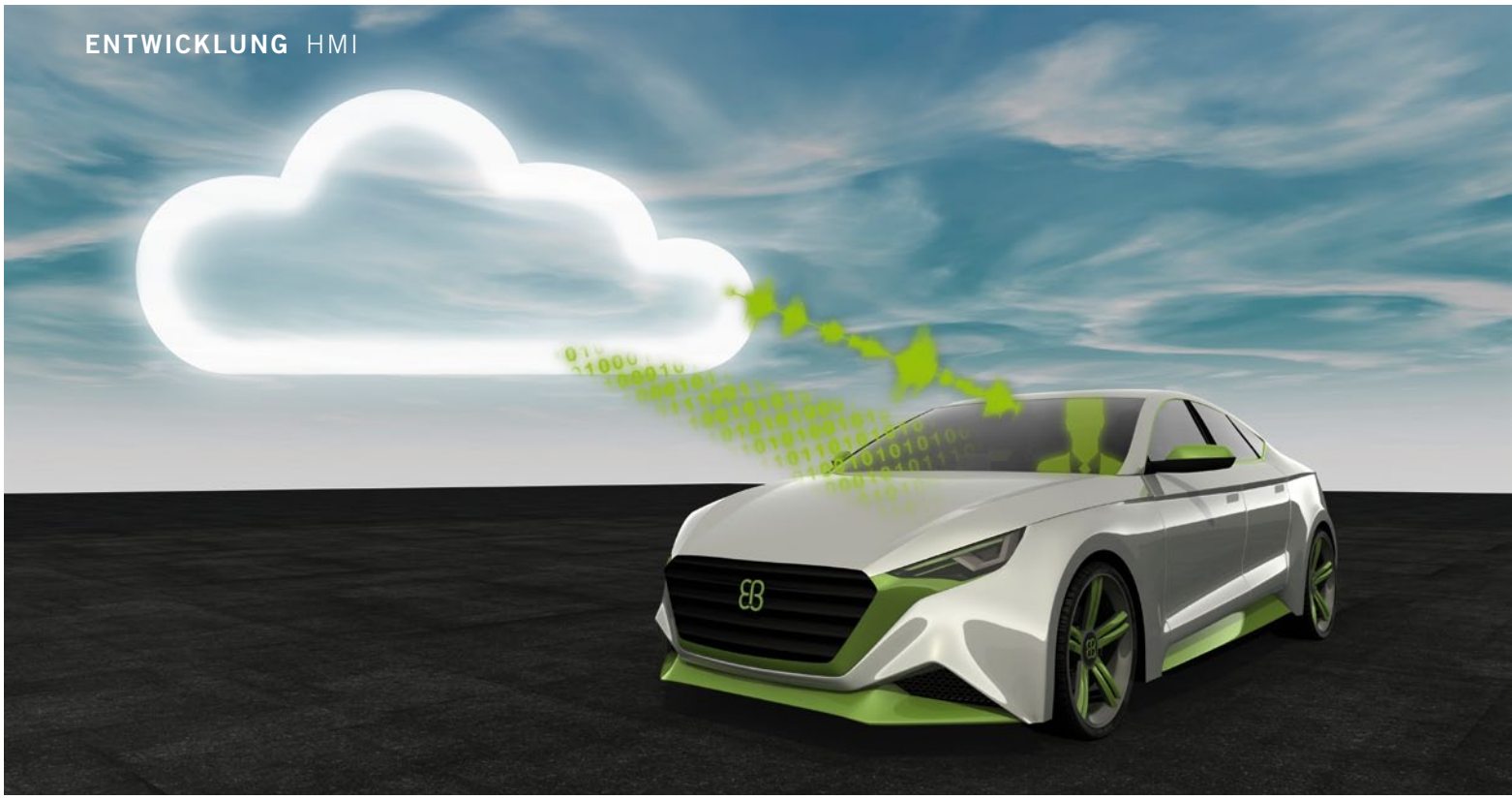


AUTOREN

DOMINIQUE MASSONIÉ
ist Produkt Manager HMI
bei Elektrobit in Erlangen.

TIMO SOWA
ist Software-Entwickler HMI
bei Elektrobit in Erlangen.

