Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Projektbericht - SoSe 2009

Kristoffer Witt

Spracherkennung im Kontext Living Lab Hamburg

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhaltsverzeichnis

1 Einführung

Das sich im Aufbau befindende Living Lab Hamburg der HAW Hamburg bietet eine Vielzahl von möglichen Szenarien aus vielen Bereichen der Informatik. Ein besonderer Schwerpunkt ist allerdings auf den Bereich Human Computer Interaction gelegt. Sei es bei der Verrichtung alltäglicher Arbeiten, wie zum Beispiel beim Kochen, der Pflege von Pflanzen und/oder Haustieren, der Planung der Abendgestaltung , aussuchen von Restaurants oder durchsuchen des Kinoprogramms. In vielen Fällen kann ein Computer mit Verbindung zum World Wide Web unterstützend tätig werden. Im Kontext der oben genannten Tätigkeiten wäre zum Beispiel die Analyse des Kühlschrankinhalts resultierend in einem Rezeptvorschlag denkbar, oder aber die Abfrage des aktuellen Kinoprogramms über das Internet unter Berücksichtigung der Vorlieben des Bewohners.

Ein besonderes Anliegen bei dieser Bedienung ist, dass sie möglichst unmittelbar, direkt und simpel von statten gehen soll. Als Bewohner soll man nicht örtlich an Notebook und Tastatur gebunden sein, sondern in allen Lebenslagen die Möglichkeit behalten mit den vorhandenen, möglichst unsichtbaren, Computer zu interagieren. Die Interaktionsmodalität die die Anforderungen nach Unmittelbarkeit und Einfachheit, für den Benutzer, wohl am besten erfüllen kann ist die der Spracherkennung. Sprache und Laute sind neben Gestik die ersten Kommunikationsmittel mit denen sich ein Mensch verständlich machen kann. Sprache ist direkt, unmittelbar und bietet eine hohe Informationsdichte, dadurch ist allerdings auch eine gewisse Interpretationsfreiheit und Vieldeutigkeit gegeben, die der einfachen Verarbeitung entgegensteht.

1.1 Gliederung

Diese Arbeit beschäftigt sich mit der Eingabemodalität Sprache im Kontext des Living Labs Hamburg. Für die Analyse der Leistungsfähigkeit und die Einarbeitung in die zur Verfügung stehenden Technologien wurden verschiedene Prototypen implementiert.

Der Abschnitt Analyse zeigt wie Sprachverarbeitung ihren Einsatz finden könnte und welches die Hauptproblematiken dabei sind.

Im Abschnitt Prototypen werden kurz die implementierten Programme vorgestellt und in den Zusammenhang mit dieser Arbeit und dem angestrebten Szenario gebracht.

Das Kapitel Problematiken beschäftigt sich mit den aufgetauchten Schwierigkeiten und bestehenden Problemen.

Im Ausblick und Fazit sind die angestrebten nächsten Schritte sowie eine Zusammenfassung der geleisteten Arbeit zu finden.

Zugehörige Arbeiten

Diese Arbeit baut auf bzw. ist Bestandteil von verschiedenen weiteren Projekten des Living Labs Hamburgs. Genauer den Arbeiten von Ali Rahimi, Matthias Vogt und André Goldflam.

Ali Rahimi und Matthias Vogt beschäftigen sich mit der Eingabemodalität Touch/Multitouch. Ihr Dashboard bietet die aktuelle Testplattform Oberfläche für einen Teil der entwickelten Prototypen.

Der von André Goldflam entwickelte XML-Wrapper ermöglicht einen von der Implementierungssprache unabhängigen Zugriff auf die Zentrale System Komponente „EventHeap“. Das Kernstück hinsichtlich der Entkopplung der einzelnen Komponenten.

Für Nähere Informationen sei auf die jeweiligen Projektberichte verwiesen.

2 Szenario/Ziel

Dieser Abschnitt beschreibt das umzusetzende Szenario und damit das angestrebte Ziel dieser und der verbunden Arbeiten.

