Mobile Informationssysteme mit integrierter Spracherkennung und Sprachsynthese

Jens Estelmann, Uwe Koloska

voiceINTERconnect

estelmann@voiceinterconnect.de

Zusammenfassung: In den letzten 20 Jahren wurden große Fortschritte in den Bereichen Rechentechnik, Spracherkennung und Sprachsynthese erzielt. Es ist dadurch möglich geworden, Spracherkennung und Sprachsynthese auf mobilen Geräten wie PDAs zu implementieren. Hieraus ergeben sich völlig neue Möglichkeiten für den Applikationsentwurf. In diesem Beitrag wird erst auf den allgemeinen Aufbau eines mobilen Informationssystems eingegangen, um dann in den folgenden Abschnitten die einzelnen Komponenten näher zu beleuchten. Abgeschlossen wird dieser Bericht mit zwei exemplarischen Beispielanwendungen, die die Funktionsweise und Vorteile eines mobilen Informationssystems mit Spracherkennung und Sprachsynthese verdeutlichen sollen.

1 Einleitung

Heutige mobile Informationssysteme werden in der Regel über wenig intuitive Eingabemedien wie Tastatur oder Touchscreen bedient. Für diese Art der Mensch-Maschine-Kommunikation werden die Hände und die volle Aufmerksamkeit der Augen benötigt. In vielen Einsatzgebieten, bei denen Augen und Hände schon anderweitig gebunden sind, stellt dies einen bedeutenden Nachteil dar. Durch die Integration von Spracherkennung und Sprachsynthese in mobile Informationssysteme lässt sich dieses Problem lösen. Somit erschließen sich viele neue Anwendungsfelder, beispielsweise in den Bereichen Telefonie, Automotive, im industriellen Bereich oder für touristische Anwendungen. Die Einbindung der Spracherkennung und Sprachsynthese in die mobilen Anwendungen erfolgt mittels eines Dialogsystems. Das Dialogsystem ermöglicht dabei eine parallele Nutzung unterschiedlicher Kommunikationskanäle wie z.B. Sprache, Text, taktile Eingaben und Bildinformation. Durch die intelligente Verknüpfung unterschiedlicher Ein- und Ausgabemedien können wesentlich intuitivere und flexiblere Anwendungen geschaffen werden, was die Ergonomie und Akzeptanz mobiler Informationssysteme im Alltag erhöht.

2 Die Komponenten eines mobilen Informationssystems

Im folgenden Abschnitt soll der prinzipielle Aufbau eines mobilen Informationssystems betrachtet werden und auf die Erweiterungen eingegangen werden, die nötig sind um Spracherkennung und Sprachsynthese sinnvoll in ein solches System zu integrieren. Mobile Informationsanwendungen können in folgende vier Teile unterteilt werden:

1. Dialog,
2. Wissensbasis,

3. Eingabemodul,
4. und Ausgabemodul.

Bei der Wissensbasis handelt es sich im Normalfall um eine Datenbank, in der die für die Anwendung nötigen Nutzerinformationen abgelegt sind. Beispielsweise würde bei einem mobilen Lexikon die Wissensbasis aus einer Datenbank mit den Lexikoneinträgen bestehen. Es ist auch möglich die Wissensbasis nicht lokal auf dem mobilen Endgerät zu speichern, sondern sie auf einen externen Server abzulegen. So zum Beispiel bei Navigationssystemen die zusätzlich Stauinformationen über einen externen Server abfragen.

Die Aufgabe des Dialoges ist es, alle Komponenten des mobilen Informationssystems zu koordinieren und zu steuern. Bei dem Dialog handelt es sich um einen endlichen Zustandsautomaten, dessen Zustand über User- bzw. Systemeingaben verändert werden kann. Abhängig vom Zustand des endlichen Zustandsautomaten werden Informationen über die Ausgabeschnittstelle ausgegeben und Eingabeoptionen aktiviert bzw. deaktiviert.

