Sprachgesteuerte Navigation in komplexen Strukturen am Beispiel eines MP3-Players.

Diplomarbeit

An der
Technischen Universität Dresden
Fakultät Informatik
Dozentur Kooperative Multimediale Anwendungen

in Zusammenarbeit mit

Harman/Becker Automotive Systems GmbH
Speech Dialog Systems
Dialog Research & Tools

Diese Arbeit wurde eingereicht am
06. Januar 2006
von

Stefan Schulz,

geboren am 25. Februar 1980 in Jena.

Autor:	Stefan Schulz
Matrikelnummer:	2702058
Geburtsdatum, Geburtsort:	25. Februar 1980, Jena
E-Mail:	diplomarbeit@stefanschulz.info

Betreuender Hochschullehrer:	Doz. Dr. rer. nat. Hilko Donker
Betreuer Harman/Becker:	Dipl.-Inform. Stefan W. Hamerich
Eingereicht am:	06. Januar 2006

Wettbewerbsrechtlicher Hinweis
Die Aufführung von Namen und Produkten von Herstellern, Dienstleistungsunternehmen und Firmen dient lediglich als Information und stellt keine Verwendung des Warenzeichens sowie keine Empfehlung des Produktes oder der Firma dar. Daher wird auch für die Verwendung und Nutzung solcher Produkte, Dienstleistungen und Firmen keine Gewähr übernommen.

Danksagung

An dieser Stelle möchte ich einigen Personen ganz besonders danken, ohne die diese Arbeit so nicht möglich gewesen wäre.

Insbesondere gilt mein Dank Stefan Hamerich, meinem Betreuer bei Harman/Becker, welcher mir, vom ersten bis zum letzten Tag dieser Arbeit, ein stets kompetenter Ansprechpartner, wirkungsvoller Motivator und sympathischer Kollege war. Gerade das Gefühl, dass er den Erfolg dieser Arbeit genauso wollte wie auch seine Ehrlichkeit gerade auch über meine Schwächen halfen mir, zielstrebig und voller Motivation diese Arbeit voranzutreiben.

Das ich überhaupt die Möglichkeit hatte, diese Arbeit durchzuführen, verdanke ich meinem verantwortlichen Hochschullehrer, Dr. Hilko Donker. Mit der Arbeit konnte ich mich in einem Thema verwirklichen, dass mir sowohl fachlich als auch von seiner Praxisrelevanz als wunderbarer Abschluss meines Studiums erschien. Darüber hinaus halfen mir die präzisen und Widersprüche aufdeckenden Anmerkungen während der Konsultationen auch den wissenschaftlichen Anspruch der Arbeit nie aus dem Auge zu verlieren.

Zum Gelingen der Arbeit wesentlich beigetragen hat auch Patrick Langer, welcher durch beharrliche Fragen nach Nutzen geplanter Untersuchungen und zu Sonderfällen konzipierter Systementwürfe half, Schwachstellen früh zu eliminieren und sich auf das Wesentliche zu beschränken. Seine ausgeprägt andere Vorstellung von Musikauswahl und die daraus entstehenden Diskussionen trugen in großen Teilen zur Sicherstellung von Objektivität in den Untersuchungen bei. Unbezahlbar schließlich war seine Hilfsbereitschaft, in Stresssituationen auch auf Kosten der eigenen Freizeit zu helfen.

Schließlich möchte ich auch allen meinen vielen Korrekturlesern danken, deren Mühe es zu verdanken ist, dass die Arbeit in Struktur, Satzbau und Wortwahl die nun vorliegende Qualität erreicht hat.

Dresden, Januar 2006 *Stefan Schulz*

Aufgabenstellung

Thema

„Sprachgesteuerte Navigation in komplexen Strukturen am Beispiel eines MP3-Players.“

Hintergrund

Bereits mit heute verfügbaren mobilen MP3-Playern können große Mengen an Titeln gespeichert und abgespielt werden. Auf einem Desktop PC und in künftigen mobilen MP3-Playern ist die Zahl noch wesentlich größer und es wird schwierig, die Titel ohne eine durchdachte Benutzeroberfläche zu verwalten. Eine Steuerung des MP3-Players allein mit Sprache bietet große Vorteile, der Sprachdialog muss jedoch mit Sorgfalt gestaltet werden. Eine Lösung in diesem Bereich böte die Möglichkeit, allgemeine Rückschlüsse über auditive Navigation in komplexen Strukturen zu formulieren.

Zielstellung

In dieser Arbeit wird die intuitive Navigation in größeren Datenbeständen untersucht. Dabei wird eine prototypische Sprachapplikation für einen MP3-Player erstellt und evaluiert. Im Rahmen dieser Evaluation werden neben verschiedenen Sprachprompts auch non-verbale auditive Interaktionsobjekte untersucht. Der Fokus liegt auf der intuitiven Navigation in größeren Datenbeständen und der einfachen Auswahl von Titeln.

Konkret stellen sich also folgende Aufgaben:

- Erhebung von Einsatzszenarien durch Befragung zukünftiger Nutzer eines MP3-Players,
- Klassifizierung von Einsatz-Szenarien, Bewertung von Vor- und Nachteilen sowie Rahmenbedingungen,
- Entwicklung eines Konzeptes für die Umsetzung des sprachgesteuerten MP3-Players,
- Nutzer-Test einzelner Navigations-Konzepte für einen MP3-Player mittels Prototypen,
- Konzeption der Navigationsstruktur Sprachsteuerung des MP3-Players,
- Implementierung verschiedener Prototypen der MP3-Applikation,
- Nutzer-Tests, Bewertung des Ergebnisses.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Dialogverzeichnis

1 Einleitung

In diesem Kapitel soll zunächst dargestellt werden, was die Motivation für diese Arbeit bildete und welche wesentlichen Problemstellungen sich bei der Bearbeitung ergaben. Anschließend wird ein kurzer Einblick in die Vorgehensweise und Gliederung der Arbeit gegeben.