Das grundsätzliche Ziel dieser Arbeit ist es, Sprachverabeitung als Bedienungsmodalität für das Living Lab Hamburg zu etablieren beziehungsweise nutzbar zu machen. Das bedeutet es soll möglich werden alle Aspekte der Wohnung, wenn Sinnvoll, durch Sprache steuerbar zu machen. Da dies im zeitlichen Rahmen des ersten Teil des Projektes nicht umsetzbar ist konzentriert sich diese Ausarbeitung auf ein prototypisches Teilszenario. Es handelt sich um einen Kontextwechsel und die damit verbundenen Befehle und Aktionen. Die aktuelle Tätigkeit eines Bewohners lässt sich in einen oder mehrere Kontexte einordnen. Beispielkontexte sind somit:

Arbeit Die Person nutzt die Wohnung als Homeoffice und beantwortet Korrespondenz, plant Termine oder führt sonstige Jobbezogene Aktionen durch.

Leisure/Recreation Der Bewohner hält sich in der Wohnung auf und erholt sich z.B. durch Musik hören, Fernseh gucken oder ähnlichem.

Ein Wechsel vom Leisure- in den Arbeitskontext könnte nun durch einen Anruf auf der Geschäftsleitung der Wohnung ausgelöst werden. Der Nutzer sitzt aktuell auf der Couch und sieht fern als ihn ein Anruf eines Geschäftspartners erreicht. Nach expliziter Bestätigung des Kontextwechsels werden automatisch verschiedene Aktionen durchgeführt: das Licht wird eingeschaltet, das Fernsehprogramm wird per Time-Shift aufgezeichnet und die Arbeitsoberfläche wird geladen.

Dieses Szenario wurde gewählt um möglichst alle beteiligten Arbeiten integrieren zu können und daraus resultierend eine Demonstrationsapplikation zu erhalten.

3 Analyse

Wie in der Einführung bereits beschrieben bietet die Erkennung und Synthese von Sprache hohes Potential für den Einsatz in der Umgebung des Living Labs Hamburg. Im Folgenden wird eine exemplarische Übersicht über die verschiedenen Ansatzpunkte gegeben.

3.1 Mögliche Anwendungen von Sprachverarbeitung im Living Lab Hamburg

Eingabe

Haupteinsatzbereich von Spracherkennung ist seit jeher die Erkennung von diktiertem Text. Sie ermöglicht die direkte Interaktion zwischen Rechner und Bediener. Szenarios im Living Lab die auf diese Funktion zurückgreifen wären zum Beispiel das Diktieren von Einkaufslisten, E-Mails und Notizen. Weiter lassen sich natürlich auch verschiedene Kommandos definieren um spezielle Aktionen auszulösen. Beispielhaft sei hier die Haussteuerung genannt, also Regelung von Hausequipment oder Multimedia-Technik (TV, Stereoanlage usw.) durch einfache Befehle wie „Fernseher einschalten auf Kanal XY“.

Als weitere Anwendung wäre die Nutzung der durch Information-Mining gewonnen Daten denkbar. Das bedeutet, durch die Aufnahme und Analyse von Äußerungen können Profile des Bewohners gebildet werden um diese bei verschiedenen Aufgaben besser unterstützen zu können. Äußerungen wie „Am liebsten bestelle ich Pizza bei Lieferdienst XY“ oder „Bei dieser Serie schlafe ich immer ein“ würden also beim nächsten Essensbestellvorgang bzw. Auswahl des Abendlichen Fernsehprogramms berücksichtigt werden können. Auch denkbar wäre die Echtzeit-Analyse von Telefongesprächen:

Beispiel-Dialog zwischen Bewohner (B), Anrufer (A) und System (S):

...

B Wir waren lange nicht mehr im Theater.

(S) Erkennt die Absicht eines Theaterbesuchs und stellt das Theaterprogramm für B dar.

A Stimmt, das sollten wir mal wieder machen. Wie sieht es denn zeitlich aus bei dir?

B Gute Frage ich schau mal in meinen Kalender.

(S) Zeigt den Kalender von Benutzer B an ohne dessen direkte Eingabe.

...

Kritisch muss hierbei die Einschränkung der Privatsphäre betrachtet werden. Neben der Konfigurierbarkeit entsprechender Funktionen muss auf die Sicherung der Sprachinhalte gegen nicht autorisierten Zugriff entsprechend geachtet werden.

Ausgabe von Sprache

In Bereichen bzw. zu Zeitpunkten, in denen Informationen dem Hausbewohner nicht über ein Display zugänglich gemacht werden können, sei es wegen nicht Verfügbarkeit von Anzeigegeräten oder weil die Augen des Anwenders anderweitig beschäftigt sind, ist es denkbar diese Informationen per Audioausgabe verfügbar zu machen. Dies hängt natürlich stark von Art der Informationen ab, nur solche Informationen die sich für Sprachsynthese eignen sollten auf diese Weise präsentiert werden. Welche dies sind müsste bestimmt werden. Als Beispiel wäre folgender Dialog zwischen System(S) und Anwender(A) denkbar:

...