Das Eingabemedium ist bei klassischen mobilen Informationssystemen ein Touchscreen oder einer Tastatur. Als Ausgabmedium wird in der Regel ein Display verwendet. Bei den in dieser Abhandlung betrachteten mobilen Informationssystemen werden zusätzlich Spracheingabe bzw. Sprachausgabe als Ein-/ bzw. Ausgabemedien benutzt. Für die Spracheingabe wird ein Spracherkenner verwendet. Auf die verschiedenen Typen von Spracherkenner und ihrer spezifischen Vor- bzw. Nachteile wird im Abschnitt 2.2 näher eingegangen. Die Sprachausgabe wird mittels Sprachsynthese realisiert. Auch hier gibt es unterschiedliche Implementationsvarianten, die sich in punkto Sprachqualität und Ressourcenbedarf unterscheiden.

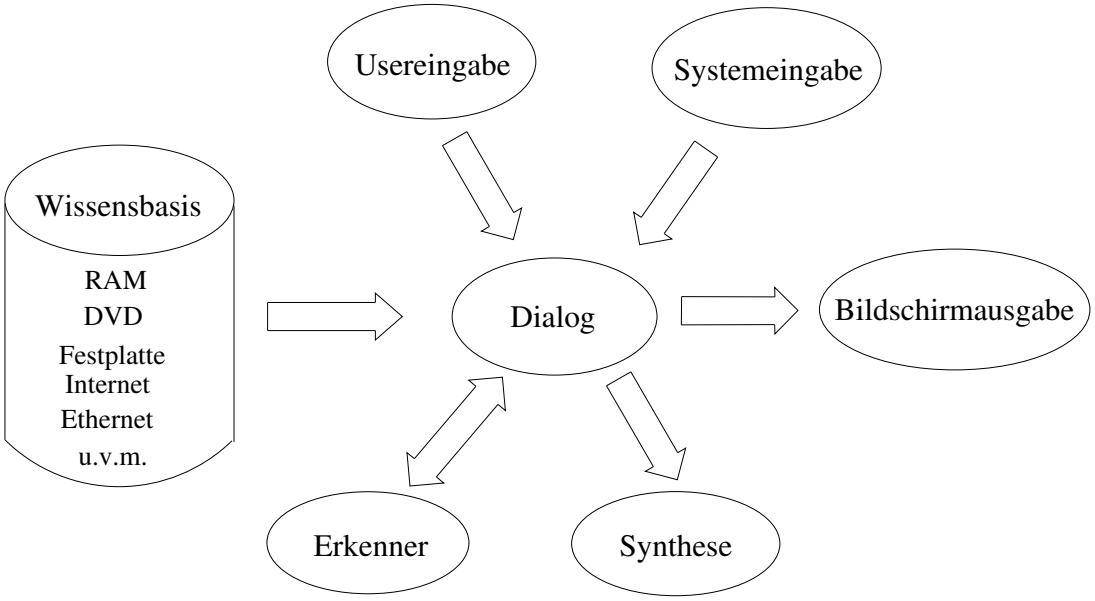

Abbildung 1 - Komponeten eines mobilen Informationssytems

2.1 Dialog

Wie bereits im letzten Abschnitt erwähnt, ist die Aufgabe des Dialoges die Steuerung des Programmflusses und aller Komponenten des mobilen Informationssystems. Bei dem Dialog handelt es sich um einen endlichen Zustandsautomaten. Endliche Zustandsautomaten sind gerichteten Graphen, bei dem jeder Knoten einem Zustand entspricht und jede Kante

einem Zustandsübergang. Endliche Zustandsautomaten lassen sich als Mealy oder Moore Automat realisieren. Bei Moore Automaten hängt die Ausgabe des Systems nur vom vorhergehenden Zustand ab, während bei einem Mealy Automaten die Ausgabe von vorhergehenden und vom aktuellen Zustand abhängt. Da Moore Automaten in den meisten Fällen komplizierter zu realisieren sind als Mealy Automaten, werden in der Regel Mealy Automaten zur Repräsentation des Dialoges verwendet. In Abbildung 2 ist beispielhaft ein einfacher Dialog als Mealy Automat dargestellt. Der Startzustand des Mealy Automa-

