Vergleich zwischen automatisiertem Labortest und Live-Probandentest zur Beurteilung der Erkennungssicherheit sprachbedienter Systeme im Kfz

Dr. Gudrun Klasmeyer, Martin Herrenkind, Bobby Haferburg

IAV GmbH, 38518 Gifhorn, Deutschland, Email: gudrun.klasmeyer@iav.de, martin.herrenkind@iav.de, bobby.haferburg@iav.de

Einleitung

Die Integration von sprachbedienten Systemen in Kraftfahrzeuge ist ein vielschichtiger Prozess, der in der Regel auch eine entwicklungsbegleitende Erprobung umfasst. Als Gütemaß für den Spracherkenner wird dabei eine Erkennungsrate in Prozent ermittelt.

Üblicherweise finden Spracherkennertests mit verschiedenen Sprechern bei verschiedenen Fahrbedingungen im Kraftfahrzeug statt. Solche *Live*-Probandentests sind nur schwer reproduzierbar. Zudem ist die Wiederholung eines Tests, z.B. nach einer Modifikation des Systems in der Entwicklungsphase, nur mit dem gleichen Zeit- und Personalaufwand, sowie den damit verbundenen Kosten realisierbar. Eine Automatisierung unter reproduzierbaren Laborbedingungen bringt hier Vorteile. Durch Trennung von in reflektionsarmer Umgebung aufgenommenen Sprachkommandos und im Fahrzeug aufgezeichneten Fahrgeräuschprofilen lassen sich Sprachsignale wiederholt - auch in unterschiedlichen Fahrzeugkabinen – einsetzen [siehe auch : Herrenkind et al. 2006].

Der automatisierte Testprozess unterscheidet sich jedoch in einer Hinsicht vom *Live*-Probandentest: Wenn ein Proband das System in mehreren Testdurchläufen im Fahrzeug erprobt, ist der Proband selbst einer **Lernphase** unterworfen, in der er – bewusst oder unbewusst – herausfindet, mit welcher Aussprachevariante, Artikulationsgenauigkeit, Sprechgeschwindigkeit, Stimmlage usw. er vom Gerät am besten verstanden wird. Bei der Aufzeichnung von Sprachproben im reflektionsarmen Raum lässt sich diese Lernphase kaum realistisch simulieren.

Wir untersuchen in der vorliegenden Pilotstudie, wie stark die gemessenen Erkennungsraten unter sonst gleichen Bedingungen bei automatisiertem Test im Vergleich mit einem *Live*-Probandentest mit mehreren Testdurchläufen voneinander abweichen.

Testdesign

1. Sprecherauswahl: "Statische" Sprechermerkmale

Sprechermerkmale, die neben der Muttersprache in direktem Zusammenhang mit den akustischen Merkmalen der Sprachbefehle stehen, mit denen das System bedient wird, sind Alter, Geschlecht und dialektale Färbung. In der vorliegenden Pilotstudie wurden 12 Versuchspersonen (5w/7m) rekrutiert, die das mögliche Spektrum natürlich nicht vollständig abdecken können.

2. Geräuschkulisse & "dynamische" Sprechermerkmale

Ein weiterer Aspekt bei der Testplanung ist die Geräuschkulisse in der Fahrzeugkabine. Das Geräusch hat je nach Pegel und Zusammensetzung nicht nur direkten Einfluss auf den Spracherkenner, sondern auch Einfluss auch die Artikulation des Sprechers (Lombard-Effekt) und damit indirekt auf den Spracherkenner. Für die vorliegende Pilotstudie wurde ein Mittelklassefahrzeug ausgewählt. Die Performance des Spracherkenners wurde in 3 Fahrsituationen untersucht:

- Fahrzeug steht, Motor läuft, Lüfter Stufe 2 (**v0**)
- 130 km/h, Autobahnfahrt, Lüfter Stufe 2 (**v130**) (*vergleichsweise hoher TaskLoad der Vp durch die Fahraufgabe*)
- 50 km/h, wenig frequentierte, glatt asphaltierte Landstraße (**v50**) (*vergleichsweise geringer TaskLoad der Vp durch die Fahraufgabe*)

Von jeder Versuchsperson wurden zunächst Sprachbefehle in reflexionsarmer Umgebung aufgezeichnet. Das jeweilige Hintergrundgeräusch der 3 Fahrsituationen wurde während der Sprachaufnahme pegelrichtig über ein offenes Kopfhörersystem präsentiert, welches die akustische Eigenwahrnehmung nicht weiter beeinträchtigt.