1.1 Motivation

Als führender Anbieter von Sprachdialogsystemen für den Automobilbereich beschäftigte sich der Bereich Speech Dialog Systems von Harman/Becker Automotive Systems[1] schon länger mit verschiedenen Entertainment-Funktionen im Auto, zu deren Bedienung immer mehr sprachbedienbare Produkte angeboten werden konnten. Dabei existierte bereits Sprachsteuerung für einfache Audiofunktionen, jedoch gab es bis dahin keine Möglichkeit, die Titel direkt über ihren Namen auszuwählen.
Durch die wachsende Verbreitung mobiler Musiknutzung, getrieben durch den Siegeszug der MP3-Player, die durch Anschluss dieser oder der Benutzung fest eingebauter Geräte auch im Auto benutzbar wurden, entstand das Bedürfnis, eine umfassende Möglichkeit für sprachgesteuerte Musikauswahl im Auto zu entwickeln. In einem in der Arbeit von Wang et al. [WHHS05] beschriebenen Ansatz wurde dabei die technische Basis für ein solches System geschaffen. Dabei stand jedoch die Entwicklung der intuitiven und empirisch ableitbaren Dialogstruktur nicht unmittelbar im Mittelpunkt der Betrachtung. Um die Vorteile der direkten Auswahl von Titel jedoch umfassend nutzen zu können, entstand der Bedarf einer umfassenden Betrachtung zum Dialogdesign eines solchen sprachgesteuerten MP3-Players im Auto. Im Rahmen dieser Arbeit sollte dafür ein Dialogdesign mit dem Schwerpunkt einer einfachen Interaktion erarbeitet und getestet werden.

[1] Im Verlauf dieser Arbeit wurde die ehemals unter dem eigenständigen Namen Temic SDS operierende Tochterfirma der Harman/Becker Automotive Systems GmbH in den Bereich Speech Dialog Systems von Harman/Becker umbenannt, weswegen Teile der Materialien im Anhang noch den Schriftzug von Temic SDS tragen. Im Rahmen der Arbeit wird jedoch immer der neue Name benutzt, wenn die Firma erwähnt wird, egal ob dies jeweils einen Zeitpunkt vor oder nach der Umbenennung bezeichnet.

Als eine Möglichkeit bot sich dafür die Verwendung so genannter non-verbaler Interaktionselemente an, was den Ansatzpunkt für die Motivation der Dozentur Kooperative multimediale Anwendungen der TU Dresden bildete. Diese non-verbalen Interaktionselemente hatten bereits in mehreren Forschungsarbeiten ihren Mehrwert gegenüber rein verbalen Dialogkonzepten unter Beweis gestellt. Dadurch bestand die Möglichkeiten, dem Nutzer Informationen zu übermitteln, die grafisch so nicht hätten präsentiert werden können. Insbesondere in der Benutzung dieser non-verbalen Interaktionselemente für das „Erlebbar machen“ der großen Datenbestände üblicher MP3-Sammlungen wurde ein Potenzial für die Verbesserung des Dialogs vermutet.
Ob aus dem im Rahmen der Arbeit entwickelten Dialogkonzept auch allgemeiner Rückschlüsse abgeleitet werden könnten, stellte eine weitere interessante Fragestellung dar.

Beide Ansatzpunkte (von Harman/Becker und der TU-Dresden) ließen sich im Rahmen dieser Arbeit vereinen.

1.2 Wesentliche Problemstellungen

Eine wesentliche Problemstellung stellte zunächst die Abgrenzung des Themas im Bereich Musik dar. Da die meisten Menschen Musik jeweils anders begreifen und verschiedene Ansichten darüber haben, was unter einer Musikauswahl zu verstehen ist, musste dieses Thema sehr grundsätzlich und ausführlich angegangen werden. Ziel war es, eine Systematisierung zu finden, auf deren Basis vorhandene Systeme diskutiert werden und die Anregungen und Ideen für die weitere Entwicklung bereitstellt werden konnten.
Die strikte Orientierung auf das Dialogdesign half zwar den Implementationsaufwand überschaubar zu halten, verhinderte aber andererseits die Verwirklichung vieler Ideen. So musste sich der Dialog nach und nach immer mehr den technischen Gegebenheiten anpassen, ohne dass die Ideale des Anfangskonzepts vernachlässigt wurden, aber unter Einbeziehung der jeweiligen Ergebnisse aus den Nutzertests.
Gleichzeitig musste eine Vorstellung für den Einsatz non-verbaler Interaktionselemente im Umfeld automobiler Nutzung entwickelt werden, welche sich ebenfalls an diesen Voraussetzungen orientieren musste. Dabei konnten die endgültigen Entscheidungen über die konkreten Töne erst sehr spät getroffen werden, da zuerst die Struktur der Anwendung nahezu vollständig bekannt sein musste. Erst damit waren alle Parameter des Verwendungszwecks, der Positionierung und gewünschten Wirkung bekannt. Ab diesem Zeitpunkt blieben nur einige Tage, ein entsprechendes Konzept zu entwickeln. Die Idee, zusammen mit einem Musikstudenten der Hochschule für Musik „Carl Maria von Weber“ Dresden [mus06a] zu diesem Konzept passende Töne zu erstellen, konnte wegen dieser zeitlichen Einschränkungen nicht berücksichtigt werden.

Neben diesen fachlichen ergaben sich auch organisatorische Problemstellungen.
Die Vereinbarkeit der kommerziellen und wissenschaftlichen Ziele dieser Arbeit, wie im Abschnitt zuvor diskutiert, gestaltete sich nicht so einfach, wie dies vor Beginn der Arbeit vermutet wurde. Denn aus den verschiedenartigen Zielen resultierten zum Teil widersprüchliche Vorgehensweisen. So ist es für die Forschung eher nebensächlich, wie ein bestimmtes System als Ganzes funktioniert, solange es realistisch genug bleibt, um den Testgegenstand bewerten zu können. Dagegen ist für eine Firma vor allem wichtig, einen

Eindruck davon zu bekommen, wie verschiedene Funktionen zusammenspielen. Die ganz konkreten Ergebnisse für eine Funktion sind jedoch meistens nicht entscheidend. Hier einen Kompromiss zu finden, war nicht immer einfach.
Hinzu kam ein durch feste Durchführungstermine für die Nutzertests begrenzter starrer Zeitplan, der es erforderte, schon am Anfang alle Termine, Entwicklungsphasen und Meilensteine zu planen. Während der Bearbeitungszeit selbst war wenig Spielraum, diese Planung zu verändern. Ebenfalls mussten Möglichkeiten gefunden werden, den Aufwand für Vorbereitung und Auswertung der Nutzertests zu begrenzen. Die Wiederverwendung des Fragebogens in den nachfolgenden Untersuchungen oder die Benutzung standardisierter, aber schnell auszuwertender Usability-Befragungen wie des SUS-Bogens sind Beispiele für solche Optimierungen.