A Welche Filme laufen heute Abend auf Sender XY?

(S) Sender XY zeigt heute um 19:00 die Serie XYXY. Anschließend den Spielfilm X. Wollen Sie wissen, was später läuft?

A Nein. Was läuft denn zu der Zeit auf Sender YX?

(S) Sender YX zeigt eine Life Sport Übertragung (...)

...

3.2 Analyse der möglichen Anwendungen

Aus den genannten Anwendungen und Beispielen lassen sich gut Anforderungen an das System ableiten.

Genauigkeit der Erkennung

Um Eingaben zu erkennen muss eine hohe Qualität der Erkennung vorausgesetzt werden. Diese ist bei aktuellen Spracherkennungs-Systemen immer mit einem Personengebundenen Profil verbunden. Zusätzlich zu dem vorher angelegten trainierten Profil müssen die zu erkennenden Phrasen spezifiziert werden. Der Sprachraum ist zu groß, als dass eine eindeutige Zuordnung von geäußerter Sprache auf Kommandos zuverlässig durchgeführt werden kann (siehe auch (Rabiner, 2003)).

Semantik/Ontologie/Dexis

Wichtig für die Verarbeitung von Äußerungen ist deren Bedeutung bzw. Semantik. Anhand des Inhalts wird festgelegt welche Aktionen ausgeführt werden sollen. Daher muss ein Begriffsraum gebildet werden der möglichst klein ist, aber genügend Freiraum lässt, dass eine natürliche Bedienung per Spracheingabe gewährleistet bleibt. Um dies zu erreichen sollten sogenannte Aktionen definiert werden. Analog zu aus der OO Entwicklung bekannten Methoden haben diese Aktionen eine eindeutige Signatur, also eine Bezeichnung und eine Anzahl typisierter Parameter. Die Systemkomponenten des Gesamtsystems sollten jeweils ihr Repertoire an Aktionen öffentlich machen, bzw. an zentraler Stelle registrieren. Dadurch wird eine Verknüpfung von Spracheingaben und eine Abbildung der Worte auf Aktions-Parameter ermöglicht. Als generisches Beispiel könnte hier die hypothetische Aktion „Zeige(ort|gegenstand|person|objekt,parameter)“ dienen. Bezüglich Dialog 1 könnte die Äußerung die das System zum Anzeigen des Kalenders veranlasst also übersetzt werden in „Zeige(kalender,“heute“)“. Um Diese Art Aktionen definieren zu können, bedarf es einer Systemweit gültigen Begriffswelt (Ontologie). Diese definiert z.B. die als Parameter verwendbaren Typen. Einen besonderen Stellenwert bezüglich der Semantik nehmen die sogenannten deiktischen Begriffe ein. Also Worte die erst in Kombination mit einem bestimmten Kontext des Nutzers Sinn ergeben. „Meine“, „hier“ und „dann“ sind solche Begrifflichkeiten. In Sätzen wie: „Zeige mir hier meine Kontakte“ muss das System den Sprecher identifizieren und wissen dass sich „hier“ zum Beispiel auf das aktuelle Ausgabemedium bezieht um die gewünschte Aktion durchführen zu können.

Feedback/Dialogsteuerung

Problematisch bei der Interaktion nur mit Sprache neben fehlendem visuellen Feedback auch die nicht eingeschränkte Wahl der Steuerungsbefehle. Da es für das System schwierig ist, zwischen deliberativen Eingaben und anderen Äußerungen des Nutzers zu unterscheiden, muss ein Feedbackmechanismus geschaffen werden. Dieser vermittelt dem Nutzer z.B. visuell das sich das System im Aufnahmemodus befindet und auf Spracheingabe wartet. Dazu gehört auch die Quittierung von nicht erkannten Eingaben.

Da aktuelle Spracherkenner eine starke Einschränkung der zu erkennenden Wörter benötigen um zuverlässig zu arbeiten, müssen die Verfügbaren Sätze und Phrasen dem Nutzer verfügbar gemacht werden. Als Beispiel wäre ein Meta-Kommando denkbar, wie „Gib mir meine Optionen.“. Je nach aktuellem Kontext würde das System dann passende Äußerungen darstellen oder vorlesen. Mithilfe dieser Funktion wäre es auch einem Erstnutzer ohne Vorkenntnissen möglich mit dem System zu arbeiten.