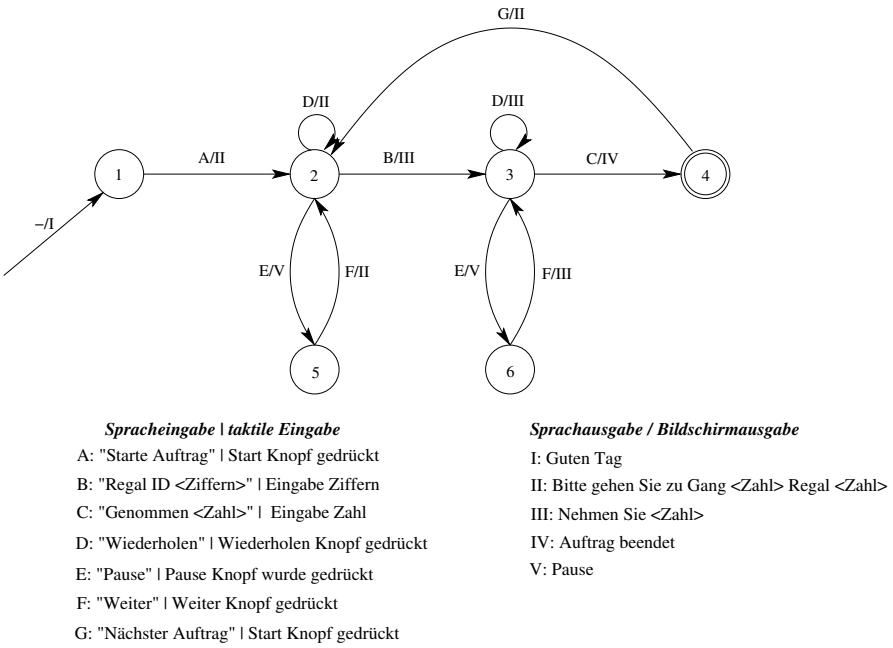

Abbildung 2 - Mealy Automat für einen einfachen Dialog

ten ist mit einem eingehenden Pfeil markiert und der Endzustand mit einem Doppelkreis. Die Symboltupel über jedem Zustandsübergang repräsentieren die Transitionsbedingung (links vom Schrägstrich) und die Ausgabefunktionen (rechts vom Schrägstrich) des jeweiligen Zustandsübergangs. Die Aufgabe der Ausgabefunktion ist nicht nur die Bildschirm- und Sprachausgabe, sondern auch die allgemeine Aktualisierung des Systemzustandes. Dazu gehören Aktivierung bzw. Deaktivierung des Mikrophons, Aktualisierung des Erkennerwortschatzes, Aktivierung bzw. Deaktivierung verschiedener GUI Elemente und vieles mehr.

Definiert wird der Dialog mittels einer Dialogdefinitionsdatei, welche als Textfile abgelegt werden kann. Zur vollständigen Definition des dem Dialog hinterlegtem Mealy Automaten werden eine Kantenliste, der Index des Startknotens und die Indizes der Endknoten benötigt. Die Kantenliste ordnet jeder Kante ihren Quellknoten, ihren Zielknoten, die Transitionbedingung sowie die Ausgabefunktion zu. Ein Beispiel für eine solche Dialog Definition ist im Listing 1 dargestellt.

2.2 Erkenner

Die Aufgabe eines Spracherkenners in einem mobilen Informationssystem ist es, die einem akustischen Signal innenwohnende Information dem System in maschinenlesbarer Form zur Verfügung zu stellen. Um dieses zu erreichen, muss das akustische Signal in einem ersten Schritt in ein digitales Signal gewandelt werden. In einem zweiten Schritt,