GRUNDLAGEN VON SPRACHDIALOGSYSTEMEN

Sprachdialogsysteme sind ein wichtiger Baustein dafür, die Bedienung von Informations- und Kommunikationssystemen im Auto einfacher und damit auch komfortabler und sicherer zu machen. Der Fahrer kann die Augen auf die Straße gerichtet lassen und wird so bei der Bedienung von Infotainment-Systemen weniger vom Verkehrsgeschehen abgelenkt. Je mehr Freiheiten die Sprachbedienung dem Fahrer bei der Formulierung seiner Wünsche lässt, umso weniger Einarbeitungszeit in die Systembedienung ist erforderlich. Im Idealfall sind keine Vorkenntnisse notwendig, um die gewünschten Funktionen per Sprache zu steuern. „Natural Language Understanding“ oder kurz NLU ist deshalb ein wichtiges Ziel.

Allerdings ist es wichtig zu verstehen, dass NLU keine Technologie per se ist, sondern ein Designprinzip. Um dieses Ziel zu erreichen können unterschiedliche Spracherkennungsmethoden eingesetzt werden.

Der Ausgangspunkt eines Sprachdialogsystems ist eine phonem-basierte Spracherkennnung: Worte werden in ihre kleinsten akustischen Bestandteile zerlegt – die Phoneme, ähnlich wie Silben. Die Erkennung basiert darauf, das Eingangssignal mithilfe von Phonem-Wörterbüchern der wahrscheinlichsten Sequenz von Phonemen zuzuordnen, ❶. Die Ergebnisse sind in den vergangenen fünf Jahren deutlich robuster geworden. Und auch die bisher problematische Konstellation von multilingualen Eingaben (etwa bei fremdsprachigen Musiktiteln, Eigennamen oder Navigationszielen in einem ansonsten deutschsprachigen Kontext) ließ sich durch den Einsatz mehrsprachiger Wörterbücher deutlich verbessern.

Die Interpretation der gesprochenen Eingaben erfolgte bislang in erster Linie grammatikbasiert. Die zulässigen Spracheingaben waren dabei durch Konstruktionsregeln streng vorgegeben, was die Analyse der gesprochenen Befehle und Antworten vergleichsweise einfach machte, aber nur vorher festgelegte Satzfolgen zuließ.

In moderneren und flexibleren Systemen wird die grammatische Interpretation deshalb durch eine separate semantische Analyse (auch als „Topic Classification“ bezeichnet) ergänzt. Hier steht die Frage im Mittelpunkt, was der Benutzer mit seiner Eingabe gemeint hat. Dies stellt insbesondere bei der Interpretation längerer Sätze große Herausforderungen an die Verständnisfähigkeit des Systems, zumal abhängig vom Kontext eine mögliche Ambiguität, also Mehrdeutigkeit,

