Sprachgesteuerte Fahrerassistenz durch einstufig-probabilistisches Verstehen natürlich gesprochener Sprache

Voice controlled driver assistance by single-stage probabilistic natural language understanding

Dr.-Ing. Robert Neuss (1), Dipl.-Phys. Jörg Hunsinger (2), Dipl.-Inform. Richard Stenzel (3), Prof. Dr. rer. nat. Manfred Lang (2)

(1) Usaneers GmbH, Paul-Lagarde-Str. 18, D-80686 München, Tel. 089/ 51 72 89 35, email robert.neuss@usaneers.de
(2) Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, D-80290 München, Tel. 089/ 289 285 48, email {hunsinger, lang}@ei.tum.de
(3) GMD-IPSI, AMBIENTE, Dolivostr. 15, D-64293 Darmstadt. Tel. 06151/ 869 4952, email stenzel@darmstadt.gmd.de

Kurzbeschreibung

Heutige sprachgesteuerte Mensch-Maschine-Interfaces in Automobilen sind auf schlüsselwortbasierte Eingaben mit drei wesentlichen Nachteilen beschränkt: Spontansprachliche Äußerungen werden zurückgewiesen, Anwender müssen eine gewisse Lernkurve durchlaufen, und Langzeitnutzer stört die Eintönigkeit der verfügbaren Eingabemuster. Dieser Beitrag beschreibt die Implementierung eines sprachverstehenden Systems zur Steuerung von Telefon und Navigationssystem im Kraftfahrzeug mittels spontansprachlicher Eingabe. Der erste Schritt bestand aus dem Aufbau einer natürlichsprachlichen Datenbasis. Hierzu wurde ein Wizard-of-Oz-Experiment mit 49 selektierten Versuchspersonen durchgeführt. Der verfolgte Lösungsansatz basiert auf einem einstufigen, erwartungsgetriebenen semantischen Decodierungsverfahren, das auf allen beteiligten Abstraktionsebenen von der akustischen bis zur syntaktisch-semantischen Ebene eine probabilistische Wissensmodellierung beinhaltet. Das vorgestellte System hat bei guten akustischen Bedingungen eine Erkennungsrate von 90%, wobei ganze Sätze insgesamt besser interpretiert werden können als Einzelkommandos.

Summary

Current voice controlled human-machine interfaces in cars are limited to keyword based input which is characterized by three major drawbacks: People who are not accustomed to ASR systems experience difficulties when natural utterances are rejected (general entrance barrier), "normal" users have to pass a certain learning curve to find out the supported commands (first time usage) and many non-professional long-time users are simply annoyed by the monotonous input patterns (long time usage). This paper describes the implementation of a robust voice driven interface to control a phone and a navigation system by spontaneous speech in a mobile environment. The first step was to generate a database of natural utterances. This was accomplished by Wizard-of-Oz experiments in a driving simulator with 49 selected human subjects. Our approach is based on an approved single-stage top-down semantic decoder which utilizes probabilistic knowledge bases from the signal up to the syntactic-semantic abstraction levels. A recognition accuracy of approximately 90% results under good background noise conditions, whereby processing of whole sentences works better than single commands.

Quelle: VDI-Bericht 1678 "USEWARE 2002, VDI-Verlag, Düsseldorf

Das Sprachverstehende System

An der Technischen Universität München wurde ein sprachverstehendes System entwickelt, welches spontansprachliche Eingaben im Rahmen einer spezialisierten Domäne in semantische Gliederungen überführt [1]. Diese Gliederungen können durch einen Intentionsdecoder in konkrete Anweisungen zur Steuerung einer Applikation transformiert werden. Die ursprüngliche Domäne bestand aus einem Grafikeditor zur Erzeugung und Manipulation grafischer Objekte („NASGRA“ = „Natürlichsprachlicher Grafikeditor“).