Listing 1 - Dialog-Definitionsdatei für den in Abbildung 2 dargestellten Dialog

```
<startNode>: 1
<endNodes>: 6
% arc list
% <arc x>: destNode,srcNode,trans. cond.,out. func.
<arc 1> : 2, 1, cond_A, out_II
<arc 2> : 3, 2, cond_B, out_III
<arc 3> : 4, 3, cond_C, out_IV
<arc 4> : 2, 2, cond_D, out_II
<arc 5> : 3, 3, cond_D, out_III
<arc 6> : 5, 2, cond_E, out_V
<arc 7> : 2, 5, cond_F, out_II
<arc 8> : 6, 3, cond_E, out_V
<arc 9> : 3, 6, cond_F, out_III
<arc 10>: 2, 4, cond_G, out_II
```

der Merkmalsextraktion werden aus dem digitalen Sprachsignal die Merkmale extrahiert, die für eine spätere Klassifikation einen maximalen Informationsbeitrag bringen. Bei der Merkmalsextraktion handelt es sich um eine verlustbehaftete Kompression des Sprachsignals. Erst durch die Merkmalsextraktion ist eine Klassifikation des Sprachsignals mit vertretbaren Rechenaufwand möglich. Im letzten Abschnitt des Erkennungsprozesses, der Klassifikation, wird die eingehende Merkmalsvektorfolge mit allen Referenzmodellen des Erkenners verglichen und das Referenzmodell ausgewählt, das am besten zur Merkmalsvektorfolge passt. Die Referenzmodelle müssen vor dem Betrieb des Erkenners in einer sogenannten Trainingsphase angelegt worden sein. Für den Vergleich der Referenzvektoren

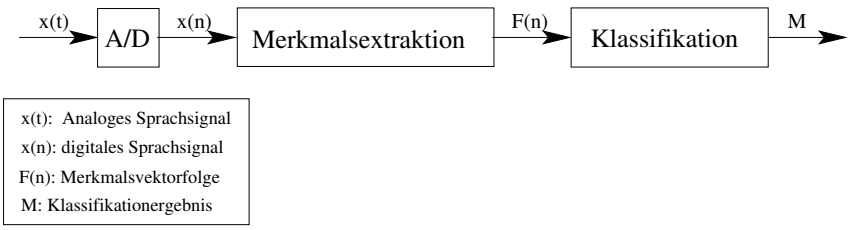

Abbildung 3 - Prinzipieller Aufbau eines Spracherkennungssystems

mit den Merkmalvektoren werden effiziente Algorithmen benötigt, um den Rechenaufwand in Grenzen zu halten. In den vergangen 40 Jahren haben sich in der Spracherkennung zwei Verfahren zur Klassifikation einer Merkmalsvektorfolge durchgesetzt und zwar:

- Klassifikation mittels des *Dynamik Timewarping* Algorithmus (DTW)
- und Klassifikation auf Basis von *Hidden Markov Modellen* (HMM).

In den folgenden zwei Abschnitten werden diese Verfahren kurz vorgestellt. Vertiefende Information zu diesen Algorithmen finden sich in [1, 2].

2.2.1 Der *Dynamic Timewarping (DTW)* Algorithmus

Bei Erkennern, die auf dem DTW Algorithmus basieren, liegen die Referenzmodelle als diskrete Vektorfolgen vor. Die Schwierigkeit bei dem Vergleich der Merkmalsvektorfolge mit den Referenzvektorfolgen liegt nun darin, dass die Referenzvektorfolgen sich bezüglich ihres zeitlichen Aufbaus von der Merkmalsvektorfolge unterscheiden können. Der DTW

Algorithmus ermöglicht die Merkmalsvektorfolge auf die Referenzvektorfolge so abzubilden, dass die Distanz der beiden Vektorfolgen zueinander minimiert wird. Um den Rechenaufwand des DTW Algorithmus in Grenzen zu halten, wird das Prinzip der dynamischen Programmierung (DP) für den Vergleich der Vektorfolgen angewendet.

Der DTW Algorithmus eignet sich besonders für Anwendungen, bei denen nur kleine begrenzte Wortschätze erkannt werden müssen. In diesen Fällen arbeitet der DTW Erkenner sehr ressourcenschonend was Speicherbedarf und Rechenzeit betrifft. Ein großer Vorteil des DTW Algorithmus ist, dass er sich sehr einfach im Fixpointformat realisieren lässt. Somit können Erkenner, die auf dem DTW-Algorithmus basieren, mit sehr wenig Aufwand auf kostengünstige Mikroprozessoren und DSPs portiert werden. Des weiteren bieten DTW-Erkenner den Vorteil, dass sie mit sehr wenig Aufwand trainiert werden können. Dies ermöglicht beispielsweise ein nachträgliches anlernen von Wörten durch den Nutzer.