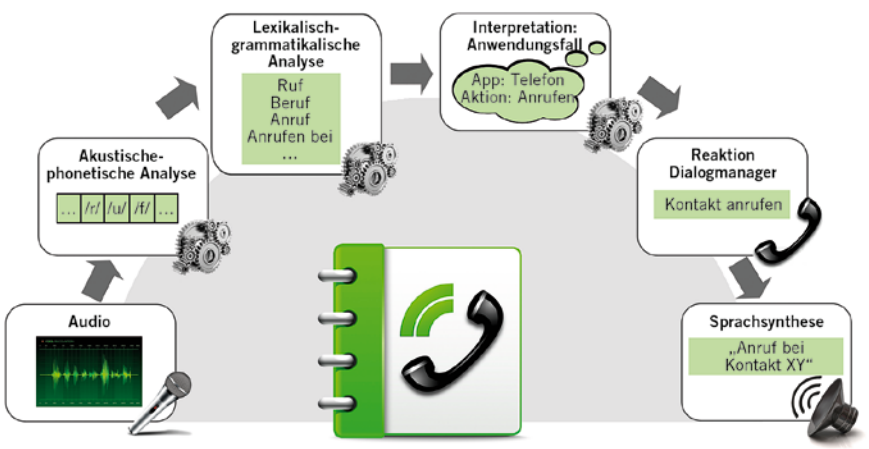
CLOUD-UNTERSTÜTZT
AUTOS IM DIALOG MIT IHREM FAHRER

Der Funktionsumfang und die Möglichkeiten von internetbasierten Infotainment-Systemen im Auto wachsen beständig. Ein entscheidender Aspekt bleibt in diesem Zusammenhang aber die Bedienung. Der Fahrer soll sich auf das Verkehrsgeschehen konzentrieren, anstatt abgelenkt zu werden. Sprachsteuerungssysteme tragen hier zu effizienten Lösungen bei. Sind diese Cloud-basiert, hat dies nicht nur Vorteile. Elektrobit diskutiert das Für und Wider und stellt neue Lösungen der Spracherkennung vor.

gelöst werden muss. Ein praktikabler Ansatz ist es, Unsicherheiten im Dialog zu lösen. Das System stellt also bei Bedarf Rückfragen, um Kontext und Bedeutung zu klären. Dadurch werden die möglichen Dialogabläufe komplexer, wirken für den Fahrer aber natürlicher und somit angenehmer und stressfreier.

Die Sprachausgabe im Rahmen von Dialogsystemen basiert in der Regel auf Text-to-Speech (TTS), also Sprachsynthese. In der Regel erlaubt die Eingabe dadurch mehr Freiheitsgrade hinsichtlich Aussprachevarianten, während die Sprachausgabe nur eine einzige Aussprachevarinate vorsieht. Eine logische Forderung an eine leistungsfähige, „natürlich" wirkende Lösung wäre, dass Input und Output zu 100 % kompatibel sind – das System also zum Beispiel bei mehrsprachigen Dialogen in seinen Antworten für jedes Wort dieselbe Sprache wählt wie sein Nutzer bei der Eingabe. Das stellt Sprachdialogsysteme jedoch bislang vor kaum lösbare Aufgaben.

Ein weiterer interessanter Trend bei der Modellierung von Human Machine Interfaces (HMI) heißt Multimodalität: Heute wird eine Spracheingabe in erster Linie durch Drücken einer Sprachtaste am Lenkrad aktiviert. In Zukunft wären zusätzliche oder alternative Bedienschritte möglich, etwa in Kombination mit Touchscreens oder kamerabasierter Gestenerkennung. Dadurch würden neuartige Bedienkonzepte ermöglicht – der Fahrer könnte zum Beispiel auf einem Touchscreen einen Punkt auf der dort angezeigten Karte antippen und dazu fragen „Was ist das?" oder dem Navigationssystem befehlen: „Bringe mich hier hin."



❶ Sprachdialogprozess am Beispiel „Anruf bei Kontakt" aus Adressbuch

AKTUELLE TECHNOLOGIEN ZUR SPRACHERKENNUNG

Wie bereits erwähnt, verschmolz bei älteren Lösungen die grammatikbasierte Spracherkennung mit einer eingeschränkten semantischen Analyse. Aus dem Dialogablauf ergibt sich eine mehr oder weniger starr vorgegebene Formulierung der Eingabe, die das System bei seiner Analyse voraussetzt, ❷. Moderne Systeme trennen diese Bearbeitungsschritte – die semantische Analyse wird dabei auf die Ergebnisse der grammatischen Interpretation angewandt.

Eine wichtige Voraussetzung für natürliches Verstehen beziehungsweise flexiblere Spracherkennung ist die Erkennung auf Basis statistischer Sprachmodelle. Man spricht auch von Statistical Language Models (SLM). Dabei werden zulässige Eingaben anhand eines statistischen Modells beschrieben, das aus Beispielen

gewonnen wird. Die Systeme lassen sich durch Verarbeitung großer Beispielmengen trainieren – das Training füllt die Modelle mit Wahrscheinlichkeitswerten für das Auftreten bestimmter Wortkombinationen beziehungsweise das Auftreten eines Wortes an einer bestimmten Stelle im Satz. So ermittelt das System bei der Analyse einer Eingabe, welches Wort an der jeweiligen Stelle und im jeweiligen Kontext am meisten Sinn ergeben würde. Ein solches Modell muss auf Basis eines Trainings-Corpus für jede unterstützte Sprache erzeugt werden, was zu vergleichsweise großen Datenmengen von mehreren Hundert Megabyte pro Sprache führt. Umfangreiche Trainings-Corpora werden ebenso wie die statistischen Modelle von Drittanbietern wie Nuance oder Voicebox geliefert.

