DESIGN VON SPRACHDIALOGEN FÜR DAS KFZ - STAND DER TECHNIK

Bärbel Jeschke

Harman/Becker, Ulm
BJeschke@harmanbecker.com

Abstract: In der Oberklasse von DaimlerChrysler, BMW und Audi gehört die Sprachbedienung schon seit vielen Jahren zu den Sonder- bzw. Serienausstattungen. Inzwischen ist sie auch in der Mittelklasse weit verbreitet. Nicht nur eine stabile und hohe Erkennrate macht ein Sprachdialogsystem attraktiv für den Kunden, sondern insbesondere dessen einfache Handhabung und Bedienbarkeit.

Dieser Beitrag beschäftigt sich mit dem Design von Sprachdialogen, der Schnittstelle zwischen Mensch und Sprachdialogsystem und gibt einen Überblick über den aktuellen Stand der Entwicklung. Untersuchungen haben ergeben, dass Autofahrer erwarten, dass sie sich einfach ins Auto setzen und die Sprachbedienung bedienen können, ohne vorher in die Bedienungsanleitung geschaut zu haben. Das stellt eine besondere Herausforderung an das Design von Sprachdialogen dar. Diese so genannten Erstbenutzer abzuholen und durch den Sprachdialog zu führen, aber auch erfahrenen Nutzern die Möglichkeit zu geben, über Shortcuts und flache Menüstrukturen das System effektiv zu bedienen.

Anhand von Beispielen, wie der Eingabe einer kompletten Adresse per Sprache oder der Erkennung von dynamischen Texten wie z.B. Radiosender oder SIM-Karten Einträge, wird die Realisierung einer benutzerfreundlichen Sprachbedienung dargestellt. Dafür sind Datenmengen mit über 80 000 Einträgen (Städte Deutschlands) zu verarbeiten, aber auch Dialogstrategien zu entwickeln, um Homophone und Homographen aufzulösen, die ein noch so guter Spracherkenner niemals auseinander halten wird.

1 Einleitung Sprachbedienung im KFZ

Seit nunmehr 11 Jahren gibt es im KFZ sprecherunabhängig Sprachbediensysteme. Angefangen hat es mit einer einfachen Telefonbedienung in der Mercedes S-Klasse. Hinzu kamen bald die Bedienung des Radios, CD-Spielers und Kassettendecks. Seit 2002 kann auch eine Stadt und Straße in das Navigationsgerät per Sprache eingegeben werden.

Möglich wurde diese Entwicklung einerseits durch immer bessere Vorverarbeitung und Erkennung des Sprachsignals, jedoch auch durch immer höhere Rechenleistung und der immer höheren Vernetzung der Komponenten im Fahrzeug.

Eine besondere Rolle beim Einsatz eines Spracherkenners spielt der Sprachdialog – die Schnittstelle zwischen Mensch und Fahrzeug. Ein Sprachdialog besteht immer aus Spracheingabe / Sprachausgabe / Systemreaktion und evtl. Texten auf dem Display. Die Komplexität eines Sprachdialogs ist mit der Komplexität der Geräte und der Spracherkennungstechnologien deutlich angewachsen.

2 Stand der Technik

Aktuell werden im Fahrzeug Sprachbediensysteme angeboten, die es ermöglichen das gesamte Menü des Infotainmentsystems (Radio, Medien, Navigation, Telefon, Klima) zu bedienen. Interessant wird es für die Sprachbedienung allerdings erst, wenn in großen

Datenmengen gesucht werden soll. Wie z. B. über alle Städte Deutschlands, Musiktitel in MP3 Verzeichnissen oder Namen in Adressbüchern. Hier zeigt sich der enorme Vorteil einer Sprachbedienung. Die Sprachbedienung ist hierfür nicht nur schneller, sondern auch deutlich sicherer, da die Suche per Haptik sehr ablenkend ist und in manchen Ländern sogar schon während der Fahrt verboten ist.