Schließlich musste im Verlauf der Arbeit ein wenig von der Aufgabenstellung abgewichen werden. So erschien es angesichts der völligen Unwissenheit über den Nutzer und seine MP3-Gewohnheiten vorteilhafter, den Fragebogen hauptsächlich der Ermittlung dieser Daten zu widmen. Die Nutzungsszenarien wurden stattdessen aus der theoretischen Diskussion von Musik und Musikauswahl ermittelt und später mit Hilfe des Wizard-of-Oz(WOZ)-Tests überprüft. Weiterhin wurden im WOZ-Test nicht mehrere einzelne Konzepte getestet, sondern mittels der WOZ-Technik und einer relativ offenen Struktur den Nutzern eine intuitive Nutzung ermöglicht. Aus der Auswertung dieses Tests, zusammen mit den Erkenntnissen aus der dazu begleitenden Powerpoint-Befragung, konnte eine Idealvorstellung für das System geschlussfolgert werden. Schließlich wurden auch nicht mehrere Prototypen implementiert, sondern verschiedene Modi der Prototypen gegeneinander getestet. Das brachte den Vorteil mit sich, das genau nur eine unabhängige Variable im System verändert wurde und damit eine saubere Vergleichbarkeit gegeben war.

1.3 Vorgehensweise und Gliederung der Arbeit

Im Rahmen dieser Arbeit wurde zunächst in vier Grundlagenkapiteln auf Sprachdialogsysteme, Musik, Non-verbale Interaktionselemente und Usability Engineering eingegangen. Dabei sollten für jedes dieser Gebiete die Abgrenzung im Rahmen der Arbeit wie auch die Besonderheiten bei der Betrachtung im Umfeld von Sprachdialogsystemen und dem automobilen Einsatz diskutiert werden. Bereits existierende Arbeiten und Systeme sowie die Abgrenzung des Themas im jeweiligen Fachbereich wurden dabei nicht separat, sondern im Rahmen der Einordnung und Diskussion des jeweiligen Kontextes diskutiert. Mit Kapitel 5 wurden die grundlegende Betrachtungen abgeschlossen und ein Vorgehensmodell für diese Arbeit entwickelt.
Dabei wurde zunächst in Kapitel 6 auf die Befragung mit Fragebogen und den WOZ-Test eingegangen, deren Ergebnisse die Grundlage für die weiteren Überlegungen bildeten. Basierend auf einer Diskussion der bereits vor und während der ersten Tests entwickelten Systemideen und -vorstellungen wurde die prototypische Umsetzung des Dialogs beschrieben. Kapitel 8 beschäftigte sich schließlich mit der Evaluation dieses Prototypen und gab aus den Ergebnissen abgeleitete Überarbeitungsempfehlungen für den Dialog. Eine Betrachtung der Verallgemeinerbarkeit der Ergebnisse schloss sich an.
Die Zusammenfassung in Kapitel 9 ermöglichte eine abschließenden Betrachtung der Ergebnisse. Weiterhin wurde eine Bewertung der Qualität der erreichten Ergebnisse vorgenommen und ein Ausblick auf mögliche weiterführende Arbeiten gegeben.

2
Sprachdialogsysteme

Sprachdialogsysteme ermöglichen dem Nutzer mithilfe seiner Sprache Geräte zu bedienen oder mit Ihnen zu interagieren. Zunächst werden in diesem Kapitel Sprachdialogsysteme eingeordnet und definiert, danach wird nach einem Überblick über bereits vorhandene Systeme ein Einblick in den prinzipiellen Aufbau solcher Systeme gegeben. Die speziellen Anforderungen und Besonderheiten der Sprachdialogsysteme im Auto werden im Abschnitt 2.4 diskutiert. Abschließend wird eine Abgrenzung der Betrachtung von Sprachdialogsystemen im Kontext der Arbeit vorgenommen.

2.1 Einordnung & Definition

„Es wird alles immer gleich ein wenig anders, wenn man es ausspricht.“
Hermann Hesse

Die Magie der Benutzung einer urmenschlichen Eigenschaft, wie die der Sprache als eine Möglichkeit für die Kommunikation mit Computern, Maschinen oder technischen Geräten ist beeindruckend und auch noch sehr neu (für kompakte Betrachtung dieser kurzen bisherigen Entwicklung siehe beispielsweise [Bak05]). Doch abseits der allgemeinen Visionen aus Science-Fiction-Filmen wie „2001“ oder „Star Trek“ herrscht in der allgemeinen Wahrnehmung hauptsächlich die Vorstellung von einer Person vor, die mit einem unhandlichen Headset auf dem Kopf vor einem PC sitzt und diesen mit einfachen Sprachkommandos steuert. Doch ist dieses Vorstellung vielfach gar nicht zutreffend, wie Michael F. McTear in seinem Buch über Sprachdialogsysteme [McT04] anmerkt. Realistischer scheint der Fall, über ein Telefon mit einem sprachgesteuerten System zu kommunizieren, da dieses schon den benötigten Sprachkanal anbietet und fast von überall bedienbar ist. Aber durch zunehmende Verbreitung von Mikrophonen und fortschreitende Miniaturisierung wie zum Beispiel mobilen Geräten wie PDAs, Smartphones oder gar als Teil des Autos werden auch ganz andere Anwendungsgebiete möglich. Insbesondere in der Heimautomatisierung sind hier Anwendungen denkbar. Diese Anwendungsmöglichkeiten werden in Abschnitt 2.2 näher erläutert.