Zusätzliche Herausforderungen stellen dynamische Inhalte dar – also Namen und Begriffe, die nicht von vornherein im Wörterbuch enthalten sind, sondern sich zum Beispiel aus dem Kontaktverzeichnis des Nutzers ergeben. Letzteres kann bei modernen Systemen mehrere Tausend Namenseinträge enthalten und soll im Rahmen des Sprachdialogs adressierbar sein. Ähnliches gilt für Kalendereinträge, Musiktitel, Point-of-Interest-Einträge in der Navigationsdatenbank und vieles mehr. Da die zur Erkennung notwendige Lautschrift der eingetragenen Namen in der Regel nicht hinterlegt ist, wird sie beim Import solcher Daten ins Infotainment-System anhand von Regeln generiert („Phonemisierung“). Außerdem werden in diesem Arbeitsschritt abgekürzte Einträge der Kontakt- und Adresslisten wie zum Beispiel „Dr.“, „Hbf“ oder „geb.“ für die Spracherkennung normalisiert – also so adaptiert, dass die Spracherkennung sie Eingaben wie „Doktor“, „Hauptbahnhof“ oder „geborene“ zuordnen kann.

Neue Dienste der Infotainment-Systeme wie Nachrichten aus Politik und Sport, Spritpreise und ähnliches führen weitere Inhalte ein, die von der Spracherkennung verstanden werden müssen und die das Sprachdialogsystem bei seinen Ausgaben in Sprache umsetzen können muss.

SPRACHERKENNUNG IN DER CLOUD

Solche Aufgaben sowie insgesamt der Bedarf an höherer Rechenleistung und Speicherkapazität, wie sie etwa zur besseren Unterstützung von Multilingualität benötigt werden, führten zu der Überlegung, Teile der Spracherkennung in die Cloud auszulagern. Die Endnutzer kennen solche Lösungen aus der Smartphone-Welt von Systemen wie Apples Siri, Google Now oder Samsungs S-Voice.

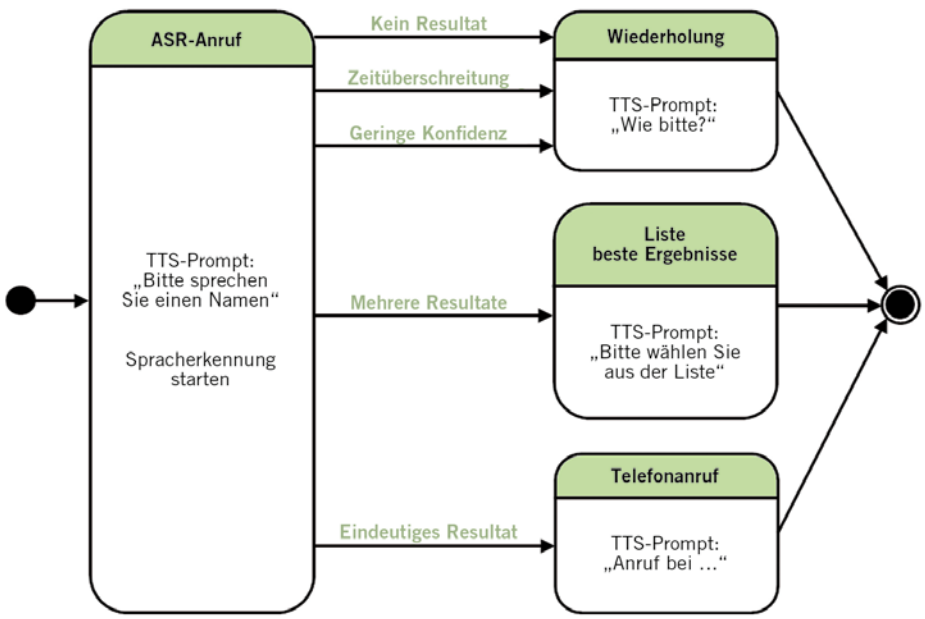
Die Vorteile liegen auf der Hand: Die verfügbare größere Leistung erlaubt komplexere Eingaben und ist zudem in der Lage, die Erkennungsleistung schnell an neue, aktuelle Inhalte wie etwa Nachrichtenmeldungen, Sportergebnisse, Börsendaten und ähnliches anzupassen. Da Tausende von verschiedenen Benutzern das System in unterschiedlichem Kontext nutzen, verbessert sich dadurch auch kontinuierlich die Erkennungsleistung. Auch die Erkennung und Handhabung mehrerer Sprachen profitiert von der Nutzung Cloud-basierter Lösungen in mehreren Ländern.