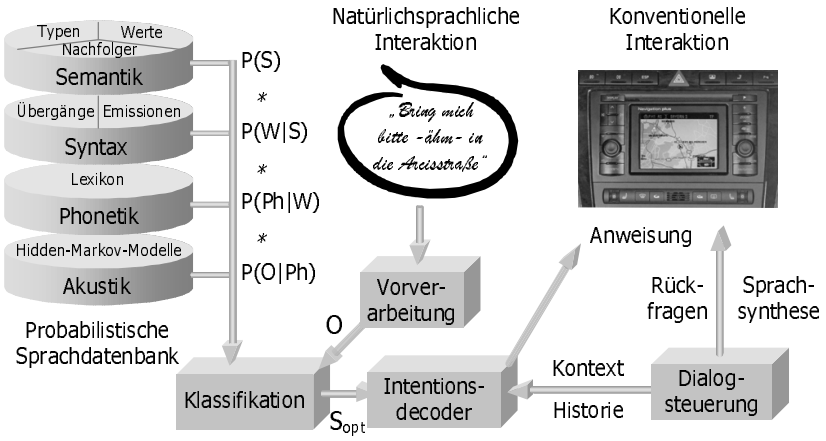

Abb. 1: Schematischer Aufbau des einstufigen sprachverstehenden Systems in der Fahrzeugdomäne. Die probabilistische Sprachdatenbank liefert statistische Bewertungen für mögliche Bedeutungsinhalte P(S), zugehörige Wortketten P(W|S) und Phonemfolgen P(Ph|W), die phonembasierte Spracherkennungsschicht bewertet deren Übereinstimmung P(O|Ph) mit dem vorverarbeiteten Sprachsignal O. Als Erkennungsergebnis wird die beste Semantische Gliederung S_{opt} an den Intentionsdecoder weitergeleitet.

Der Hauptvorteil dieses Systems liegt im einstufigen Ansatz, da verschiedene Beobachtungshypothesen einer Äußerung unmittelbar mit einer statistischen Sprachdatenbasis verrechnet werden. Auf diese Weise werden die Wahrscheinlichkeiten verschiedener möglicher Bedeutungsinhalte und zugehöriger Wortkombinationen schritthaltend mit dem vorverarbeiteten Sprachsignal (Beobachtungsfolge) verglichen, was dazu führt, dass Ambivalenzen bei der Erkennung anhand der Datenbasis aufgelöst werden. Im Gegensatz hierzu arbeiten konventionelle Systeme zweistufig, d.h. zuerst liefert ein Spracherkenner ein Erkennungsergebnis, welches anschließend durch einen Syntax-Parser ausgewertet wird. Dies ist zwar prinzipiell leichter durchzuführen, da Spracherkenner und Parser entflochten sind, jedoch werden unter Umständen brauchbare Erkennungshypothesen frühzeitig (aufgrund des fehlenden syntaktisch-semantischen Zusatzwissens) verworfen.

Versuchskonzept

Zur Gewinnung von authentischen Sprachäußerungen als Grundlage für eine Sprachdatenbasis wurde ein Wizard-of-Oz-Versuch im Usability-Labor der TU München mit 49 Teilnehmern gestartet. Die Anzahl der Versuchspersonen spiegelt die Anforderung an die Datenbasis wieder, eine ausreichend hohe Varianz von Sprachäußerungen zu enthalten, welche signifikant für die Qualität des resultierenden Systems ist. Um die Domäne abzugrenzen, wurden die Testteilnehmer mit einem prototypischen Bordcomputer konfrontiert, welcher ein

Display besaß und den Probanden somit indirekt Begriffe anbot. Der Versuchsleiter stellte Aufgaben wie das Programmieren eines Navigationszieles, welches per Sprache schrittweise oder durch einen einzigen Satz gelöst werden sollte. Der Bordcomputer wurde vom Versuchsleiter in Abhängigkeit der Spracheingaben in kooperativer Weise ferngesteuert, wobei Sprachfeedbacks vorbereitet waren und „Fehleingaben“ gemäß Testskript künstlich provoziert wurden, um eine möglichst realistische Illusion zu schaffen.