Doch für welche konkreten Funktionen kann Sprache überhaupt eingesetzt werden? McTear zeigt dafür die im Folgenden vorgestellten Möglichkeiten auf [McT04]:

Gerätesteuerung
Oft wird diese Art der Sprachfunktionen auch „Command-and-Control“ genannt. Dabei löst ein Kommando immer direkt eine Funktion bei einem bestimmten System, mag es nun ein Computer, eine Fabrikmaschine oder eine Haushaltsgerät sein, aus. Besonders lohnenswert ist der Einsatz dieser Möglichkeit in Situationen, in denen die Hände mit anderen Tätigkeiten gebunden sind. Sowohl im Auto, wo der Fahrer mit der Primäraufgabe Fahren beschäftigt ist, als auch bei der Maschinenbedienung in der Industrie, bei der der Arbeiter eventuell ein Werkstück mit beiden Händen greifen muss, ist ein solcher Einsatz von Sprachsteuerung von Vorteil. Ebenfalls nützlich ist dieses Paradigma für Personen mit Behinderungen, die andernfalls Schwierigkeiten /Probleme hätten, ein System zu bedienen. Üblicherweise benötigen solche Systeme ein kleines, aber robustes Vokabular, das die beschränkte Funktionalität in all ihren Äußerungsmöglichkeiten abdeckt.

Dateneingabe
Anders als bei der Gerätesteuerung wird bei der Dateneingabe keine direkte Funktion ausgeführt, sondern vielmehr eine Möglichkeit gegeben, Daten per Sprache einer Software mitzuteilen. Hier ist der Einsatz vor allem dann sinnvoll, wenn mehrere Tätigkeiten gleichzeitig ausgeführt werden sollen. Damit ist es zum Beispiel möglich, Geräte zu überwachen, während man Daten über deren Zustand eingibt. Weitere Anwendungsmöglichkeiten für diese Möglichkeit stellen das Ausfüllen von Formularen, das Melden von Staus auf der Autobahn oder die Verschiebung eines Termins im Terminplaner dar. Auch hier wäre nur ein beschränktes Vokabular nötig, nur manche Anwendungen, wie das erwähnte Stau-Melder-System, erfordern die Einbindung von größeren Mengen von Vokabeln (Straßennamen).

Informationsabruf
Das Komplement zu der Dateneingabe ist der Informationsabruf, wobei per Sprache aus umfangreichen Informationssammlungen Daten abgerufen werden. Dazu gibt es eine ganze Reihe von Beispielen meist telefonbasierter Dienste; von Wetter- über Reise- bis hin zu Börsenkursinformation. Da diese Möglichkeit eher für gelegentliche Nutzung ausgelegt ist, muss hier ein größeres Vokabular bereitstehen, um möglichst viele Varianten und Strategien für mögliche Anfragen zu verstehen. Außerdem müssen solche Systeme ein Vokabular unterstützen, dass Zugriff auf die Datenbank-Felder und ihren Inhalt bietet. Da es sich meist um größere Datenbanken handelt, ist das Vokabular dementsprechend auch umfangreicher als bei den bisher vorgestellten Möglichkeiten.

Um Sprachdialogsysteme zu entwickeln, müssen diese Funktionen kombiniert und in einem Dialog zusammenführt werden. In den bisher vorgestellten Funktionen bestand die Interaktion zwischen Computer und Benutzer meist darin, dass der Benutzer eine Äußerung macht, das System darauf reagiert und damit die Interaktion endet.
Die Komplexität des Dialogs erfordert jedoch oftmals eine umfangreichere Interaktion. Dafür muss ein richtiger Dialog zwischen Mensch und Computer entwickelt werden. Dieser sollte einerseits natürlich sein, andererseits auch zielführend. Eine vertiefende Betrachtung

zu Dialogdesign folgt am Ende dieses Kapitels.

Neben diesen grundsätzlichen Betrachtungen ist zusätzlich zu hinterfragen, wann denn ein Einsatz von sprachgesteuerten Systemen sinnvoll ist. Cameron [Cam00] benennt als Antwort im Wesentlichen vier Punkte, in welchen Fällen Sprache für die Kommunikation mit Maschinen vorzuziehen ist:

- Wenn es keine andere Wahl gibt.
- Wenn es sich in die Privatsphäre ihrer Umgebung und die lösende Aufgabe einfügt.
- Wenn die Hände oder Augen mit einer anderen Aufgabe beschäftigt sind.
- Wenn es schneller ist als jede andere Alternative.

Während der erste Punkt meist zwingend aus dem Anwendungsgebiet hervorgeht (z.B. Telefon-Applikation), beeinflussen die folgenden zwei Gründe meist die Bereitschaft des Nutzers, überhaupt mit Maschinen zu sprechen. Den letzten Grund identifiziert Cameron als den besten, der aber noch viel zu selten benutzt wird.
Alle diese Punkte sollten berücksichtigt werden, wenn entschieden wird, ob ein System mit Sprachsteuerung ausgestattet werden soll. Gleichzeitig sollten sie aber ebenfalls bei der Entwicklung beachtet werden, um zu verhindern, dass in einem eigentlich sinnvollen Umfeld falsche Prioritäten für die Sprachbedienung gesetzt werden.

Allerdings muss nicht zwangsläufig mit einem Gerät mit Sprachfunktionen nur per Sprache kommuniziert werden.
Gerade die Kombination mit klassischen Ein-/Ausgabegeräten ist eine viel versprechende Möglichkeit. Diese so genannte Multimodalität ist prinzipiell die Kombination von verschiedenen Modalitäten[1], wie zum Beispiel die visuell-taktile (über einen Bildschirm und Tastatur/Maus), die auditive (Sprachein- und -ausgabe) oder eine rein taktile Modalität. Im Kontext dieser Arbeit wird der Begriff Multimodalität insbesondere für die Kombination visuell-taktiler und auditiver Modalität benutzt.
Bei solchen multimodalen Anwendungen wird Sprache üblicherweise immer nur dort eingesetzt, wo die Verwendung von Sprache anderen Modalitäten überlegen ist bzw. durch redundante Anordnung der Modalitäten eine Benutzung je nach Präferenzen von Nutzer und Aufgabe möglich ist.

Weiterhin ist zu betrachten, wie in einem sinnvollen Kontext all diese vorher angesprochenen verschiedenen Möglichkeiten dem Nutzer möglichst natürlich und leicht verständlich bereitgestellt werden können. Pieaccini und Huerta [PH05] unterscheiden dabei zunächst grundsätzlich zwischen zwei Zielen, die bei einer solcher Entwicklung erreicht werden sollen: Die Schaffung einer möglichst realistischen Simulation eines menschlichen Dialogs und Erreichung möglichst vieler Usability-Ziele[2]. Oft wird aber beides verwechselt.

