EINE MOBILE INTERAKTIONSPLATTFORM FÜR MULTIMODALE INTERAKTION

Giancarlo Boi[1], Klaus Kasper[3], Lubos Krejsa[1], Kerstin Reichel[2], Herbert Reininger[1] und Bernd Schindler[2]

[1]ATIP GmbH, [2]MediaInterface Dresden GmbH, [3]Fachhochschule Darmstadt
Herbert.Reininger@atip.de

Abstract: Es wird eine mobile Interaktionsplattform für multimodale Dialoge beschrieben. Dies ist eine Softwareumgebung, die auf einem mobilen Endgerät vom Typ Pocket PC die gleichzeitige Kommunikation per Sprache und Stift/Grafik mit einer Anwendung ermöglicht. Wesentlicher Aspekt der Lösung ist eine vollständige Integration sowohl von Dialogsteuerung als auch von Spracheingabe und Sprach/Audioausgabe. Zur Steuerung des Dialogablaufs wurde eine einfache Dialogbeschreibungssprache definiert, deren Tags während der Dialogausführung interpretiert werden. Eine einfache Anbindung von Applikationsteilen wird durch Kapselung der Kommunikationsmodalität erreicht.

1 Einleitung

Die Möglichkeit mit der Maschine reden zu können, ist zunehmend eine Voraussetzung für die Realisierung von einfachen und vor allem intuitiven Benutzerschnittstellen. Zahlreich sind bereits heute die Applikationen, die mit Sprache gesteuert werden können und immer besser wird die Technik auf der sie basieren. Applikationen mit automatisierten Sprachdialogen kommen heutzutage zunehmend bei telefonischen Auskunfts- und Bestellsystemen zum Einsatz.

In bestimmten Situationen ist die natürliche Sprache wiederum nicht als Kommunikationsmittel geeignet bzw. ausreichend. Dies fällt in der alltäglichen Kommunikation insbesondere dann auf, wenn verbale Erläuterungen mit Gesten, Mimik oder sonstigen grafischen Hilfsmitteln, wie z.B. Skizzen, unterstützt werden. Aus diesen Gründen ist es notwendig und sinnvoll, komplementär oder ergänzend zur Sprache, weitere Kommunikationsformen auch für den Mensch-Maschine-Dialog einzusetzen. Multimodale Dialoge sind solche, die verschiedene Eingabe-Modalitäten unterstützen, z.B. Sprache, Text, Zeigegesten, etc., und mehrere Medien zur Ausgabe, z.B. Audio, Grafik, etc., verwenden [1].

In diesem Beitrag wird eine mobile Interaktionsplattform für multimodale Dialoge beschrieben [2]. Damit wird hier eine Softwareumgebung bezeichnet, die auf einem mobilen Endgerät (PDA) die gleichzeitige Kommunikation per Sprache und Stift/Grafik mit einer Anwendung in einem multimodalen Dialog ermöglicht. Wesentlicher Aspekt der hier vorgestellten Interaktionsplattform ist eine vollständige Integration sowohl von Dialogsteuerung als auch von Sprachein- und -ausgabe, so dass diese Realisierung als „Embedded Lösung“ bezeichnet werden kann. Zunächst werden die strukturelle Architektur der Interaktionsplattform vorgestellt und anschließend anhand eines Beispieldialogs die Prozessabläufe erläutert.

2 Architektur der multimodalen Interaktionsplattform

In Abbildung 1 ist die Struktur der Plattform schematisch dargestellt. Konzeptionell basiert die mobile Plattform auf einer für multimodale Dialoge erweiterten Sprachdialogplattform, die in [3] dargestellt wurde. Die Plattform sorgt für ein Framework, in dem die Ausgabe von Prompts, die Erkennung der Sprache und der Ablauf des Dialogs gesteuert wird. Der multimodale Dialog ist in einer eigens definierten Dialogsbeschreibungssprache (MMXML) codiert. Ein Dokument mit dem Dialog wird vom MMXML-Interpreter gelesen und bei der Dialogausführung interpretiert. Hierbei werden die Statements des Dialogs in Nachrichten an die Devices (ASR, TTS) und Frontends (Grafik Engines, Browser) umgesetzt.