2.2.2 Erkennung mit *Hidden Markov Modellen* (HMM)

Bei der Erkennung mit Hidden Markov Modellen wird versucht, ein Sprachsignal mittels eines statistischen Modelles zu beschreiben. Bei den verwendeten statistischen Modellen handelt es sich um doppelt verschachtelte stochastische Prozesse. Die erste Ebene dieser Prozesse besteht aus einer Markov Kette, deren Zustände mit diskreten Zeitschritten durchlaufen werden. Beim Durchlaufen generiert ein jeder Zustand ein Ausgabesymbol $\mathbf{O_t}$ gemäß einer dem Zustand angehängten Wahrscheinlichkeitsdichtefunktion. Da für einen Beobachter nur die entstehende Sequenz der Ausgabesymbole $\mathbf{O_t}$ sichtbar ist und nicht die durchlaufende Zustandsfolge, werden diese Modelle als *Hidden* Markov Modelle (HMM) bezeichnet.

Für das Training der HMMs werden sehr große Datenmengen benötigt, was das Training sehr aufwendig macht. Dies stellt neben dem erhöhten Rechenaufwand gegenüber DTW-Erkennern den Hauptnachteil der HMM-Erkenner dar. Dafür weisen HMM-Erkenner eine wesentlich bessere Sprecherunabhängigkeit als DTW-Erkenner auf. Außerdem haben HMM Erkenner besonders bei der Erkennung von kurzen akustischen Ereignissen, wie zum Beispiel Phonemrealisierungen, einen Vorteil vor Erkennern, die auf Basis des DTW-Algorithmus klassifizieren. Dies kann damit erklärt werden, dass HMMs besser die Varianz unterschiedlicher Wortrealisierungen erfassen können, was besonders bei kurzen Worten wichtig ist.

2.3 Sprachsynthese

Die Aufgabe der Sprachsynthese ist die künstliche Erzeugung menschlicher Sprache. Im wissenschaftlichen Bereich haben sich drei unterschiedliche Methoden zur Sprachsynthese durchgesetzt und zwar,

- die artikulatorische Synthese,
- die Formantsynthese
- und die konkatenative Synthese.

Praktischen Sprachqualitätsansprüche kann derzeit nur die konkatenative Sprachsynthese genügen. Dies hat zur Folge, dass parametrische Synthesesysteme nur in der Forschung und in Laborsystemen zum Einsatz kommen.

Bei der konkatenativen Synthese wird eine Äußerung erzeugt, indem menschliche Sprachsamples aneinandergehängt werden. An den Verkettungsstellen wird das sich ergebende Signal digital nachbearbeitet, um wahrnehmbare Diskontinuitäten zu minimieren. Abhängig von der Länge der verwendeten Sprachsamples können unterschiedliche Varianten der konkatenativen Synthese unterschieden werden.

Vollsynthese Bei der Vollsynthese werden als Bausteine Bi- oder Triphone verwendet. Diese Systeme bieten ein Maximum an Flexibilität, da es mit ihnen möglich ist, beliebige Texte in Sprache zu wandeln. Vollsynthesesysteme haben einen relativ hohen Speicherbedarf, der aber unabhängig vom eingesetzten Wortschatz ist. Synthesesysteme mit großen Wortschätzen lassen sich auf mobilen Informationssystemen effektiv nur mit einer Vollsynthese realisieren. Nachteil der Vollsynthese ist, die im Vergleich zu den anderen Synthesevarianten, schlechtere Sprachqualität, sowie ein realtiv hoher, dafür jedoch einmaliger Aufwand bei der Erstellung einer neuen Sprache oder Stimme.