Doch auch die Nachteile sind zu bedenken: Der Zugriff auf Cloud-Dienste erfordert eine schnelle Internetverbindung, was im Auto gleichbedeutend mit sehr guter Mobilfunkabdeckung ist. Insbesondere in Roaming-Situationen kommt zum Netzempfang die Frage nach den Übertragungskosten. Zudem wirft die Nutzung von Cloud-Diensten auch Fragen der Privatsphäre auf: Erkannte Eingaben wie Navigationsziele, Kontakte oder Kalendereinträge sind private Daten des Nutzers, lassen sich dem einzelnen Kunden aber eindeutig zuordnen. Neben den eigentlichen Diensteanbietern sind etwa auch die Mobilfunkprovider sowie insbesondere die Betreiber von Cloud-Infrastrukturen wie Amazon oder Microsoft an Transport, Speicherung und gegebenenfalls Verarbeitung der persönlichen Nutzerdaten beteiligt. Dabei müssen OEMs bedenken, dass für ihre Angebote in verschiedenen Ländern unterschiedliche technische und juristische Rahmenbedingungen gelten.

Die Entscheidung ob, und in welchem Umfang Cloud-basierte Spracherkennungssysteme eingesetzt werden sollen, obliegt somit letztlich dem OEM, der eine entsprechende Lösung in Auftrag gibt.

ÜBERLASTUNG DES FAHRERS VERMEIDEN

Wie bereits eingangs erwähnt, gilt Spracheingabe als wichtiger Ansatz zur Verminderung der Ablenkung des Fahrers beziehungsweise zur Reduktion seiner Belastung in bestimmten Fahr- und Bediensituationen („Driver Distraction“



❷ Ablaufdiagramm der Sprachdialogschritte am Beispiel Adressbuch

und „Driver Workload“). Allerdings muss auch konstatiert werden, dass Sprache im Auto nicht nur Teil der denkbaren Lösung, sondern auch Teil des Problems ist. Weltweit sind die Gesetzgeber derzeit dabei, eine mögliche Verschärfung der Vorschriften im Hinblick auf die Nutzung von Informations- und Kommunikationssystemen während der Fahrt zu überprüfen. Und gerade die zunehmende Integration von Nachrichten und Informationen aus dem Internet kann dazu führen, dass der Fahrer trotz Bedienkonzepten wie Spracheingabe in bestimmten Situationen überlastet wird. Im übrigen empfehlen die jüngsten Forschungsergebnisse aus diesem Bereich, dass der Fahrer in einem optimalen Belastungskorridor gehalten werden muss, in dem er von den Fahrzeugsystemen ebenso wenig unterfordert wie überfordert wird – auch Langeweile und Monotonie wirken sich negativ auf Aufmerksamkeit und Leistungsfähigkeit aus. Beim Design von Mensch-Maschine-Schnittstellen, zu dem auch die Nutzung von Spracherkennungssystemen zählt, gilt es deshalb, Fahrsituation und Kontext zu berücksichtigen. Andere Bordsysteme steuern Informationen über Verkehrsaufkommen, Strecke und Fahrsituation bei, die auf Funktionsumfang und Bedienoptionen Einfluss nehmen können. Und auch eine Müdigkeitserkennung und -Warnung, die ihren Input aus Lenkbewegungen oder anderen Bedienvorgängen gewinnt, ist in manchen Fahrzeugen schon Stand der Technik.

Schon heute sind bestimmte Funktionen wie etwa TV-Empfang während der Fahrt gesperrt. In Zukunft könnte das Infotainment-System in feineren Abstufungen auf die Situation sowie den Konzentrations-/Müdigkeitsstatus des Fahrers reagieren und davon abhängig den Zugriff auf einzelne Funktionen regulieren.