4 Eingesetzte Technologien

Im Folgenden werden kurz die eingesetzten Technologien erläutert die für die Umsetzung des Szenarios als Sinnvoll erachtet werden.

SAPI Die Microsoft Speech API bietet die Möglichkeit über eine in DOT.NET definierte Schnittstelle auf die in vielen Microsoft Betriebssystemen nativ vorhandene Spracherkennung zuzugreifen. Sie ermöglicht das trainieren auf einen bestimmten Benutzer mit Hilfe von Lerntexten. Leider ist ein es nicht möglich verschiedene Sprecher-Profile gleichzeitig zu laden bzw. mit Hilfe der API zu wechseln. Die API bietet ebenfalls Methoden für die Synthese von Sprache.

WPF Windows Presentation Foundation ist die in DOT.NET verfügbare gestaltungs API für Oberflächen. Mit Hilfe dieser API ist es mit geringem Aufwand möglich verschiedenste Medien darzustellen und ansprechend zu präsentieren.

5 Prototypen

Im Zuge der Einarbeitung in die vorgestellten Technologien sind verschiedene Prototypen entstanden. Diese repräsentieren größtenteils einen Aspekt des Szenarios oder können als hinführende Vorstufen betrachtet werden. Der folgende Abschnitt listet diese Prototypen sowie die Beweggründe die zur Erstellung dieser geführt haben und ordnet sie in das Szenario ein.

5.1 Dashboard

Das Programm „Dashboard“, frei übersetzt in etwa Armaturenbrett, ist der erste Ansatz einer Multimodal bedien baren Benutzerschnittstelle. Im Szenario würde diese überall dort ihren Einsatz finden, wo eine genügend große Anzeigefläche vorhanden ist. Dabei sind Single/Multitouch -Kapazitäten der Anzeigehardware optimal aber optional. Folgende Funktionsweise wird von der aktuellen Ausbaustufe unterstützt:

(Multi)Touch Interaktion mit multimedialen (Grafiken, Videos, 3D-Modellen) und Interaktiven (Schaltflächen, Bildergalerien u.a.) Inhalten

Sprachein-/ausgabe

- Erstellung von Notizen durch Diktat. (Qualität stark korreliert mit Sprachtrainingsprofil)
- Sprachsynthese der Notizen

- Interaktion mit Dashboard-Elementen, das bedeutet hinzufügen(laden), löschen, editieren von neuen GUI-Elementen mit Sprachbefehlen

Verteilung

- Präsenzindikation, die aus populären IM-Programmen bekannte Anzeige der aktuell verfügbaren verteilten Instanzen.
- Versand von Elementen an die verfügbaren Instanzen per Drag and Drop des Elements(z.B. eine Notiz oder ein Bild) auf einen sichtbaren Avatar.

Motivation

Die Entwicklung der Sprachsteuerungskomponente des Dashboards wurde vor allem motiviert durch die Einarbeitung in die Microsoft Speech API-Technologie(SAPI). Insbesondere der dynamische Charakter der Oberflächenbedienung stellte eine Herausforderung dar. Statik ist für eine Spracherkennung besonders wichtig, da die Sprach- und Audiomodelle nur für einen bestimmten Wortschatz trainiert werden. Aufgrund der Möglichkeit Bedienelemente hinzuzufügen, mit Namen zu versehen und sie dadurch ansprechbar zu machen wird diese Statik aufgebrochen. Diese Schwierigkeiten ließen sich mit der Microsoft Speech API lösen, da sie unterstützende Funktionen für dynamische Anpassungen der Modelle auch während des Betriebs bietet.

Einordnung

Das Dashboard ist die Hauptbenutzeroberfläche, also das Betriebssystem das für die Darstellung verschiedenster Komponenten zuständig ist.

