbte auf die Effizienz-

Beurteilung ab. Ein Halo-Effekt könnte auch die schlechtere Bewertung für die Genauigkeit der Systemantwort bei hohen Silbenraten trotz gleicher Fehleranzahl erklären.

Ein möglicher Bias der Ergebnisse ergibt sich aus der Reihenfolge, in der die Systeme mit unterschiedlichen Sprechtempi den Probanden präsentiert wurden. Probanden, die zunächst die Bedingung mit langsameren Sprechtempi erfahren hatten, kannten Teile der Prompts bereits und waren somit im Vorteil gegenüber Nutzern, die direkt mit einem System mit hohem Sprechtempo interagieren müssen. Eine Alternative wäre gewesen, dass die Tempi 10,5 s/s und evtl. auch 9 s/s immer zuerst getestet werden. Hier musste beim Design des Experiments ein Kompromiss gemacht werden zwischen Lerneffekten und der Verteilung von Reihenfolgeeffekten über alle Bedingungen. Da die Lerneffekte leichter abschätzbar sind, wurde ein diesbezüglicher Fehler bevorzugt in Kauf genommen.

Hinsichtlich des Gültigkeitsbereichs der Studie ergeben sich einige Einschränkungen. Zunächst wurde das Experiment als Labortest durchgeführt, die Probanden befanden sich also in einer ruhigen Umgebung, in der keine Störungen des Dialogs auftraten. Durch die vorgegebene Aufgabenstellung hatten sie zudem Informationen über die Struktur des Interaktionsablaufs, die normale Nutzer des Systems möglicherweise aus den Prompts ableiten müssten. Beides wirkt sich tendenziell zum Vorteil höherer Silbenraten aus.

Ferner war das verwendete System nicht in allen Punkten realistisch. Da in dem Experiment keine Spracherkennungsfehler auftraten (bzw. simuliert wurden), diese jedoch die kognitive Beanspruchung der Nutzer stark erhöhen können, ist eine entsprechende Folge-Studie angezeigt, um diese Ergebnisse zu bestätigen. Dabei sollte auch berücksichtigt werden, dass durch die Telefonübertragungsqualität Einbußen in der Verständlichkeit möglich sind, und dass diese sich bei unterschiedlichen Sprechtempi unterschiedlich auswirken können. In dieser Studie wurde hingegen das Audiosignal weitgehend ungestört übertragen.

Schließlich handelt es sich bei den Testnutzern um eine relativ kleine und homogene Gruppe, bestehend aus größtenteils jungen Menschen ohne bekannte Gehörprobleme. In der Praxis werden Sprachdialogsysteme meist für viele unterschiedliche Nutzertypen designt. Das Sprechtempo sollte in diesem Fall auch für die schwächste Gruppe noch verständlich sein. Eine Adaption an den Nutzer wäre hier ggf. sinnvoll.

Literatur

[1] Ward, A. Rivera, K. Ward, D. Novick: Some Usability Issues and Research Priorities in Spoken Dialog Applications. Technical Report UTEP-CS-05-23, 2005.

[2] Moos, A., Trouvain, J.: Einzelfallstudie zu Grenzen der Verständlichkeit ultra-schneller Sprachsynthese. In: Tagungsband der ESSV 2008, Frankfurt am Main, 2008, S. 207–214.

[3] Moos, A., Trouvain, J.: Comprehension of ultra-fast speech – blind vs. "normally hearing" persons. In: Proc. 16th ICPhS, Saarbrücken, 2007, S. 677–680.

[4] Trouvain, J.: On the comprehension of extremely fast synthetic speech, Saarland Working Papers in Linguistics 1, 2007, S. 5-13.

[5] Stent, A., Syrdal, A., Mishra, T.: On the Intelligibility of Fast Synthesized Speech for Individuals with Early-Onset Blindness. In: Proc. ASSETS '11, Dundee, Schottland, 2011, S. 211-218.

[6] Dioubina, O. I.: Prosody of Dialogues: Influence of Recognition Failure on Local Speech Rate. In: Proc. Speech Prosody 2004, Nara, Japan, 2004, S. 275-278.

[7] Hone, K. S., Graham, R.: Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, *6*(3-4), 2000, S. 287-303.

[8] Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology, 6,* 2003, S. 365-377.