[1] Allgemein bezeichnet Modalität die Art und Weise eines Vorgangs, des Denkens oder Daseins bzw. die Art der Ausführung [Tim97]. Im Speziellen bedeutet es in welcher Art und Weise mit dem Rechner interagiert wird.

[2] Eine genaue Definition von Usability und die Erläuterung von Usability-Zielen speziell für Sprache finden sich in Kapitel 5 dieser Arbeit

Dabei gibt es eine ganze Reihe von Anwendungen, bei denen Natürlichkeit und Redefreiheit die Usability eher behindern ([Ovi95],[WW04]). Aus diesem Grund wählte die Sprach-Industrie sehr früh den Weg, Benutzbarkeit in den Mittelpunkt zu stellen, während viele Forschungsanstrengungen weiterhin die Schaffung eines möglichst natürlichen Dialogs anstrebten.
Technisch wäre ein solcher natürlicher Dialog dadurch zu erreichen, dass sehr viel Aufwand in das Verstehen des syntaktischen und semantischen Aufbaus eine Äußerung investiert und damit eine möglichst umfassende Datenbasis bereitgestellt wird, aufgrund derer das System daraufhin mit allen vorliegenden Informationen die perfekte Antwort oder Reaktion zu einer Äußerung produzieren kann. Dies soll an dieser Stelle natürlichsprachliche Dialogsysteme genannt werden, da sie zuerst die Natur der Sprache nachbilden, und dann erst sich mit den Einschränkungen technischer und inhaltlicher Art beschäftigen.
Bei diesem Vorgehen, welches vor allem durch Arbeiten aus den Gebieten der Linguistik und künstlichen Intelligenz vorangetrieben wird, ist allerdings ein hoher Aufwand nötig, der nicht in jedem Fall wirklich angemessen ist.

Um eher dem Ziel der Usability Genüge zu tun, beschreibt McTear [McT04] im Gegensatz dazu eine eher pragmatische zweite Möglichkeit. Soll nicht mehr in jedem Fall alles verstanden werden, was der Nutzer sagt, sondern nur, was zu dem Zeitpunkt wichtig ist, reicht dieses meist auch aus, um eine hinreichend realistische Simulation eines Dialogs zu erschaffen. Dafür reichen meist auch technisch einfachere Systeme aus. Dies ermöglicht weiterhin die klare Trennung in verschiedene Subsysteme.
Hierbei soll also nicht die Natürlichkeit der Sprache erreicht werden, sondern lediglich ein System, das im speziellen Kontext natürlich genug ist, um seine Aufgabe zu erfüllen. Ebenso wie von einer grafischen Schnittstelle nicht ständig eine Virtual Reality Simulation erwartet wird, so ist auch die Sprache nur eine spezielle Form der Schnittstelle, mit der es umzugehen gilt.

Ein Beispiel soll das verdeutlichen: Ein solches Sprachdialogsystem würde eventuell auf eine Äußerung wie in Dialog 2.1 richtig reagieren können, obwohl es nicht den gesamten Sinnzusammenhang der Nutzeräußerung auflösen könnte. Ein natürlichsprachliches System dagegen würde zunächst versuchen, die Referenzen (das, ihm) aufzulösen (was nicht unmöglich ist, in Jurafsky [JM00] werden Möglichkeiten für dieses Problem diskutiert).

Dialog 2.1: Beispiel für Äußerung mit Referenzen auf vorher Gesagtes (frei nach [JM00])

sys: Wie lange soll ihr Termin mit John bei Bills Autohändler dauern?
usr: Ich möchte mir das mit ihm eine Stunde lang anschauen.

Im Rahmen dieser Arbeit stand nun hauptsächlich die Usability-orientierte Sicht im Mittelpunkt des Interesses, deshalb soll der Begriff der Sprachdialogsysteme hier wie folgt definiert werden:

Definition 2.1 *Ein **Sprachdialogsystem** im Rahmen dieser Arbeit ist ein aufgabenorientiertes, sprachliches System, welches die Usability des Systems in den Mittelpunkt stellt, um lediglich eine hinreichend realistische Simulation eines natürlichen Dialogs zu erschaffen. Es ermöglicht dabei die Kommunikation zwischen Mensch und Computer und kann auch durch multimodale Elemente erweitert sein.*

McTear [McT04] weist allerdings darauf hin, dass die Grenze zwischen den hier unterschiedenen natürlichsprachlichen und Sprachdialogsystemen immer fließender wird, in heutigen Sprachdialogsysteme werden auch immer mehr Bestandteile aus dem Gebiet des natürlichen Sprachverstehens und der künstlichen Intelligenz eingebaut, und umgekehrt. Welche Elemente für ein System benutzt werden, muss jeweils aus dem aktuellen Kontext abgeleitet werden. Weiterhin soll auch eine mögliche multimodale Ausgestaltung der Systeme in der weiteren Betrachtung mit einbezogen werden.

2.2 Beispiele vorhandener Systeme

Sprachdialogsysteme werden heute hauptsächlich im Rahmen von Hotline-Lösungen eingesetzt. Sie sollen dabei einfache Vorgänge automatisieren helfen und lösen meist Vorgängersysteme ab, welche lediglich auf Eingabe von Tastentönen mit vorgefertigten Antworten reagierten. Solche telefonbasierten Systeme nehmen heute den Großteil der bisher realisierten Sprachdialogsysteme ein.
Ein bekanntes Beis8piel ist die automatische Bahnauskunft, welche kostenfrei unter der Telefonnummer (+49) (0) 800 1507090 Interessierten zur Verfügung steht. Unter der Nummer kann auf die gesamte Verbindungsauskunft der Bahn zugegriffen werden, die mit einem schon mehrmals verbesserten Sprachdialog zur Information über Verbindungen genutzt werden kann.