5.2 MP3-Player für das Dashboard

Wie bereits beschrieben ist Dynamik für eine Spracherkennung äußerst schädlich. Um eine bessere Abschätzung für die Auswirkungen zu erhalten wurde ein Sprachgesteuerter Mp3-Player für das Dashboard konzipiert und umgesetzt. Dieser unterstützt folgende Funktionalität:

- Laden eines spezifizierten Verzeichnisses und einlesen der Medien Informationen aus ID3-Tags
- Steuerung des Abspielverhaltens über Sprachbefehle, also „Start“ und „Stop“

- Anwahl und hinzufügen von verschiedenen bekannten Musikstücken anhand von Interpret, Songtitel, Genre oder Albumtitel mit Hilfe von Sprache(z.B. „Öffne Album „Jazz ist anders““ oder „Öffne Genre Rock“)
- Visualisierung der Abspielliste durch ein durch Touch steuerbares GUI-Element das anhand des etwaig verfügbaren Albumtitels und Interpreten, das Albumcover des Musikstücks durch den Amazon Webservice lädt und darstellt.

Motivation

Durch die hohe Varianz in Musikdateien bezüglich Song und Albumtiteln, Eigennamen (Künstler) und unterschiedliche Sprachen wurde eine hohe Schwierigkeit für die Erkennungs-Engine geschaffen. Die durchgeführten Tests verliefen zum Großteil positiv, da die Beschränkung auf einen geringen Datenbestand wenige bzw. keine Kollisionen bei den zu erkennenden Wörtern verursachte.

Einordnung

Der Mp3-Player ist stellvertretend für den Leisurekontext. Er ergänzt die Abspielkapazitäten des Dashboards, also z.B. Videos und 3D-Modelle, um eine Wiedergabelisten gesteuerte Musikwiedergabe.

5.3 Datenbankgestützte dynamische Spracherkennung

Für die Simulation einer zentralen Datenbank wurde ein System entwickelt das aus Textdaten eine in-memory Datenbank schafft, aus der dynamisch Sprachmodelle gebildet werden können. Die Datenbank enthält die von der Spracherkennung zu erkennen Sätze und Begriffe. Sie bietet zusätzlich über eine in relationale Struktur gebrachte Hierarchie eine simplifizierte Verknüpfung mit Semantischen Elementen.

Anwendung Google Maps Steuerung

Als Testfall für die Datenbankgestützte Spracherkennung wurde eine Steuerung für Google Maps entwickelt. In der Datenbank wurden verschiedene Phrasen abgelegt und über eindeutige IDs mit Semantischen Elemeten verknüpft. Als Beispiel der Satz: „Zeig mir <ort>“. <ort> ist eine Kollektion aus in der Datenbank abgelegten Begriffen die jeweils eindeutig auf ein Element mit Koordinaten verweisen. Der Ablauf einer typischen Anfrage ist Abbildung **??** dargestellt.

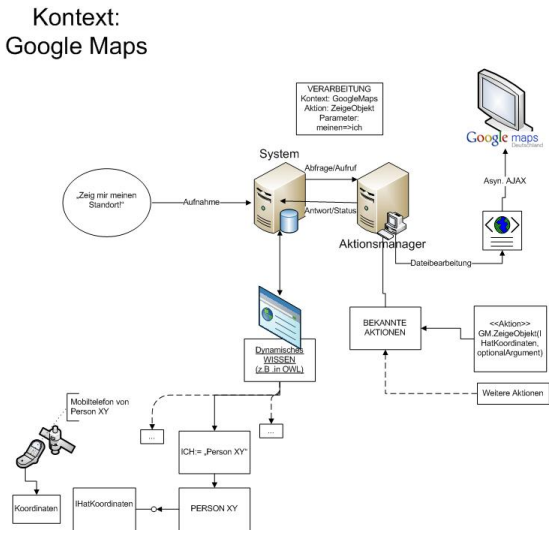

Abbildung 1: Ablauf einer Anfrage an den Google Maps Prototypen

Einordnung

Die Visualisierung von Standorten bzw. die Planung von Routen ist sowohl für Freizeit- wie auch Geschäftlichenaktivitäten Sinnvoll. Im Szenario repräsentiert diese Komponente den Arbeitskontext. Also z.B. die Anfrage „Zeig mir die Hamburger Straße Nr. 32“. Um den Nächsten Bahnhof für Anfahrt mit öffentlichen Verkehrsmitteln zu planen.

Die Datenbank gestützte Erkennung ist das Kernstück der Sprachmodalität. Sie stellt die Funktionalität zur Verfügung, die später von allen in der Wohnung befindlichen und per Sprache steuerbaren Komponenten genutzt werden soll. Insbesondere die Definition von Aktionen und Semantischen Begriffen ist dafür von großere Bedeutung.