Der Dialog Manager führt dann die entsprechenden Aktionen aus und verwaltet die interne Kommunikation der Plattform. Er empfängt die Erkennungsergebnisse vom Multimodal Input Collector und leitet diese an den MMXML-Interpreter weiter. Der Interpreter kommuniziert dann dem Dialog Manager, welche Aktionen ausgeführt werden müssen. Audio-Prompts werden dem Input/Output-Channel übertragen, während Outputs zu dem visuellen Frontend per TCP/IP kommuniziert werden.

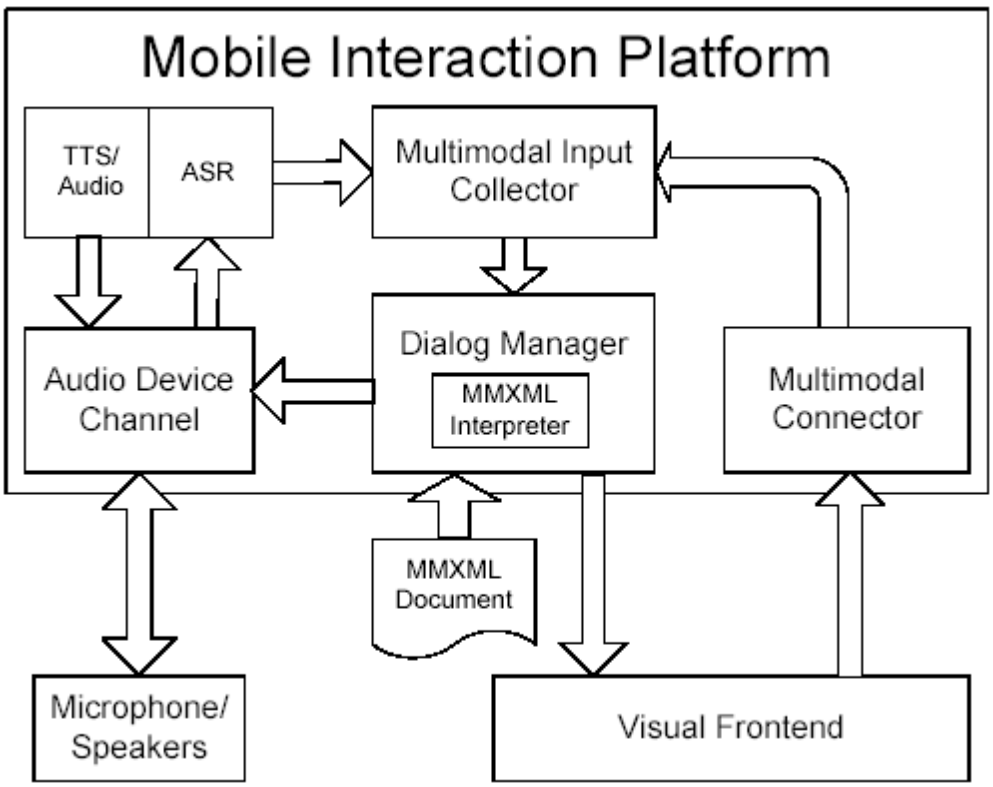

Abbildung 1 Architektur der mobilen Interaktionsplattform

Der Multimodal Input Collector hat die Aufgabe, sowohl Spracherkennungsergebnisse als auch Stifteingaben zu kombinieren und dem Dialog Manager zu kommunizieren. Hierbei werden die Eingaben in sog. Slots gefüllt, die als Variablen im Dialog referenziert werden. Über im Dialog einstellbare Zeitparameter kann erreicht werden, dass z.B. Spracheingabe und Stifteingabe, die innerhalb einer vorgegebenen Zeitspanne erfolgen, sich zu einer vollständigen Dateneingabe ergänzen. Alternativ können diese Modalitäten konkurrierend

verwendet werden, so dass ein Benutzer kontextabhängig die adäquate Modalität verwenden kann.