Der erste Schritt in der Bedienung von großen Datenmengen per Sprache bestand darin, ein Ziel (Ort/Straße) per Sprache einzugeben.

3 Adresseingabe per Sprache

3.1 Ansatz 1: Buchstabieren

Deutschland hat 80 000 Städtenamen, Berlin 20 000 Straßennamen und USA 200 000 Straßennamen in Kalifornien. Seit 2002 kann in solch großen Datenmengen per Buchstabieren gesucht werden (siehe Tabelle 1).

Schritt	Spracheingabe	Sprachausgabe	Display
1	Ort buchstabieren		
2		Bitte buchstabieren Sie den Ortsnamen	
3	S-T-U-T-T		
4		Bitte wählen Sie eine Zeilennummer	
5			1. Stuttgart 2. Stutengarten 3. Stutenfurt

Tabelle 1 - Beispiel für das Buchstabieren eines Ortes

Da die Erkennrate für Einzelbuchstaben viel zu schlecht ist und es häufige Verwechslungen gibt, wie z.B. A und E, E und I, M und N, braucht man für eine sichere Erkennung mehrere Buchstaben hintereinander. Mit einen s. g. Matcher, der mit den Verwechslungswahrscheinlichkeiten der Einzelbuchstaben arbeitet, kann dann in großen Listen (Städte- / Straßennamen) das richtige Ergebnis gesucht werden (siehe Abbildung 1).

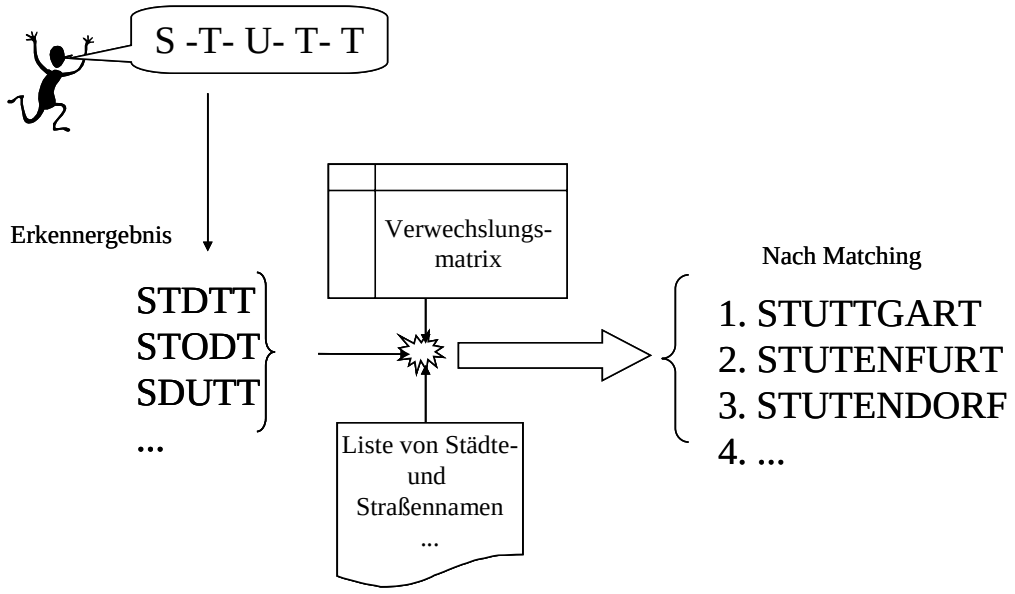

Abbildung 1 – Erkennungsvorgang Buchstabieren

Der Buchstabieransatz hat den großen Vorteil, dass er weitgehend von der Sprache der Listen unabhängig ist. Dafür müssen in der Matrix allerdings die Zeichen aller Sprachen berücksichtigt werden, die abgedeckt werden sollen. Beispiel: Spracherkenner Deutsch, in der Liste sind deutsche oder französische Namen. Das deutsche Alphabet beinhaltete aber kein À, Ê oder Û. Somit muss ein A auf ein À und ein Ê auf ein E abgebildet werden, um die Orte, wie z. B. Saint-Étienne, erkennen zu können. Die Abbildung beruht auf Wahrscheinlichkeiten, wie diese „Sonderzeichen“ in der Ursprungsprache ausgesprochen werden.