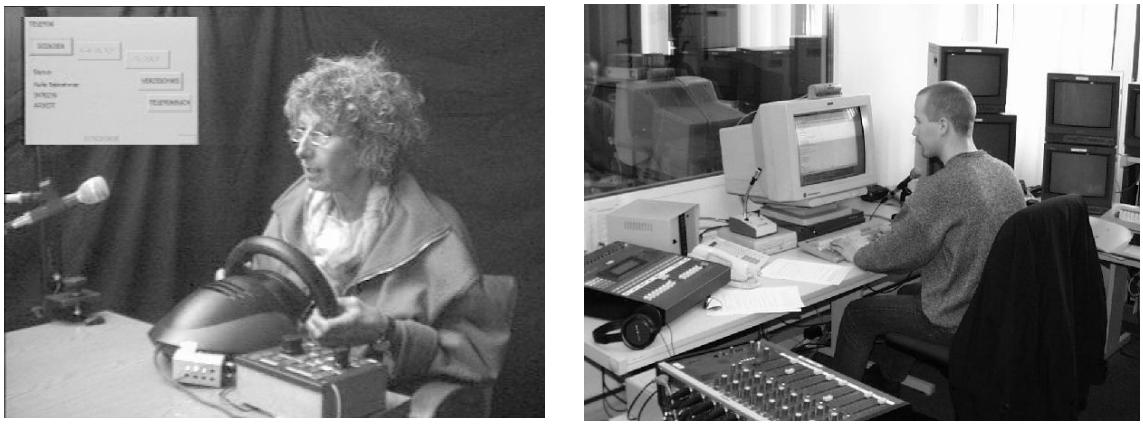

Abb. 2: Links: Versuchsperson in der Aufzeichnungskammer, links oben ist die Ansicht des „Bordcomputers“ eingeblendet. Rechts: Versuchsleiter im Regieraum.

Damit eine ungünstige Wahl von Funktionsnamen den Wortschatz nicht unnötig einschränkte, wurden zwei komplette Sätze von Schaltflächenbeschriftungen vorgehalten, die interpersonell gewechselt wurden. Weiterhin waren zwei verschiedene Arten von Sprachausgaben implementiert (maschinelle „Roboterstimme“ mit Samplefrequenz 8 kHz bzw. professionelle Synchronsprecherin in CD-Qualität), die ebenfalls interpersonell ausgetauscht wurden. Diese Variation basierte auf der These, dass die Systemstimme Einfluss auf das „Weltbild“ des Benutzers vom System und somit sein Wording haben könnte. Da die Annahme bestand, dass Menschen gegenüber einem „menschlich“ anmutenden System zu mehr Höflichkeit und Formulierungsvielfalt tendieren als bei einem „maschinellen“ System, sollte die Variation des Stimmtyps zeigen, ob die Verwendung einer menschlichen Stimme nicht Nachteile mit sich bringt (Eine Animation der Benutzer zu „blumigeren“ Formulierungen wäre aus Sicht des sprachverstehenden Systems unvorteilhaft). Sinngemäß wurde das Bordcomputersystem explizit als „Computer“ oder „System“ bezeichnet, und den Versuchspersonen erläutert, dass der Adressat der Sprachäußerungen der Computer und nicht der Versuchsleiter sei.

Eine Hauptschwierigkeit bei der Durchführung der Tests bestand darin, die Probanden zu instruieren, ohne Wortwahl und Formulierung durch die Aufgabenstellung zu suggerieren. Da frühere Versuche gezeigt haben, dass Anweisungen der Art „Wählen Sie die Nummer 089-12345“ oder „Programmieren Sie das Fahrziel XY-Strasse“ die Sprachäußerungen stark beeinflussen, wurden alternative Beschreibungen wie Visitenkarten gewählt. Für Rückfragen an die Probanden enthielt das Testskript, aus welchem der Versuchsleiter wortgenau vorlas, vorgefertigte Antworten wie „Das System kennt die Firma X nicht, aber auf der Visitenkarte haben Sie noch eine weitere Information.“. Es wurden etwa 10 verschiedene Aufgaben zur elementaren Bedienung des Telefons und des Navigationssystems gestellt. Um ausreichende Varianz zu erhalten wurden die Aufgaben im Verlauf der Versuchssitzung im Rahmen von zwei Versuchsabschnitten bis zu dreimal gestellt. Im ersten Teil wurden die Probanden in-

struiert völlig frei zu formulieren, wobei das Bordcomputersystem noch unbekannt war. Nach diesem Erstkontakt wurde neben einer Einführung in das Bordcomputersystem die Beschränkung vorgegeben, keine Nebensätze zu verwenden. Der Versuch selbst wurde in einer schallisolierten Aufzeichnungskammer und nicht im Fahrsimulator durchgeführt, um bessere Aufnahmebedingungen zu gewährleisten. Es wurde dabei keine Fahraufgabe gestellt, sondern nur Fahrsequenzen ohne Ton per separatem Videobildschirm eingespielt, um die Aufmerksamkeit etwas vom Bordcomputerdisplay abzulenken.