Der Multimodal Connector empfängt die Stifteingabe direkt von dem grafischen Applikationsteil und leitet diese an den Multimodal Input Collector weiter. Die Kommunikation zwischen Applikation und Multimodal Connector erfolgt via Sockets und versendet Daten über das TCP/IP Protokoll. Die Kommunikation zwischen dem Multimodal Connector und dem Multimodal Input Collector erfolgt mittels eines Shared-Memory Segments. Dies ist ein Speicherbereich, auf den mehrere, unabhängige Prozesse zugreifen können. Damit ist es möglich sehr effizient Nachrichten zwischen verschiedenen Prozesse auszutauschen.

Für die Anbindung von Applikationen an die Interaktionsplattform wird davon ausgegangen, dass die zur Sprache alternative Kommunikationsmodalität in einem separaten Teil der Applikation gekapselt ist. In einem derartigen Applikationsteil sind für den Dialogablauf relevante Objekte definiert. Diese können Aktionen ausführen, wie z.B. „anzeigen“, oder gewisse Zustände annehmen, wie z.B. Zustand „selektiert“. Das erlaubt die Integration einer beliebigen Eingabemodalität, ohne dass spezifische Änderungen in der Plattform nötig werden. Im realisierten Prototyp wird Stifteingabe und grafische Ausgabe als Kommunikationsmodalität parallel zur Sprachkommunikation verwendet.

3 Embedded Spracherkenner

Der im Rahmen dieser prototypischen Implementierung verwendete Spracherkenner (ASR) ist der Very Smart Recognizer (VSR) 4.0 der Firma Siemens AG. der durch die MediaInterface Dresden GmbH bereitgestellt wurde. VSR ist eine Spracherkennungssoftware, welche speziell für mobile Endgeräte (Handy, PDA, Smart Device, etc.) entwickelt wurde. Die Hauptmerkmale des VSR sind:

- sprecherunabhängige phonembasierte Spracherkennung
- Hidden-Markov-Model Technologie
- kontinuierliche Erkennung
- Grammatik Support
- Word Spotting
- Support für hohe Störgeräusche (Spracherkennung im Auto)
- optimierte akustische Modelle für kontinuierliche Zifferneingabe
- optimierte akustische Modelle für kontinuierliches Buchstabieren
- Hinzfügen neuer Worte in geschriebener Form (TypeIn)
- Hinzfügen neuer Worte in gesprochener Form (SayIn)
- VGEditTool für die grafische Erstellung von Wortschätzen und Grammatiken
- verschiedene Programmierinterface (APIs) auf einem unterschiedlichem Abstraktionsniveau
- verfügbar in den Sprachen (Deutsch, Englisch(UK, US), Französisch, Italienisch, Holländisch, Polnisch, Spanisch)
- portierbarer Source Code in ANSI-C

Für die Integration des VSR in die Interaktionsplattform wurde das COM-Modul VSRXCOM verwendet. Das VSRXCOM-Interface kapselt die low-level Funktionalität des VSR, womit

eine einfache und schnelle Integration möglich wurde. Mit den entsprechenden Grammatiktools wurden Wortschätze und Grammatiken in Deutsch und Englisch(UK) definiert. Aus diesen wurden kompilierte Packages erstellt, die Grammatik-Regeln enthalten, die innerhalb des Dialogs aktiviert oder deaktiviert werden.