Doch erschöpft sich die Bandbreite der Anwendung von Sprachdialogsystemen nicht auf Telefonapplikationen. Auch im Bereich des Ubiquitous Computing, der allgegenwärtigen Informationsverarbeitung, bietet sich die Verwendung von Sprache an.
Zuerst angedacht wurde dies im Bereich der Heimautomatisierung. Erste Überlegungen zu diesen Smart Homes gab es in den 70er Jahren, konkreter formuliert dann ab Mitte der neunziger Jahre [Ven96] konnte auch die Verwendung von Sprache zu der Zeit auch erstmals angedacht [Zag95] werden. Doch aufgrund fehlender Standardisierung und hoher Kosten für notwendige Infrastrukturmaßnahmen in diesem Bereich ist dort außer einigen Testhäusern bisher nicht viel vorangekommen.
Die Gerätesteuerung, welche in diesen Häusern angedacht war, wurde aber an anderer Stelle umgesetzt, vor allem in der industriellen Produktion. Solche Systeme ermöglichen, die Maschine oder ein Werkstück mit beiden Händen zu bedienen, aber gleichzeitig weitere Kommandos abzusetzen. Ein solches, sich bereits im praktischen Einsatz befindliches System, ist beispielsweise SpeaKING Control der Firma MediaInterface Dresden GmbH [Med06].
Im Consumer-Bereich haben sich Sprachtechnologien in den letzten Jahren vor allem als Erweiterungen der schon vorhandenen Modalitäten etabliert. Ob PDAs, Handys, eBooks oder gar Wearables (in die Kleidung integrierte Geräte), die Grundidee war immer gleich. Da meist keine großen Displays zur Verfügung standen und auch die Eingabemöglichkeiten meist klein und schwer zugänglich waren, wurde Sprache als eine weitere Modalität zusätzlich zu den bisher bekannten Möglichkeiten benutzt, es entstanden multimodale Anwendungen. Beispielsweise sind heute Handys verfügbar, die neben der vergleichsweise einfachen Freisprecheinrichtung und der Nummernwahl auch für komplexere Funktionen sprachbedient werden können (zum Beispiel PocketPC2003-Handys wie das XDA Mini). Ebenso wird auf PDAs die umständliche Stifteingabe immer mehr auch durch Sprachfunktionen ergänzt (z.B. für Navigationsanwendungen wie den „Navigon MobileNavigator|5“)

und bei Wearables gibt es ebenfalls Ansätze, Sprachbedienung anstatt der teilweise umständlich in die Kleidung zu integrierenden haptischen Bedienelemente zu verwenden (z.B. Normadic Radio [SS00]).
Ein ganz besonderes mobiles Gerät soll hier extra betrachtet werden: Das Auto. In diesem ist seit der Vorstellung des weltweit ersten verfügbare Systems 1996 [Hei01] zu beobachten, dass in immer mehr Baureihen Sprachtechnologie Einzug hält. Heute bieten fast alle namhaften Hersteller solche Systeme an [Han04]. Der Vorteil der Sprache liegt hierbei auf der Hand. Der Fahrer muss zur Bedienung der immer komplexeren Funktionen seines Autos nicht mehr die Augen von der Straße nehmen. Ein Überblick über den genauen Aufbau solcher Systeme und weitere Informationen zu Sprachdialogsystemen im Auto wird Abschnitt 2.4 gegeben.
Über diese bisherigen Anwendungsmöglichkeiten hinaus gibt es auch noch weitere denkbare Einsatzmöglichkeiten im Rahmen des Ubiquitous Computing. So entstand beispielsweise in einem Projekt an der TU-Dresden ein Prototyp eines Abstimmungssystems für das interaktive Kino [JSF05], bei dem Zuschauer im Kino durch Reinrufen an definierten Stellen den weiteren Ablauf des Films beeinflussen können.

Neben diesen bisher erwähnten Möglichkeiten gibt es auch Anstrengungen, den klassischen Computer um Sprachbedienung zu erweitern. Lange war diese Entwicklung auf einfache Diktiersysteme und einfaches Abbilden von Sprachbefehlen auf vorhandene grafische Interaktionselemente beschränkt. Doch in letzter Zeit zeichnen sich zwei Entwicklungen ab: Einerseits wird die Sprachbedienung nicht mehr nur als Aufsatz, sondern als Bestandteil des Computers betrachtet (die nächste Windows Version Vista wird beispielsweise standardmäßig Sprachfunktionen für rund ein Dutzend üblicher Sprachen mitbringen [Bro06]), was dazu führen könnte, dass wirklich integrierte Anwendungen entstehen.
Auf der anderen Seite entstand das Bedürfnis, Webseiten nicht nur grafisch-haptisch zu bedienen, sondern zusätzlich auch per Sprache. Anwendungsmöglichkeiten dieser Technologie bestehen im Anreichern normaler Webanwendungen mit Sprachein- und -ausgabe, was sich bei einigen umfangreichen Eingaben z.B. in Formulare als sehr sinnvoll erweisen kann.

2.3 Prinzipieller Aufbau

Nachdem im Abschnitt 2.1 betrachtet wurde, was Sprachdialogsysteme prinzipiell auszeichnet, und im Abschnitt 2.2 Beispiele vorhandener Systeme aufgezeigt wurden, soll folgend der Aufbau eines solchen Sprachdialogsystems beleuchtet werden.

Den typischen Aufbau eines solchen Systems zeigt Abbildung 2.1 [McT02].
Dabei ist die Spracherkennung für die Umwandlung des kontinuierlichen Signals in eine Menge von Wörtern bzw. Phonemen[3] zuständig, welche dann durch die Sprachverstehenkomponente syntaktisch und semantisch analysiert wird, um eine Repräsentation der Bedeutung zu erhalten. Diese Bedeutung benutzt der Dialogmanager, neben der Kommunikation mit externen Quellen (Datenbankzugriffe, Hintergrundprogramme), um Ablauf und Fluss des Dialoges zu steuern. Zur Ausgabe leitet der Dialogmanager die mitzuteilenden Daten an eine Spracherzeugungskomponente, die unter Berücksichtigung von Auswahl,

[3]Ein Phonem ist die kleinste lautliche Einheit der Sprache, die aber keine Bedeutung trägt.[HF04]