Dieser Ansatz hat zusätzlich den Vorteil, dass er nicht besonders speicher- und rechenaufwändig ist. Jedoch nicht in allen Ländern werden Wörter buchstabiert. Wie z. B. im Japanischen, denn dort können Städtenamen nur über Silben eingegeben werden. Neben den Einschränkungen bezüglich der Sprache, ist es für viele Menschen ungewöhnlich, lange Städte- bzw. Straßennamen zu buchstabieren.

3.2 Ansatz 2: Ganzworteingabe

In Tabelle 2 ist ein Beispiel für die Eingabe eines Ortes per Ganzwort dargestellt.

Schritt	Spracheingabe	Sprachausgabe	Display
1	Ort eingeben		
2		Den Ort bitte	
3	Frankfurt		
4		Ist Frankfurt richtig?	
5			1. Frankfurt 2. Frankenberg 3. Franzfurt

Tabelle 2 - Beispiel für die Ganzworteingabe eines Ortes

Für diesen Ansatz müssen für alle Listeneinträge Phonetiken vorhanden sein. Große Navigations-Datenhersteller haben sich die Mühe gemacht, alle Städte und Straßen Deutschlands, große Teile Europas und der USA zu phonetisieren. Diese Phonetiken werden mit dem Kartenmaterial regelmäßig aktualisiert. Für die Erkennung in solch großen Datenmengen wird ein 2-stufiger Ansatz verwendet. Im ersten Schritt werden Phoneme erkannt und aus der kompletten Liste die phonetisch ähnlichsten Einträge rausgesucht. Anschließend wird mit der aufgezeichneten Äußerung ein 2. Erkennschritt auf der Teilliste durchgeführt. (Siehe Abbildung 2).