Auswertung

Das Durchschnittsalter der Probanden betrug 42,9 Jahre, von den 49 Personen waren 21 weiblich und 28 männlich, 26 Personen waren über und 23 unter 40 Jahre alt. Insgesamt wurden 1538 Äußerungen aufgezeichnet, davon 564 im ersten und 974 im zweiten Versuchsteil. Zunächst wurden die Eingaben auf Anzahl der Bedeutungseinheiten, Satzlänge und Wortschatz untersucht, um die technischen Anforderungen an den Spracherkenner abschätzen zu können. So gab es im ersten Versuchsabschnitt 483 (85,6%) im Wortlaut unterschiedliche Äußerungen, im zweiten Teil 568 (58,0%). Bei einer Gesamtzahl von 9728 Einzelworten bestanden die Anweisungen im Mittel aus 6,32 Wörtern. Eine quantitative Auswertung bestätigt, dass Personen mit technischer Erfahrung kürzere, prägnantere Sätze verwenden.

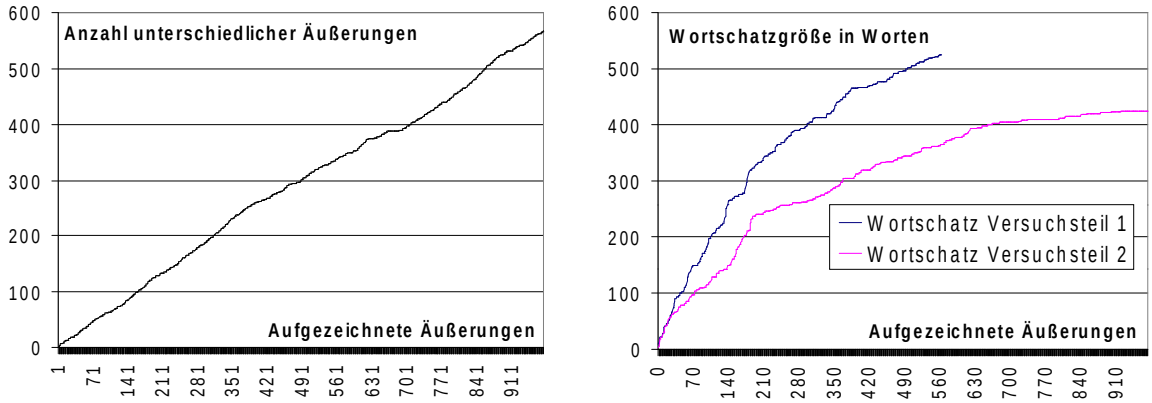

Abb. 3: Links: Die Zahl verschiedener Äußerungen (Y-Achse) pro aufgezeichneter Äußerungen (X-Achse) nimmt konstant zu. Rechts: Anwachsen des Gesamtwortschatzes mit der Anzahl ausgewerteter Anweisungen. Aus dem Diagramm lässt sich abschätzen, dass für die gewählte Domäne mindestens 1000 Äußerungen benötigt werden, um den benötigten Wortschatz abzudecken.

Im Gegensatz zur Anzahl unterschiedlicher Äußerungen gab es bei den beobachteten Wörtern eine Sättigungstendenz zu beobachten, welche im zweiten Versuchsteil bei einem niedrigeren Niveau eintritt. Dies ist vermutlich darauf zurückzuführen, dass die Testpersonen aufgrund der Nebensatzeinschränkung und der besseren Systemkenntnis zielgerichteteres Vokabular verwendeten. Das häufigste Wort war „bitte“ (407mal), dann erst kamen domänenspezifische Ausdrücke wie Ziffern und Eigennamen. Häufige Füllwörter waren „ich“ (227), „möchte“ (94) oder „hallo“ (53).