Für den *Live*-Probandentest wurden die gleichen Versuchspersonen herangezogen. Jede Vp absolvierte zuerst die Sprachaufnahme im Labor, dann einen kompletten Probedurchlauf mit allen Sprachbefehlen im stehenden Fahrzeug, der nicht in die Messung einging, und schließlich die 3 *Live*-Tests unter den oben beschriebenen Fahrbedingungen. Das Fahren des Testfahrzeugs und die Eingabe der Sprachkommandos erfolgte durch die VP, ein Testbetreuer begleitete das Experiment auf dem Beifahrersitz.

3. Erkennervokabular ("Grammatiken")

Als dritter Testfaktor kommt die Auswahl des im Spracherkenner berücksichtigten Vokabulars hinzu. Die Untersuchung wurde mit einem speziell präparierten Spracherkenner durchgeführt, aus dessen Dialogstruktur 2 Grammatiken verwendet wurden:

- "Zifferngrammatik" – enthält alle Ziffern von 0 bis 9 mit der Möglichkeit der Verkettung aller Ziffern zu Ziffernfolgen beliebiger Länge
- "Befehlswortgrammatik" – enthält insgesamt 137 Einträge bestehend aus 15 allgemeinen Navigationsbefehlen (z.B. "Nordausrichtung" oder "Topografisch") sowie einer zufälligen Auswahl von 122 Städtenamen und ermöglicht mit jeder Aktivierung eine Einzelworterkennung

4. Im Test verwendete Sprachbefehle

Im Test wurden 35 Sprachbefehle in 4 Gruppen, verwendet:

1. Gruppe A: 5 Navigationsbefehle
2. Gruppe B: 10 Einzelziffern
3. Gruppe C: 10 Ziffernfolgen
4. Gruppe D: 10 Städtenamen

Statistische Auswertung

Hypothese:

"Ein automatisierter Labortest mit im reflektionsarmen Raum aufgezeichneten Sprachbefehlen führt NICHT zu

den gleichen gemessenen Erkennerraten wie der Live-Probandentest", weil die Vp während der Testwiederholungen eine Lernphase durchläuft, d.h. mit Hilfe des Feedbacks vom Spracherkenner – bewusst oder unbewusst – herausfindet, mit welcher Aussprachevariante, Artikulationsgenauigkeit, Sprechgeschwindigkeit, Stimmlage usw. sie vom Gerät am besten verstanden wird.
Es ist anzunehmen, dass die gemessene Erkennungsrate deshalb im *Live*-Probandentest bei ansonsten unveränderten Testbedingungen anfänglich mit jeder Testwiederholung steigt.
Weiterhin ist anzunehmen, dass der Lernerfolg nach ausgiebiger Lernphase (20 oder mehr Wiederholungen) eine "Sättigung" erreicht. (*Der experimentelle Aufwand zur Überprüfung dieser zweiten Annahme konnte jedoch im Rahmen der vorliegenden Studie nicht geleistet werden*.)

v0 automat	v0 live	v0 delta	v130 automat	v130 live	v130 delta	v50 automat	v50 live	v50 delta
93,33	**91,43**	**-1,9**	**84,29**	**83,93**	**-0,36**	**94,05**	**94,29**	**+0,24**

Tabelle1: Übersicht der gemessenen Erkennungsraten, prozentual

Mit dem zugrundeliegenden Stichprobenumfang kann die "wahre" Erkennungsrate jedoch nur sehr ungenau geschätzt werden. Zur statistischen Prüfung der Unterschiedshypothese ist ein Signifikanztest erforderlich. Aufgrund der vergleichsweise kleinen Grammatik und der Länge der verwendeten Wörter ist eine Tendenz zu sehr guter bis 100% Erkennung zu erwarten. Deshalb ist eine Normalverteilung der Messwerte unwahrscheinlich. Es wurde ein verteilungsfreies Verfahren zur Überprüfung der Hypothese gewählt, welches besonders für geringe Stichprobenumfänge geeignet ist: Der Wilcoxon-Test für Paardifferenzen [Wilcoxon, 1945, 1947].
Im Rahmen des Wilcoxon-Tests werden Paardifferenzen gebildet und alle **n** von Null verschiedenen Differenzen in einer Rangordnung platziert. Aufschluss über die Wahrscheinlichkeit der Nullhypothese (die besagt, das kein Unterschied in der zentralen Tendenz der der Stichprobe zugrunde liegenden Population besteht) geben die Variablen **T und T'.** Details zu deren Berechnung in [Bortz 2004]. Je deutlicher sich T und T' unterscheiden, umso unwahrscheinlicher ist die Nullhypothese.
Wäre die Nullhypothese richtig, so ließe sich ein mittlerer T-Wert **Tm** berechnen. Je deutlicher der empirische T-Wert vom berechneten Tm abweicht, umso geringer ist die Wahrscheinlichkeit, dass der gefundene Unterschied zufällig zustande gekommen ist.
Abhängig von n sind maximale T-Werte für die Signifikanzniveaus (Irrtumswahrscheinlichkeit der Hypothese: 0,5%, 1%, 2,5%) vorgegeben.