6 Schwierigkeiten

Dieser Abschnitt beschreibt die auftretenden und erwarteten Problem sowie, zum Teil, mögliche Lösungsansätze.

Die Probleme von Spracherkennung die in Zusammenhang mit dieser Arbeit auftauchen, sind hauptsächlich die typischen Probleme von Erkennung außerhalb einer Laborumgebung. Zu den Problemen wie der Qualität der Audiodaten existieren hier zusätzlich auch Schwierigkeiten bei der Dialogsteuerung und der zugehörigen Infrastruktur.

Unterscheidung zwischen Sprache und Lärm Um ein möglichst gutes Ergebniss zu erhalten, also eine geringe Wortfehlerrate, muss Sprachinhalt von anderem Audioinhalt, wie zum Beispiel Musik, getrennt werden. Erst nach dieser Vorverarbeitung sollte das Resultat einem Spracherkenner zugeführt werden. Im Living Lab ist dies besonders wichtig da häuf situation auftreten in denen die Mikrophone Nebengeräusche wahrnehmen werden. Als Beispiel sei ein laufendes Radio oder Fernseher genannt. Um dieses Problem zu beseitigen gibt es verschiedene Ansätze Audiodaten zu klassifizieren und dann erst weiterzuverarbeiten (siehe (Spina und Zue, 1996) oder zusammenfassend (Witt, 2009)).

Trennen von Sprechern Analog zum ersten Punkt muss nach der Kategorisierung in Sprach und Lärm eine Zuordnung zu Sprechern erfolgen. Damit deiktische Begriffe wie „ich“, „meine“ und „hier“ auf die richtige Person beziehungsweise Zeigegeste bezogen werden können. Dies bezeichnet man Speaker Diarysation, also das führen einer Zuordnung von Sprecher zu Sprachsegment. Für weitere Informationen sei auf (Nguyen, 2003) oder (Grimaldi und Cummins, 2008) verwiesen.

Trainieren der Profile Um auf die sprachlichen Eigenheiten jedes Sprechers eingehen zu können müssen spezielle Sprachprofile trainiert werden. Für perfekte Ergebnisse wäre es notwendig den gewünschten Sprachraum, also die Begriffe die als Befehl zum Einsatz kommen können, zu kennen. Dies ist allerdings mit sehr hohem Aufwand verbunden und somit nicht umsetzbar. Daher muss ein Trainingstext zum Einsatz kommen der möglichst alle Laute abdeckt und somit als Grundlage für unbekannte Wörter dienen kann. Die eingesetzte Microsoft SAPI bietet zu diesem Zweck eine Trainingsfunktion an. Für andere Erkenner, wie den Hidden Markov Toolkit, werden unter Realbedingungen aufgenommene Audiocorpora benötigt. Es bleibt festzustellen, inwiefern die höhe der Erkennungsrate korreliert mit der Nutzung von bereits verfügbaren (z.B. TIMIT, (Garofolo u. a., 1993), oder der SmartWeb Corpus ((Hannes Mögele, 2006)) oder im fertigen Living Place aufgenommenen Audiodaten. Zu beachten hierbei, dass ein Großteil der verfügbaren Audiocorpora die Englische Sprache abbildet.

Unterschiedliche Erkennungsmethoden Neben der vollständigen spezifikation der möglichen Eingabesätze und Befehle gibt es die Möglichkeit die erkannten Äußerungen über einen Part-Of-Speech Tagger in ihre Bestandteile zu zerlegen und dann weiter zu verarbeiten. Das System wäre somit nicht auf im voraus bekannte Sätze festgelegt sondern könnte anhand der Satzstruktur und der Wortarten entscheiden was zu tun ist. Dafür sind allerdings weitere große Sprachcorpora und Zuordnungen zu Wortform nötig. Zum Vergleich siehe zum Beispiel (Engel, 2002).

Aufnahmebedingungen Da es sich bei den Aufnahmegeräten höchstwahrscheinlich nicht um Personengebundene Mikrophone handeln wird, sondern um Raumgebundene direktionale hat die Raumcharakteristik einen großen Einfluss auf die Ausbreitung des Schalls

und muss somit für die verschiedenen Mikros unterschiedlich normalisiert werden. Siehe (Liu u. a., 2006) für weitere Informationen.