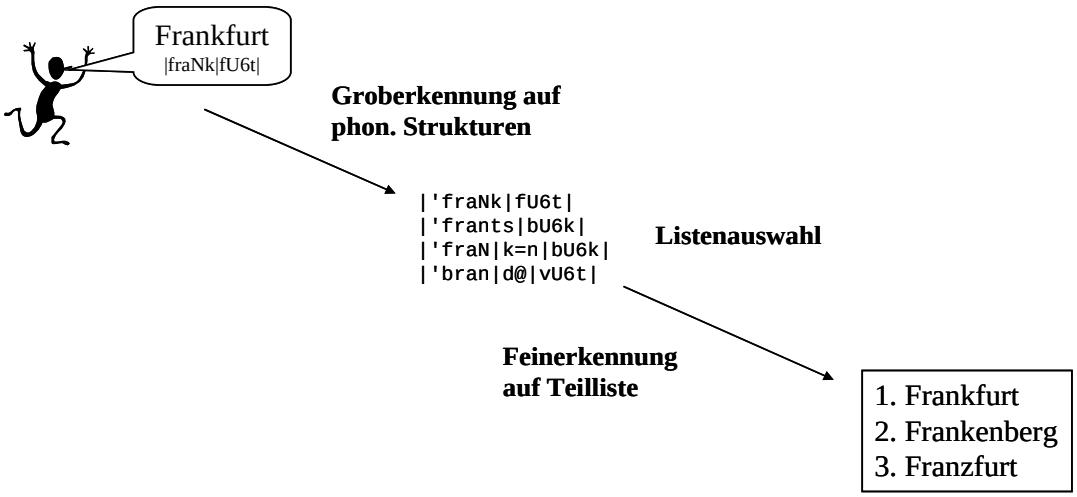

Abbildung 2 – Erkennungsvorgang Ganzworteingabe

Ein Design-Problem hierbei ist, dass der Benutzer nicht immer den vollständigen Namen kennt bzw. sprechen möchte, wie z.B. „Frankfurt am Main“. Nun ist es abhängig vom Datenlieferanten, ob „Frankfurt am Main“ auf oberster Ebene zu finden ist, oder nach der Eingabe von Frankfurt „Frankfurt am Main“ darunter aufgelöst werden kann. In Deutschland findet man drei Frankfurt im Datenmaterial:

Frankfurt / Frankfurt Oder / Frankfurt am Main

Ein weiteres Problem stellen s.g. Mehrdeutigkeiten dar. Es gibt 16 Städte in Deutschland, die mit vollständigem Namen Neustadt heißen. Um eine eindeutige Auswahl treffen zu können, hat der Datenmaterialhersteller Städte in der Nähe, Kreise, etc. den Namen hinzugefügt. Diese müssen in einem 2. Schritt ausgewählt werden. (Siehe Tabelle 3).

Schritt	Spracheingabe	Sprachausgabe	Display
1	Ort eingeben		
2		Den Ort bitte	
3	Neustadt		
4		Bitte wählen Sie den Ort	
5			1. Neustadt/Vogtl. 2. Neustadt (Dosse) 3. Neustadt I.SA. 4. Neustadt
6	Neustadt		
7		Neustadt ist mehrfach vorhanden. Bitte wählen Sie aus	
8			1. Berge 2. Bremen 3. Breuberg 4. ...

Tabelle 3 - Beispiel für die Ganzworteingabe eines mehrdeutigen Ortes

Hier ist man beim Dialog Design sehr vom Datenlieferanten und dessen Datenstrukturen abhängig und muss entsprechende Dialogstrategien entwickeln um den richtigen Ort dem Benutzer anbieten zu können. Die Dialogerfolgsrate hat dabei immer höchste Priorität.

Ein weiteres Problem stellen phonetisch sehr ähnliche bzw. identische Namen dar. Zum Beispiel Aalen und Ahlen, sogenannte Homophone. Es ist nicht möglich, diese Namen eindeutig zu erkennen. Aus diesem Grund hält der Dialog immer die Möglichkeit offen, statt dem Namen der Stadt auch die Zeilennummer zu sprechen. Das ist einer der Hauptgründe, warum die Ergebnislisten immer am Display angezeigt werden, da diese sich auch beim Vorlesen identisch anhören würden. (Siehe Tabelle 4).

Schritt	Spracheingabe	Sprachausgabe	Display
1	Ort eingeben		
2		Den Ort bitte	
3	Aalen		
4		Bitte wählen Sie den Ort	
5			1. Ahlen 2. Aalen 3. Arlen 4. ...
6	Nummer 2		

Tabelle 4 - Beispiel für die Ganzworteingabe eines Ortes mit homophonen Alternativen

Eine weitere Möglichkeit, Homophone aufzulösen besteht darin, Zusatzinformationen vom Benutzer abzufragen. Wie z.B. Postleitzahl, Ort in der Umgebung oder einen Landkreise. Bis auf die Postleitzahl unterstützt der Datenlieferant diese Zusatzinformationen nicht und auch die Postleitzahl hat der Benutzer oft bei der Eingabe einer Adresse nicht zur Hand.

Die Ganzworteingabe ist sprachenabhängig. D. h. mit einem deutschen Spracherkenner können auch nur deutsch phonetisierte Daten erkannt werden. Möchte der Benutzer ins Ausland fahren wird für die Erkennung von z.B. französischen Städten ein multilingualer Spracherkenner benötigt.

Die Erkennung von ausländischen Städte- und Straßennamen wird den nächsten großen Meilenstein in der Entwicklung von Sprachdialogsystemen im Fahrzeug darstellen.