1. Vergleich: Automatischer Labortest und *Live*-Fahrerprobung bei **v0** (*Fahrzeug steht, Motor an, Lüfter Stufe 2*):

Systematischer Unterschied: Bei der *Live*-Fahrerprobung v0 haben die Vpn bereits einen Probedurchlauf mit dem Spracherkenner absolviert.

n	**2,5% Irrtum**	**Tm**	**T**	**T'**
8	T <= 4	18	17,5	18,5

Da der empirische T-Wert ungefähr gleich T' und ungefähr gleich Tm ist, kann man davon ausgehen, dass die gemessenen Unterschiede in der prozentualen Erkennungsrate bei v0 (Tabelle 1) zufällig durch die Stichprobe zustande gekommen sind.

2. Vergleich: Automatischer Labortest und *Live*-Fahrerprobung bei **v130** (*Autobahn, 130 km/h, Lüfter Stufe 2*):

Systematischer Unterschied: Bei der *Live*-Fahrerprobung v130 haben die Vpn bereits zwei Durchläufe mit dem Spracherkenner absolviert. Einen möglichen Lerneffekt könnte aber der TaskLoad der Fahrsituation entgegenwirken, der in der vorliegenden Untersuchung bei der Sprachaufnahme im Labor nicht simuliert wurde.

n	**2,5% Irrtum**	**Tm**	**T**	**T'**
6	T = 0	10,5	9	12

Da der empirische T-Wert ungefähr gleich Tm ist und nicht stark von T' abweicht, kann man davon ausgehen, dass auch bei v130 eventuelle Unterschiede in der prozentualen Erkennungsrate zufällig durch die Stichprobe zustande kommen. In der Tat ist die gemessene Abweichung (Tabelle 1) gering.

3. Vergleich: Automatischer Labortest und *Live*-Fahrerprobung bei **v50** (*gut asphaltierte, wenig frequentierte Landstrasse, Lüfter Stufe 2*):

Systematischer Unterschied: Bei der *Live*-Fahrerprobung v50 haben die Vpn bereits drei Durchläufe mit dem Spracherkenner absolviert. Da es sich nicht um Fahranfänger handelt, sollte der TaskLoad der Fahrsituation in diesem Fall vernachlässigbar sein.

n	**2,5% Irrtum**	**Tm**	**T**	**T'**
6	T = 0	10,5	7	14

Dass der empirische T-Wert nur halb so groß wie T' ist, spricht dafür, dass im gegebenen Szenario tatsächlich ein Unterschied in der Erkennungsrate auftritt, der durch einen Lerneffekt der Probanden zu begründen ist. Die Irrtumswahrscheinlichkeit dieses Ergebnisses ist jedoch hoch, der Unterschied ist nicht signifikant.

Ergebnis

Wir konnten unsere generellen Bedenken, dass ein automatisierter Spracherkennertest eine härtere Prüfung darstellt und zu niedrigeren Erkennungsraten führt als ein *Live*-Probandentest, zumindest für den Fall ausräumen, dass die Anzahl der Testdurchläufe für die einzelne Versuchsperson gering bleibt. Mit zunehmender Anzahl von Testwiederholungen scheint sich jedoch ein Lerneffekt bei den Vpn abzuzeichnen.

Literatur

Bortz, J., *Statistik*, Berlin: Springer 2004
Herrenkind, M., Klasmeyer, G., Kreft, K.: *Evaluierung der Erkennungssicherheit von sprachbedienten Systemen im Kfz*, DAGA'06, Braunschweig (2006)
Wilcoxon, F. *Individual comparisons by ranking methods*, Biometrica 1 , p.80-83