Die wichtigste Folgerung aus diesen Beobachtungen ist, dass der Wortschatz für die gewählte Domäne zwar begrenzt ist, jedoch sehr viele syntaktische Variationen möglich sind und diese in der Praxis auch auftreten. Dies ist ein Indiz dafür, dass eine Lösung auf statistischer Basis naheliegender ist als eine statische Grammatik mit fester Struktur. Andererseits treten häufig Füllwörter („bitte“, „ich möchte“, Partikel, etc.) auf, weswegen es sich lohnen könnte, eine konventionelle Spracherkennung mit einem entsprechenden Garbage-Modell zu erweitern. Die Daten zeigen ferner, dass sich die für die hier verwendeten Algorithmen notwendige Nebensatzeinschränkung günstig auf die Mächtigkeit des Wortschatzes und die Satzlängen auswirkt, vorausgesetzt die Benutzer halten sich daran. Trotzdem besteht für das Sprachverstehen das prinzipielle Problem von Out-of-Vocabulary-Eingaben. Diese wirken sich im hier vorgestellten System jedoch nur dann aus, wenn die zugehörigen Wörter a) bedeutungstragend und b) ähnlich zu vorhandenen Wörtern mit abweichender Bedeutung aber ähnlichem Kontext sind. Nur in diesem Falle ist die Wahrscheinlichkeit einer Fehlinterpretation hoch.

Ein weiterer Aspekt der Untersuchung war die „Personifizierung“ des Systems. So kamen nur in etwa 5% der Äußerungen Anreden wie „Computer“ (20), „System“ (19), „Sie“ (19), „Auto“ (17) oder „du“ (6) vor, wobei in diesem Versuch kein Schlüsselwort wie z.B. „Computer“ notwendig war (z.B. „Computer, wähle die Nummer xx“). Etwa 16% der Äußerungen wurden in Ich-Form getätigt (z.B. „Ich möchte die Nummer xx wählen“), hingegen wurde die Imperativ-Form ohne Anrede in 20-30% der Fälle gewählt (z.B. „Wähle die Nummer xx“). Die Hälfte und somit die Mehrzahl der Eingaben erfolgten in Infinitiv-Form (z.B. „Nummer xx wählen“).
Das unterschiedliche Systemfeedback per Roboterstimme (Gruppe A mit 16VPn) bzw. menschlicher Stimme (Gruppe B mit 33 VPn) wirkte sich folgendermaßen aus: Von 1538 Äußerungen (ca. 31/VPn) enthielten 399 Sätze das Wort „bitte“ (u.U. mehrfach), davon 130 in der Gruppe A (= 8,12/VPn) und 269 in Gruppe B (= 8,15/VPn). Andererseits verwendeten Probanden der Gruppe A geringfügig kürzere Sätze (siehe Diagramm).

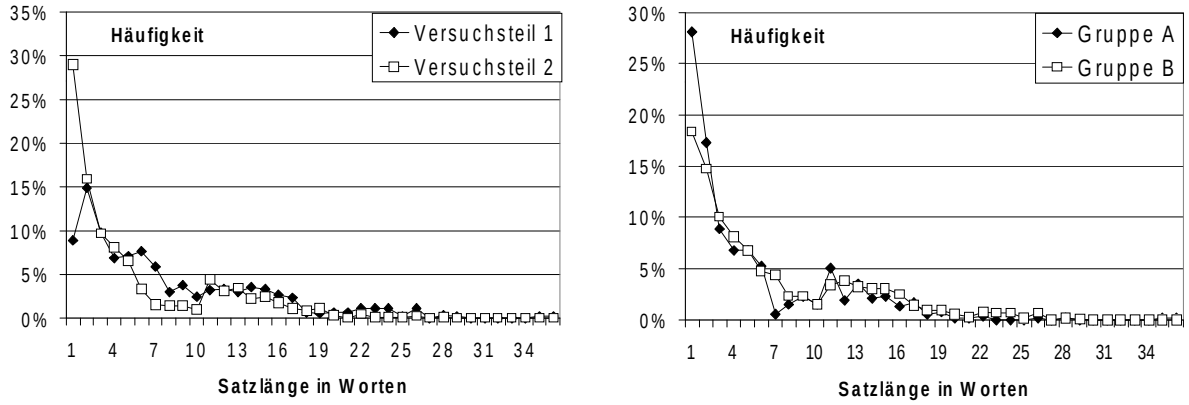

Abb. 4: Links: Satzlängen im ersten (∅ 7,99 Wörter/Satz) und zweiten Versuchsteil (∅ 5,36 W/S), deutlich ist der Zuwachs von Eingaben mit nur einem Wort. Rechts: Satzlängen für Gruppe A (Computerstimme, ∅ 5,4 Wörter/Satz) und Gruppe B (∅ 6,79 W/S). Der Unterschied ist nicht so deutlich wie links.