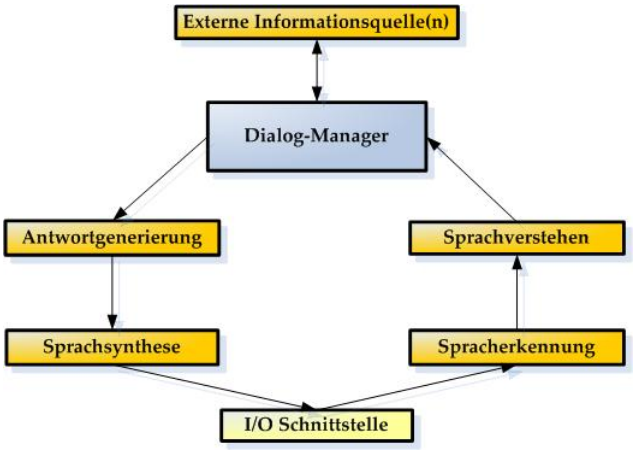

Abbildung 2.1: Architektur für Sprachdialogsysteme [McT04]

Struktur und Form der auszugebenden Informationen eine Antwort erzeugt, die die Sprachsynthese dann in eine gesprochene Form umwandelt.
Der Dialogmanager kann also als eine übergeordnete Schicht begriffen werden, der sich verschiedener Grundlagen-Techniken bedient. Die genauen Aufgaben im System werden später in diesem Abschnitt erläutert. Diese klare Trennung ermöglicht es, ebenso wie Grafik- und Tastaturtreiber von dem Design der grafischen Benutzerschnittstelle getrennt sind, den Sprachdialog von äußerst anspruchsvollen Teilgebieten Spracherkennung und Sprachsynthese zu kapseln und somit ein Interface Design ohne vertiefte Kenntnisse der unterliegenden Schichten zu ermöglichen.

Nachfolgend sollen nun die erwähnten Teile eines solchen Systems näher betrachtet werden. Es wird jeweils darauf eingegangenen werden, welche der spezifischen Eigenschaften für einen Dialogdesigner zu beachten sind und an welchen Stellen er auch die grundlegenden Komponenten beeinflussen kann. Diese Einflussmöglichkeiten sind aus Abbildung 2.2 ersichtlich und werden nun im Weiteren erläutert.
Durch das Wissen um konkrete Einflussmöglichkeiten auf die einzelnen Komponenten, können Schwächen einer Komponente des Systems eventuell mit einer anderen ausgeglichen werden, zum Beispiel eine relativ schlechte Spracherkennung mit einer geschickten Grammatik.
Zunächst werden die grundlegenden Teile beschrieben, am Ende dann der Dialogmanager als die Klammer. Diese Darstellung ist zu großen Teilen dem Kapitel 4 und 5 des Buches von McTear [McT04] entnommen.

Spracherkennung

Die Spracherkennung (ASR, Automatic Speech Recognition) sorgt innerhalb des Sprachdialogsystem dafür, die Nutzeräußerung in eine Sequenz von Wörtern umzuwandeln. Dabei sind in den letzten Jahren große Fortschritte gemacht wurden. Trotzdem bleibt das Hauptproblem der Spracherkennung weiterhin, dass sie keine Garantie geben kann, ob eine Erkennung einer Nutzeräußerung korrekt ist oder nicht. Es bleibt zu diskutieren, warum Spracherkennung so schwer ist.

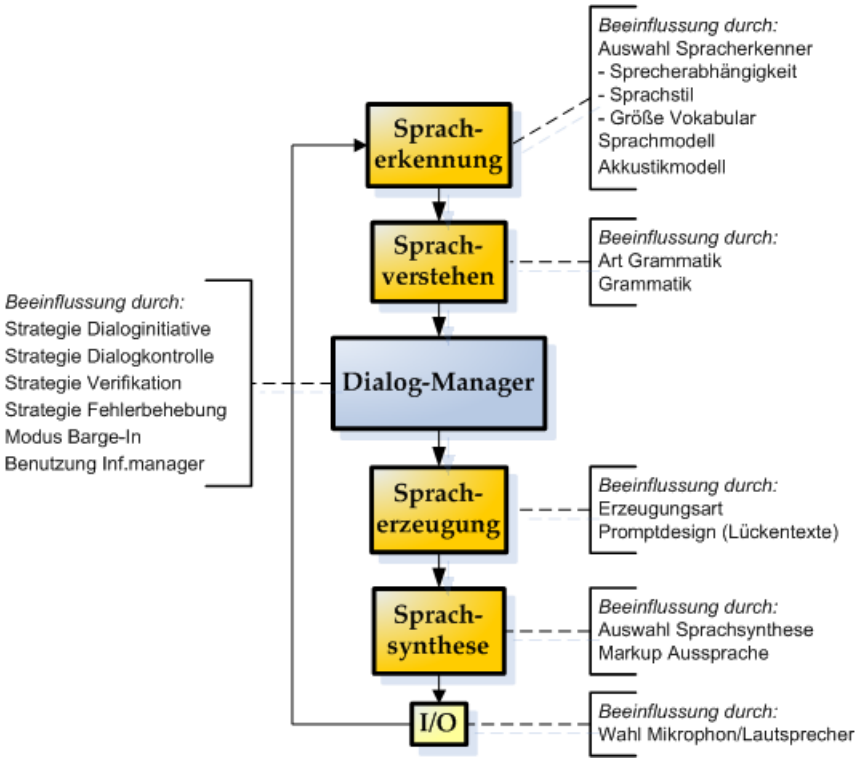

Abbildung 2.2: Einflussgrößen für Dialogdesigner auf Sprachdialogsysteme (nach [McT04])

Das hat zuerst einmal phonetische Gründe, so gibt es viele Phoneme, die je nach der Stelle in Wort, Satz oder Sinnzusammenhang eine andere schriftliche Repräsentation besitzen können. Diese Vieldeutigkeiten in allen Facetten abzufangen, ist eine der Hauptschwierigkeiten der Spracherkennung und verlangt einen großen Aufwand. Eine genauere Betrachtung zu Einzelheiten der Spracherkennung ist nicht Teil dieser Arbeit, mehr dazu findet sich zum Beispiel bei Jurafsky [JM00].
Neben diesen Problemen sind beim Einsatz von Spracherkennung auch Unterschiede der Aussprache zwischen verschiedenen Sprechern (Inter-Sprecher Varianz), ebenso wie auch Unterschiede in der Aussprache einer einzelnen Person (Intra-Sprecher-Varianz) zu beachten. Kanalstörungen und Hintergrundgeräusche bieten darüber hinaus weitere Störfaktoren für die Spracherkennung.
Einige dieser Punkte können durch Optimierungen am Spracherkenner, Nachbearbeitungen am Signal (über Akustikmodelle) oder größere Datenbasis verschiedener Stimmen und Dialekte vermindert werden. Abschalten lassen sie sich aber nicht, da Spracherkennung ein statistisches Verfahren ist, was immer nur eine Wahrscheinlichkeit ermitteln kann, mit welcher eine gewisse Äußerung getätigt wurde. Spracherkennung kann also immer nur eine Annäherung an perfekte Erkennung sein.