4 Namenssuche in Adressbüchern und Telefonbüchern

Eine bereits verfügbare neue Technologie ist das G2P (Graphem to Phonem). Damit können SIM-Karteneinträge, Adressbücher und Radiosender sprechbar gemacht werden, obwohl keine Phonetiken für diese Orthographien verfügbar sind. Dazu werden über Bluetooth oder andere Schnittstellen die Einträge ausgelesen, in Listen aufbereitet und dazu Phonetiken generiert. Für die Erkennqualität spielt die Qualität der Phonetiken eine sehr große Rolle. Vor allem auf SIM-Karten befinden sich wegen der Zeichenbegrenzung sehr viele Abkürzungen, die so niemals vom Benutzer gesprochen werden. Diese Zeichenbegrenzung macht es dem G2P-Prozess schwer, die richtige Phonetik zu generieren. Mit Hilfe von intelligenten Ausnahmelexika wird versucht, diese aufzulösen und „sprechbar“ zu machen. (Siehe Abbildung 3).

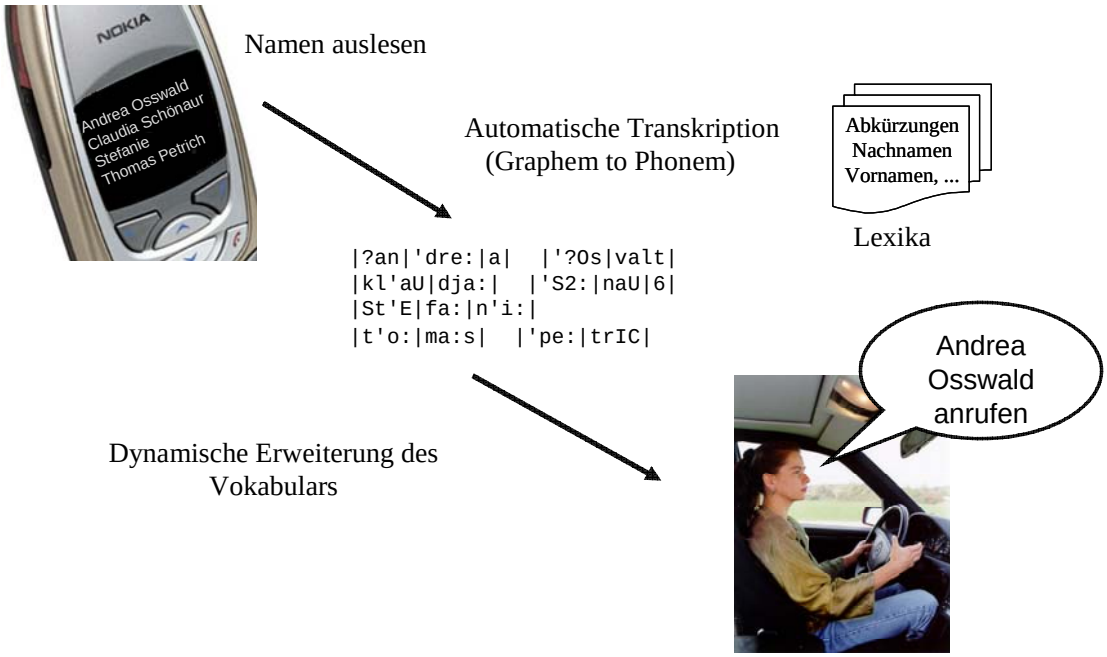

Abbildung 3 – G2P-Verfahren

In Adressbüchern und bei SIM-Karteneinträgen gibt es nicht nur das Problem mit den Homophonen (siehe Tabelle 5), sondern auch mit den Homographen. So können identische Namen mehrfach vorkommen. Diese können vom Benutzer sowohl absichtlich, als auch unabsichtlich eingetragen worden sein.