Diese Zahlen zeigen, dass die Systemstimme keine signifikante Auswirkung auf die „Höflichkeit“ der Benutzer gegenüber dem System hatte.

Ergebnisse

Von den 1538 aufgezeichneten Sprachäußerungen wurden 916 in den akustischen Trainingskorpus übernommen (es wurden diejenigen mit Nebensätzen, erfundenen Eigennamen und ungenügender akustischer Qualität entfernt). Zur Bildung einer Erkennungsrate dienten diese Äußerungen gleichzeitig als Testeingabe für das sprachverstehende System (Reklassifikation), wobei 92,58% der Äußerungen korrekt klassifiziert wurden. Betrachtet man nur die Intentionen unter Vernachlässigung ihrer Parameter, so wurde eine Zuordnungsrate von 97,7% erreicht. Der Unterschied von 5,12% wurde zu 98% durch die fehlerhafte Erkennung von Telefonnummern verursacht, da für die Ziffern keine statistischen Bindungen genutzt wurden. Die unter realen Bedingungen zu erwartende Erkennungsleistung wird sicher unter den oben genannten Werten liegen, da zum einen kein Fremdtest (mit unabhängigem Sprachmaterial) durchgeführt wurde und zum anderen die Äußerungen im Labor mit definierten akustischen Rahmenbedingungen aufgezeichnet wurden.

Das resultierende System verarbeitet Eingaben auf einem PC (400 MHz, 256 MB, Betriebssystem Windows NT) in fünffacher Echtzeit, wobei die aus dem Verbmobil-Projekt [2] stammenden Spracherkennungsalgorithmen noch verbesserungsfähig sind. Das sprachverstehende System verbessert die intuitive Bedienung bei Erstkontakt, liefert jedoch bei Eingaben im Kommandostil (Eingaben mit 1-3 Worten) und bei Telefonnummern schlechtere Ergebnisse als ein konventioneller Spracherkenner. Ein weiterer Nachteil ist die Begrenzung der Eigennamen im Test auf wenige Dutzend.

Als Hauptergebnis lässt sich festhalten, dass das Prinzip Sprachverstehen erfolgversprechend ist. Eine Vorraussetzung für den realen Einsatz wäre jedoch die Fähigkeit, zunächst unbekannte Eigennamen (z.B. für eine Navigationszieleingabe mit 100.000 Städtenamen) oder Nummernblöcke als solche zu kennzeichnen. Diese könnten dann direkt an einen zweiten, konventionellen Spracherkenner weitergereicht und von diesem verarbeitet werden.

Literatur:

[1] Stahl H., Müller J., and Lang M.: Controlling Limited-Domain Applications by Probabilistic Semantic Decoding of Natural Speech. Tagungsband „IEEE International Conference on Acoustics, Speech, and Signal Processing“ (ICASSP) 1997, München.

[2] Bub T., Schwinn J.: “Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System.”, Proceedings of the International Conference of Spoken Language Processing; pp 2371-2374; Philadelphia, PA; 1996.

[3] Stenzel R.: Natürlichsprachliche Eingabe für das Automobil, Diplomarbeit, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 1999.

[4] Neuss R.: Usability-Engineering als Ansatz zum multimodalen Mensch-Maschine-Dialog, Dissertation, Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 2001.

[5] Hunsinger J., Lang M.: A Speech Understanding Module for a Multimodal Mathematical Formula Editor. Tagungsband „IEEE International Conference on Acoustics, Speech, and Signal Processing“ (ICASSP) 2000, Istanbul, Türkei, 5.-9.6.2000, Bd. 4, S. 2413-2416.