4 Mobiles Assistenzsystem

Zur Demonstration der Funktionsweise und Prozessabläufe der mobilen Interaktionsplattform wurde ein multimodal gesteuerter Guide (FH-Info) der Fachhochschule Frankfurt realisiert. Die Applikation - dargestellt in Abbildung 2 - zeigt den Lageplan der Fachhochschule Frankfurt am Main und erlaubt dem User Informationen über die verschiedenen Gebäude oder Personen zu erfahren. Der User kann mittels Stift ein oder mehrere Gebäude auswählen. Alternativ kann er fragen, wo eine Abteilung, ein Büro oder eine Person sich befindet.

Abbildung 2 Multimodal bedienbarer Lageplan

Der sprachgesteuerte Applikationsteil von FH-Info besteht aus den Grammatiken für die Spracherkennung und den Audio-Prompts, die für die Kommunikation mit dem User benötigt werden. Alle Audio-Prompts wurden mit einer Abtastfrequenz von 11 kHz und einer 16 Bit Kodierung synthetisiert. Die Audio-Prompts wurden so ausgewählt, dass sie auf eine einfache Weise zu unterschiedlichen Sätzen verkettet werden konnten. Zum Beispiel, um den Satz:

<Das Hausmeisterbüro befindet sich im Gebäude 8>

abzuspielen werden die folgende Audiofiles sequentiell abgespielt:

1. De_Hausmeister.pcm (Das Hausmeisterbüro)

2. De_Befindet.pcm (befindet sich im)
3. De_Geb_8.pcm (Gebäude 8)

Der erste Teil eines Prompts ist immer die Sprache, in der ein Prompt synthetisiert wurde. Die Prompts können in verschiedenen Sprachen vorhanden sein und der Dialog Manager entscheidet anhand der aktuell ausgewählten Sprache, welches Audio-File abgespielt werden muss. Die Namen der Prompts werden in dem folgenden Format geschrieben:

Sprache + PromptName + Erweiterung

Derzeit werden die deutsche und englische Sprache unterstützt. Die zugehörigen Kennzeichen sind „De“ für Deutsch und „En“ für Englisch.

Die Grammatik für die Spracherkennung enthält ca. 50 Wörter, die in verschiedene Gruppen aufgeteilt sind. Z.B. enthält eine Gruppe unterschiedliche Frageformulierungen, wie z.B. wo ist, was ist, was befindet sich etc., und eine andere Gruppe enthält die Objekte, wie z.B. Gebäude 1, Sekretariat etc.

Die Applikation FH-Info zeigt auf dem Display einen Lageplan, bei dem die verschiedenen Gebäude schematisch dargestellt sind. Wenn ein Gebäude als Folge einer Sprach- oder Stifteingabe ausgewählt wird, wird eine Ellipse um das Gebäude gezeichnet und der Namen des Gebäudes angezeigt. Der Benutzer kann mit dem Stift ein „Kreis“ auf dem Display zeichnen und damit ein oder mehrere Gebäude auswählen. Jedem Gebäude wurde eine Oberfläche zugeordnet, die intern als Koordinaten-Set gespeichert ist. Wenn der Kreis gezeichnet wird, wird er mit einem Rechteck angenähert. Wenn dieses Rechteck die Oberfläche eines Gebäudes überschneidet, wird die Überschneidungsoberfläche berechnet. Ist mehr als 30% der gesamten Oberfläche des Gebäudes bedeckt, wird das Gebäude ausgewählt.