Wortsynthese Die Wortsynthese benutzt als Bausteine für die Synthese ganze Worte. Wortsynthesesysteme weisen eine bessere Sprachqualität als Vollsynthesesysteme auf. Der benötigte Speicherbedarf dieser Systeme steigt allerdings proportional mit der Größe des verwendeten Wortschatzes. Für Synthesesysteme mit kleinem Wortschatz können Wortsynthesesysteme einen kleineren Speicherbedarf bei gleichzeitig höherer Sprachqualität als Vollsynthesesysteme aufweisen.

Sprachprompting Die weitaus beste Sprachqualität wird beim Sprachprompting erzielt. Streng genommen handelt es sich hier um keine Synthese, da nur vorher aufgenommene Sprachprompts abspielt werden. Die Qualität hängt ausschließlich von der Qualität der Aufnahme und den eventuell verwendeten Kompressionsverfahren ab. Das Sprachprompting ist die unflexibelste und speicherintensivste Variante der Spracherzeugung.

Weitere Informationen zur konkatenativen Sprachsynthese können in [4, 3] nachgeschlagen werden.

3 Beispielapplikation

3.1 Mobiles Sprachkommissionierungssystem Volog

Bei Volog handelt es sich um ein System zur Verwaltung von Lagerbeständen. Das System besteht aus einem PDA, der mittels WLAN Verbindung zu einem zentralen Server hat, auf dem eine Lagerverwaltungssoftware läuft. Der PDA ist mit einem Spracherkennungssystem und Sprachausgabe ausgerüstet.

Der PDA wird von dem Kommissionierer in der Tasche oder am Gürtel getragen. Die Kommunikation zwischen PDA und Lagerarbeiter geschiet über ein Headset, das wahlweise per Kabel oder Bluetooth an den PDA angebunden ist. Die Kommunikation zwischen PDA und Kommissionierer findet ausschließlich über Sprache statt. Vereinfacht dargestellt unterhält sich der Kommissionierer mit dem Spracherkenner. Der Spracherkenner interpretiert die Äußerungen des Kommissionierers und sendet die Daten an den zentralen Server. In anderer Richtung sendet der Server Anweisungen an den PDA, der diese interpretiert und mittels Sprachausgabe an den Kommissionierer weitergibt.

Am Anfang seiner Schicht meldet sich der Kommissionierer mittels seiner ID beim Server an und tritt daraufhin in einen Dialog mit dem Server. So erhält der Kommisionierer

Listing 2 - Beispiel eines Dialoges zwischen Server und Anwender des Volog-Systems

```
Server: "Guten Tag Herr Mustermann es sind zwei Auftraege vorhanden"
Kommissionierer: "Starte ersten Auftrag"
Server: "Auftrag 1. Gang 3 Fach 14"
Kommissionierer: "Platz ID 1542"
Server: "Nimm 10"
Kommissionierer: "Genommen 10"
Server: "Auftrag Beendet. Es ist noch ein Auftrag vorhanden."
```

Aufträge vom Server, die er nach erfolgreicher Abarbeitung quittieren muss. Weiterhin kann der Kommissionierer auch Anfragen an den Server stellen wie z.B. „*wiederhole letzten Auftrag*“ oder „*Pause*“. In Listing 2 ist ein kurzes Beispiel für einen Dialog zwischen Server und Kommissionierer dargestellt.

Volog erlaubt es auf dem zentralen Server Sprachprofile von verschiedenen Sprechern abzulegen. Durch Anwendung des passenden Sprecherprofils kann die Erkennungsgenauigkeit, besonders bei Sprechern mit Akzent, deutlich erhöht werden. Für die Erstellung eines Sprachprofils benötigt der betreffende Sprecher circa 10 bis 15 Minuten.

Ein großer Vorteil von Volog besteht darin, dass der Kommissionierer das System vollständig mit Sprache bedienen kann und somit beide Hände für anstehende Arbeiten zur Verfügung hat. Weiterhin ist Volog sehr intuitiv zu bedienen, was die Einarbeitungsphase verkürzt.