Schritt	Spracheingabe	Sprachausgabe	Display
1	Anrufen bei Meier		
2		Bitte wählen sie einen Eintag aus der Liste	
3			1. Karl Mayer 2. Karl Maier 3. Ingo Meier 4. ...
4	Ingo Meier		
5		Die Nummer von Ingo Meier wird gewählt	

Tabelle 5 - Beispiel für die Auswahl eines mehrdeutigen Namens aus dem Adressbuch

Das Sprachkommando „Anrufen bei Meier“ kann durch den Sprachdialog aufgelöst werden, indem eine Liste mit allen möglichen Meier angezeigt wird, oder indem der Dialog nach dem Vornamen fragt. Aber auch die Zusatzinformation „Vorname“ löst nicht die ersten beiden Einträge auf. Da bleibt nur wieder die Lösung mit der Anzeige. Als Zusatzinformationen können auch Telefonnummer bzw. Adresse herangezogen werden, um die Homophone und Homographen aufzulösen.

5 Ausblick

Zum Abschluss soll noch darauf hingewiesen werden, dass all diese Ansätze großen Zwängen was Rechenleistung und Speicherbedarf betrifft, unterliegen. Nach wie vor ist für den Endkunden eine sehr schnelle Routenführung wichtiger als eine Sprachbedienung.

Neben der Hochintegration steht aber auch der Weltkonzeptansatz im Vordergrund. So soll eine Sprachbedienung mit geringen Anpassungen auch für die Märkte in Nordamerika oder Asien verwendbar sein. Deshalb müssen alle Ansätze, sei es für die Adresseingabe als auch das gesamte Bedienkonzept über alle Märkte hinweg einzeln beleuchtet werden. Kulturelle Unterschiede spielen hier ebenfalls eine wichtige Rolle. Ein Beispiel hierfür ist die Adresseingabe in Japan. (Siehe Abbildung 4).

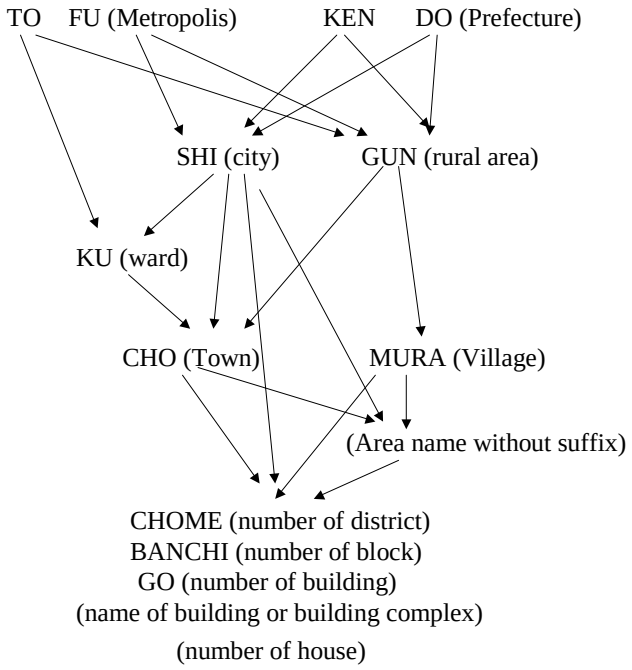

Abbildung 4 – Japanische Adresstruktur [1]

Schon heute kann dort eine Adresse „an einem Stück“ eingegeben werden. Das ist möglich, da die Datenmenge pro Level (Metropolis, City, Ward, ..) gegenüber Europa sehr gering ist und die Level immer hierarchisch aufeinander aufbauen. (47 Prefecturen, darunter 20-30 Citys, ...).

Dies soll einen kurzen Ausblick darstellen, welche kulturellen Besonderheiten bei der Entwicklung von Sprachdialogen über mehrere Kontinente hinweg berücksichtigt werden müssen.

Literatur

[1] http://www.sljfaq.org/afaq/addresses.html