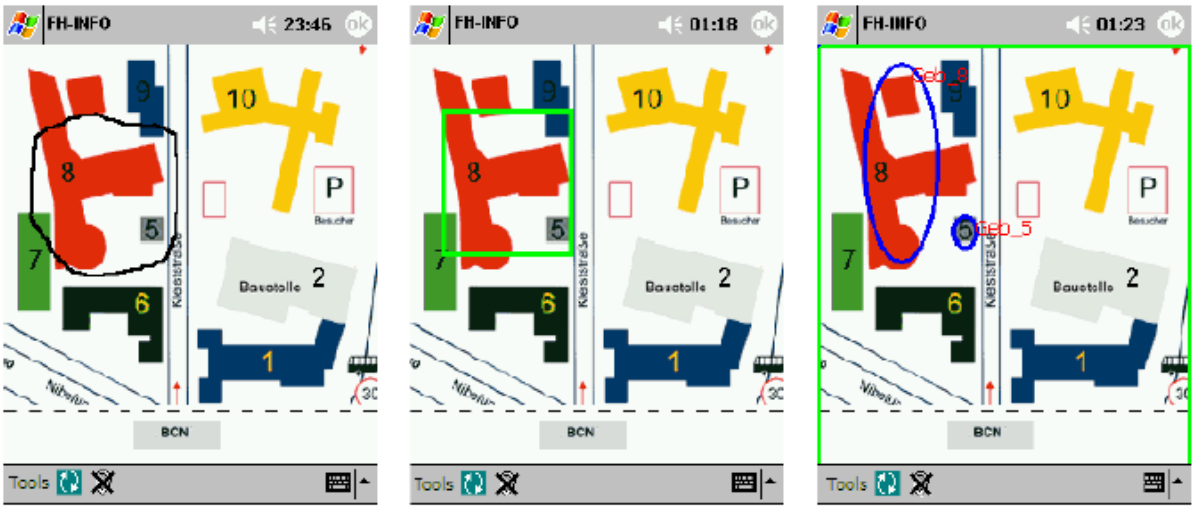

Abbildung 3: Auswahl von Objekten

Die Abbildung 3 zeigt dieses Verfahren. Der durch die Eingabe des Users entstehende Kreis wird zunächst in ein Rechteck umgewandelt, um die überlappenden Oberflächen zu berechnen. Obwohl vier Gebäude vom Rechteck überschnitten werden, resultieren nur zwei ausgewählte Gebäude (Gebäude 8 und Gebäude 5). Die Gebäude 7 und 9 werden nicht

selektiert, da die überschnittenen Oberflächen zu klein sind. Die ausgewählten Objekte werden dann als Eingaben der Plattform übermittelt und dort dem Dialog Manager kommuniziert.

Der Dialogablauf wird über die Befüllung von Slots gesteuert. Mittels dieser Variablen werden die Eingaben in den Dialog transportiert. Die Applikation FH-Info besteht aus verschiedenen Zustände, zwischen denen je nach Befüllung der Slots gewechselt wird. Wenn die Applikation startet, befindet sie sich zunächst im Zustand „START“, dem kein Slot zugewiesen ist. Es wird ein Begrüßungsprompt abgespielt, in dem kurz erklärt wird, wie die Applikation zu bedienen ist. Der User kann dann ein oder mehrere Gebäude auswählen oder eine von der Grammatik abgedeckten Frage stellen. Wird ein Gebäude ausgewählt, dann bekommt der Slot „Target“ den Name des Gebäudes zugewiesen und der Dialog geht in den Zustand „INFO“. Hier werden anschließend Audio-Prompts über die Inhalte des Gebäudes abspielt. Der Slot „Target“ wird ebenfalls gefüllt, wenn der User nach dem Inhalt eines bestimmten Gebäudes fragt.

Alternativ kann ein User kann auch ein Büro, eine Abteilung oder eine Person suchen. Wenn die Frage eines von diesen Elementen betrifft, wird der Slot „Find“ mit dem Namen des Elements gefüllt und der Dialog geht in den Zustand „SEARCH“. Das zu findende Objekt wird gesucht und seine Stelle dem User kommuniziert. Dies geschieht sowohl akustisch als auch grafisch, indem das entsprechend Gebäude mit einer Ellipse markiert wird. In Abbildung 4 wird ein typischer Dialogablauf dargestellt, bei dem ein User sich nach dem AstA erkundigt.