3.2 Sprachgesteuerte Messdatenerfassung im Bereich der Forstwirtschaft

Als zweites Beispiel soll hier eine sprachgesteuerte Messdatenerfassung für den Bereich der Forstverwaltung dienen. In der Forstwirtschaft müssen Messdaten in schlecht zugänglichen Aussenarealen mit wechselnden Witterungsbedingungen aufgenommen werden. Aus diesem Grund bietet sich in diesem Bereich die Verwendung von mobilen Informationssystemen auf Basis von PDAs an.

Die Forstanwendung besteht aus mehreren Eingabemasken in denen Baumbestand, Baumschäden und ähnliche Informationen eingegeben werden können. Neben der klassischen Eingabe über Touchscreen ist eine Eingabe mittels Sprache möglich. Hierfür wird erst der Name des zu beschreibenden Maskenfeldes gesprochen und anschließend der Eingabewert / Eingabetext. Für die Sprachkommunikation wird ein Headset verwendet. Als Spracherkennungsmodul wird ein Phonemerkenner, auf Basis von HMM, benutzt. Dadurch kann der Wortschatz bequem durch Erweiterung des Lexikon angepasst werden. Im Gegensatz zu Volog, bei dem es sich um eine dialoggetriebene Anwendung handelt, ist die Forstanwendung ereignisgesteuert. Das bedeutet, dass das Verhalten des Systems stark von der Eingabe des Nutzers abhängt und weniger vom aktuellen Systemzustand.

Für die Sprachausgabe werden Sprachprompts verwendet. Dies ist möglich, da die Forstanwendung einen eng begrenzten Wortschatz aufweist. Nach Beendigung der Messdatenaufnahme im jeweiligen Forst werden die Daten auf einen zentralen Rechner übertragen.

4 Ausblick

Durch die Integration von Spracherkennung und Sprachsynhese in mobilen Informationssystemen lassen sich intuitivere und leichter zu bedienende Anwendungen schaffen. Desweiteren bietet sich Spracheingabe besonders in Bereichen an, in denen Hände und Augen nicht für die Dateneingabe zur Verfügung stehen. So können beispielsweise mobile Informationssysteme Behinderte in vielen Alltagssituationen unterstützen. Auch im Automobilbereich bieten sich viele Anwendungsmöglichkeiten für mobile Informationssysteme an. So sind Navigationssysteme mit Spracheingabe und Sprachausgabe schon heute weit verbreitet. Die Steuerung von Klimaanlage, Hifi-Anlage und anderer Komponenten im Auto wären ebenfalls sinnvoll, und dient nicht nur der Bequemlichkeit sondern auch der Sicherheit im Straßenverkehr.

Durch steigende Rechenleistung und weitere Verbesserung der Erkennungs- und Synthesealgorithmen werden mobile Informationssysteme immer immer schneller, genauer und günstiger. Dadurch werden diese Systeme vermehrt Einzug in unseren Alltag halten. Bei richtigem Entwurf der Anwendungen, kann ein hoher Ergonomie- und Produktivitätsgewinn beim Einsatz solcher Systeme im Industrie und Landwirtschaft erzielt werden. Praktische Erfahrungen liegen bei der *voiceINTERconnect* für die Einsatzgebiete Kommissionierung, Landwirtschaft, Forstwirtschaft sowie im Bereich Tourismus vor. [1]

Literatur

[1] L.R. Rabiner, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[2] T. Rudolph, *Evolutionäre Optimierung schneller Worterkenner*, w.e.b. Universitätsverlag, 1999.

[3] M. Holzapfel, *Konkatenative Sprachsynthese mit großen Datenbanken*, w.e.b. Universitätsverlag, 2000.

[4] D. Hirschfeld, *Universelle Datenaufbereitung und Bausteinauswahl für die konkatenative Sprachsynthese*, w.e.b. Universitätsverlag, 2001.

[1] Das Vorhaben wird im Rahmen der Technologieföderung mit Mitteln der Europäischen Fonds für regionale Entwicklung (EFRE) 2000 - 2006 und mit Mitteln des Freistaates Sachsen gefördert.