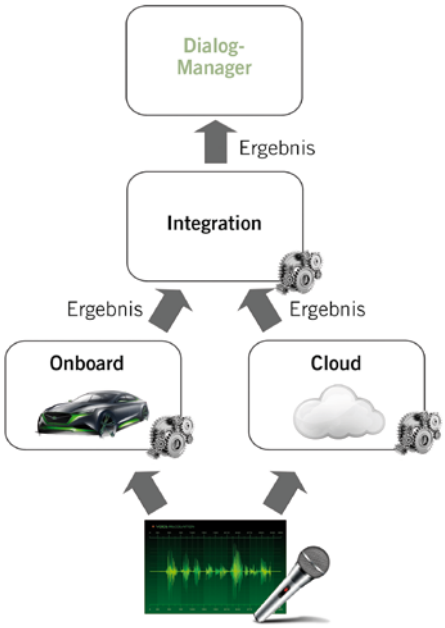
Noch Zukunftsmusik, aber in diesem Zusammenhang ebenfalls vorstellbar, sind biometrische Erkennungsmechanismen. So könnte aus dem Stimmmuster des Nutzers, einem Kamerabild und/oder anderen Einflussgrößen die Aufmerksamkeit beziehungsweise Belastung des Fahrers erkannt werden. Doch dies ist auf jeden Fall noch einige Jahre entfernt. In anderen Bereichen aber könnten biometrische Verfahren im Fahrzeug eher Einzug halten, wie zum Beispiel zur stimm- und/oder kamerabasierten Fahrererkennung.

AUSBLICK – HYBRIDLÖSUNGEN MIT ODER OHNE SMARTPHONE-EINBINDUNG

Aus praktischen Erwägungen wird sich die Spracherkennung nie allein auf die Cloud verlassen, sondern mindestens als Fallback eine lokale Implementation besitzen. Sinnvoll wäre eine Hybrid-Architektur, ❸ die einen breiten Katalog an Funktionen lokal bereitstellt und diese um weitere Inhalte und Funktionen aus der Cloud ergänzt.

Wichtig ist in diesem Zusammenhang auf jeden Fall, dass die Sprachbedienung alle anderen HMI-Anwendungen kennen muss: Alles, was sich per Sprache steuern und bedienen lässt, ob lokal oder in der Cloud, muss in den Dialogsystemen und statistischen Sprachanalysen enthalten sein. Die Komplexität und der Footprint eines Sprachdialogsystems liegen deshalb heute in derselben Größenordnung wie das primäre, grafisch dargestellte und manuell bediente HMI. Nicht umsonst sprechen in der Branche einige von der Sprachsteuerung als „zweitem HMI“.

In diesem Kontext ist auch die Integration von Mobiltelefonen und Smartphones im Fahrzeug zu sehen. Die Kontrolle behält dabei das Onboard-Infotainment-System, es kann aber auf Inhalte und künftig gegebenenfalls auch Funktionen des Smartphones zugreifen. Denkbar sind zum Beispiel die Integration des Smartphone-Displays als Teil des HMI oder die Auslagerung bestimmter Funktionen und Aufgaben des Infotainment-Systems an Smartphone-Apps. So wird das Smartphone möglicherweise in Zukunft neben dem Onboard-System und der Cloud eine dritte Größe, die beim Design von hybriden HMI-Konzepten zu berücksichtigen ist. Die aktuellen Ankündigungen von Apple („CarPlay“) und Google (Android im Auto) weisen genau in diese Richtung. Da auch diese Smartphone-Systeme eigene Spracherkennungen mitbringen, gilt es in diesem Zusammenhang also auch zu lösen, ob und wie Siri, Google Now, S-Voice und Co. mit den bordeigenen Dialogsystemen künftig zusammenarbeiten sollen.



❸ Software-Architektur für hybride Spracherkennung

DOWNLOAD DES BEITRAGS
www.springerprofessional.de/ATZelektronik

READ THE ENGLISH E-MAGAZINE
order your test issue now:
springervieweg-service@springer.com