Erstellen von benutzerfreundlichen, intuitiven Dialogen Um die Bedienbarkeit der einzelnen Bestandteile garantieren zu können müssen sollten die Dialog möglichst intuitiv sein. Das ließe sich durch sogenannte Wizard-Of-Oz Experimente erreichen. Bei denen einem Bewohner mitgeteilt wird er könne das System mit seinen Spracheingaben steuern, das System wird allerdings von einer weiteren Person bedient. Die so entstehenden Bedienungstranskripte könnten als Anhanltspunkt für leichte Bedienung herhalten.

7 Zusammenfassung

Wie bereits beschrieben wurden verschiedenste Prototypen entwickelt, anhand deren Verhalten weitere Anforderungen und Problematiken identifiziert werden konnten. Diese Prototypen können als Grundlage für die weitere Entwicklung genutzt werden bzw. als Vorlage für andere Komponenten dienen. Wichtig hierbei ist der Umstieg von Insellösungen hin zum Gesamtkonzept. Also die Vision des Gesamtsystems sollte für Identifizierung von Schnittstellen verwendet werden, damit später die Sprachverarbeitung auch flächendeckend in allen Komponenten Verwendung finden kann. Als Grundlage für die weitere Arbeit ist eine grundsätzliche Ontologie zu sehen. Die dort definierten Begriffe sind die Grundbausteine für alle Sprachbefehle und Kommandos.

7.1 Ausblick

Wie im Technologie Abschnitt beschrieben bietet die bisher verwendete Spracherkennungs API zwar gute personengebundene Ansätze und ermöglicht auch eine fein granulare Definition von zu erkennenden Äußerungen und deren Semantik, es fehlt aber an der wichtigen Skalierbarkeit hinsichtlich der Sprecher. Aktuell wäre es nötig pro Sprecher einen Rechner vorzuhalten da ein direktes umschalten der Sprachprofils nicht möglich ist. Daher ist ein Wechsel zu einer anderen Technologie wohl unausweichlich (z.B. Hidden Markov Toolkit (HTK)). Weiterhin müssen Durchführbarkeitstests konzipiert und ausgeführt werden, die zum Ziel haben das Verhalten des Systems bei realen Bedingungen zu evaluieren. Also Lärm, unterschiedliche Stimmen und verschiedene Mikrofone und Raumcharakteristiken. Die mit den Prototypen durchgeführten Tests sind alle unter Laborbedingungen geschehen und können somit nur als Indikator dienen.

Literatur

[Engel 2002] ENGEL, Ralf: ICSLP-2002: SPIN: language understanding for spoken dialogue systems using a production system approach. (2002), S. 2717–2720

[Garofolo u. a. 1993] GAROFOLO, J. S. ; LAMEL, L. F. ; FISHER, W. M. ; FISCUS, J. G. ; PALLETT, D. S. ; DAHLGREN, N. L.: *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. 1993

[Grimaldi und Cummins 2008] GRIMALDI, Marco ; CUMMINS, Fred: Speaker Identification Using Instantaneous Frequencies. In: *IEEE Transactions on Audio, Speech and Language Processing* 16 (2008), Nr. 6, S. 1097–1111. – URL http://dblp.uni-trier.de/db/journals/taslp/taslp16.html#GrimaldiC08

[Hannes Mögele 2006] HANNES MÖGELE, Florian S.: LREC06:SmartWeb UMTS Speech Data Collection, The SmartWeb Handheld Corpus. (2006), May

[Liu u. a. 2006] LIU, Fu-Hua ; STERN, Richard M. ; HUANG, Xuedong ; ACERO, Alejandro: Efficient cepstral normalization for robust speech recognition. In: *HLT '93: Proceedings of the workshop on Human Language Technology*. Morristown, NJ, USA : Association for Computational Linguistics, 2006

[Nguyen 2003] NGUYEN, Junqua J.-C.: NIST RT03: Pstl's speaker diarization. (2003)

[Rabiner 2003] RABINER, Lawrence: COMPUTER SCIENCE: The Power of Speech. In: *Science* 301 (2003), Nr. 5639, S. 1494–1495. – URL http://www.sciencemag.org

[Spina und Zue 1996] SPINA, M. S. ; ZUE, V.: Automatic Transcription of General Audio Data: Preliminary Analyses. In: *Proc. ICSLP '96* Bd. 2. Philadelphia, PA, 1996, S. 594–597

[Witt 2009] WITT, Kristoffer: AW2: Transkription von Radiospots. (2009), July