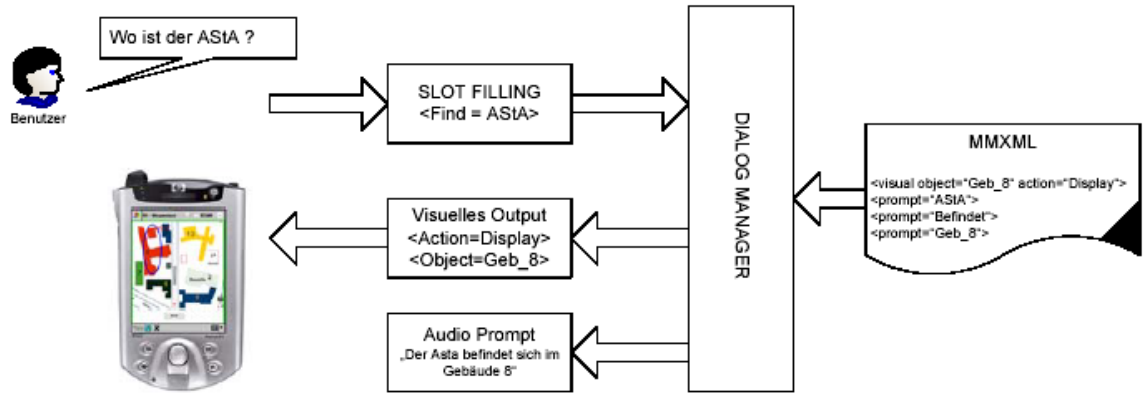

Abbildung 4 Beispiel eines Dialogablaufs

5 Zusammenfassung und Ausblick

Von der ATIP GmbH wurde eine mobile Interaktionsplattform für einen Personal Digital Assistant vom Typ Pocket-PC entwickelt, auf der multimodale Applikation mit Eingaben per Sprache und Stift/Grafik ablaufen können. Trotz der beschränkten Rechenleistung eines Pocket PCs gelang es, eine universell nutzbare Interaktionsplattform zu realisieren. Die entwickelte Interaktionsplattform besteht aus einem Dialog-Manager zur Dialogsteuerung, einem Dialoginterpreter für die Interpretation einer multimodalen Dialogbeschreibung, einem Spracherkenner Siemens AG (Bereitstellung und Unterstützung durch MediaInterface

Dresden GmbH) für die Spracheingabe und einem Audio-Plugin für die Sprachausgabe. Für die Kommunikation zwischen diesen Komponenten wurden geeignete Protokolle und Nachrichten definiert.

Zur Steuerung des Dialogablaufs wurde eine einfache Dialogbeschreibungssprache definiert, deren Tags während der Dialogausführung interpretiert werden. Für die Anbindung von Applikationsteilen müssen nur die aktiven Objekte mit ihren Zuständen oder Aktionen definiert werden. Diese Kapselung erlaubt es, eine bestehende Applikationen in einfacher Weise mit einer multimodalen Interaktion zu versehen.

Die Weiterentwicklung der Interaktionsplattform konzentriert sich derzeit auf die Realisierung von Dialogen, die aus mehrer Dokumenten bestehen. Diese sollen während der Dialogausführung auch dynamisch nachladbar sein.

Literatur

[1] Dusan, S. Flanagan, J.: Multimodal Interaction on PDA's Integrating Speech and Pen Inputs. Proc. of EUROSPEECH 2003,Genf, pp. 2225-2228

[2] Boi, Giancarlo: Entwicklung einer multimodalen Interaktionsplattform für Pocket PC. Diplomarbeit Fachhochschule Frankfurt am Main, Fachbereich Elektrotechnik, Januar 2004.

[3] Reininger, H. , Kasper K., Boi, G. et al.: Skalierbare Voice-Plattform mit Unterstützung Multimodaler Interaktion. In Kristian Kroschel (Hrsg.) Tagungsband der 14. Konferenz Elektronische Sprachsignalverarbeitung, Karlsruhe, w.e.b. Universitätsverlag 2003, pp. 307-314.