Soll nun ein Spracherkenner für ein Sprachdialogsystem ausgewählt werden, sind dabei mehrere Parameter zu beachten:

Sprecherabhängigkeit

Grundsätzlich sind Spracherkenner entweder sprecherabhängig oder sprecherunabhängig . Im ersteren Fall muss der Nutzer den Spracherkenner zusätzlich vor der

Benutzung trainieren, dafür erreicht ein solcher Spracherkenner auch bei großen Vokabular hohe Erkennraten. Wenn der Spracherkenner sprecherunabhängig sein soll (was bei den meisten Systemen Anforderung ist, da sie hauptsächlich von „Laufpublikum“ bedient werden), wird das Training ausschließlich durch eine Auswahl repräsentativer Sprecher durchgeführt. Die Qualität hängt bei diesem Vorgehen wesentlich von dem Umfang und der wirklichen Repräsentativität dieser Stichprobe ab.

Sprachstil

Bei dem angesprochenen Training ist der Sprachstil ein entscheidender Faktor. Je nach Einsatzzweck des späteren Systems soll vielleicht eher Sprache beim Lesen, spontane Sprache oder sehr befehlsorientierte Sprache trainiert werden.

Größe des Vokabulars

Ebenfalls ein wichtiger Parameter ist der Umfang des Vokabulars. Grundsätzlich ist die Erkennrate[4] desto besser, je kleiner das jeweilige Vokabular ist. Mehr Wörter im Vokabular führen dagegen zu einer höheren Wahrscheinlichkeit, dass zwei Wörter ähnlich klingen und somit verwechselt werden können. Jedoch steigt meist auch Flexibilität und Funktionsvielfalt, so dass an dieser Stelle eine Abwägung von Usability/Utility auf der einen Seite und Erkennrate auf der anderen Seite getroffen werden muss.

Auch nachdem solche Entscheidungen getroffen wurden und ein Spracherkenner ausgewählt wurde, gibt es weitere Möglichkeiten, die Erkennrate des Systems zu verbessern.
Die wohl offensichtlichste ist das Sprachmodell. Mit dem Sprachmodell steuert der Entwickler, welche Worte von dem System erkannt werden sollen und damit direkt auch die Größe des Vokabulars. Doch nicht nur dessen Umfang ist von Bedeutung, auch sollten phonetische Ähnlichkeiten (einfach gesprochen: Wörter, die ähnlich klingen) berücksichtigt werden.
Eine weitere Möglichkeit zur Verbesserung der Erkennung kann durch die Veränderung oder Ergänzung der Phonemmodelle der Spracherkenner erreicht werden. Dabei kann für Worte in der Grammatik eine eigene Phonetik festgelegt oder eine vorhandene ergänzt werden. Dies macht zum Beispiel für fremdsprachige Begriffe wie „Play“ für einen deutschen Spracherkenner Sinn.
Darüber hinaus kann auch die kontinuierliche Erkennung einschränkt werden, da bei dieser immer unklar ist, wann eine Äußerung beginnt und wann sie eigentlich für das System und nicht für andere Gesprächspartner gedacht ist. Eine Lösungsmöglichkeit bietet sich in der Verwendung eines speziellen Knopfes („Push-To-Talk“ oder PTT), den der Nutzer drücken muss, um zu sprechen.

Sprachverstehen

Die Sprachverstehenskomponente analysiert die Ausgabe des Spracherkenners und weist ihm eine Bedeutungsrepräsentation zu, die vom Dialogmanager benutzt wird.

[4] Die Erkennrate ist das Verhältnis richtig erkannter Wörter zu der Gesamtzahl aller dem System mitgeteilten Wörter.

Traditionell ist Sprachverstehen ein natürlichsprachliches Verstehen, welches erst die syntaktische und semantische Struktur analysiert, um die Bedeutung des Gesagten zu ermitteln. Dabei müssen Ambiguitäten[5] beachtet werden und das Verfahren robust gegen Fehlerkennungen, Füllworte, Satzfragmente und Selbstkorrekturen sein.

In den meisten heutigen Systemen wird dies jedoch einfacher gelöst. Dort ist in den Grammatiken nicht nur enthalten, was für Worte zu verstehen sind, sondern auch in welchem Zusammenhang. Zusätzlich ermöglichen viele Grammatiken semantische Interpretation, d.h. dort kann bestimmten ähnlichen Phrasen jeweils gleiche Bedeutung zugeordnet werden.
Die Daten für den Aufbau solcher semantisch-fachspezifischer Grammatiken werden meist aus einem iterativen Prozess gewonnen, der üblicherweise aus Datensammlung, Überarbeitung der Grammatikregeln und einem neuen Test, mit dem neue Daten gesammelt werden, besteht. Mehr zu diesem Prozess, der sich auch auf andere Teile der Sprachsysteme beziehen kann, findet sich im Abschnitt 5.3 dieser Arbeit.
Zunehmend etablieren sich auch alternative Ansätze zu Grammatiken, vor allem bei Telefoniesystemen, die mit ihren laufenden Systemen riesige Datenmengen an Sprachdaten sammeln können. Dort werden, basierend auf so genannten statistischen Sprachmodellen, automatisiert Modelle zum Sprachverstehen erzeugt.

Durch die gemeinsame Verwendung von Elementen von Sprachverstehen und Spracherkennung in den Grammatiken wird deutlich, dass diese Prozesse eng zusammengehören. Oftmals wird auch die Spracherkennung und das Sprachverstehen gekoppelt, um mit Informationen aus dem Sprachverstehen den extrem schwierigen Prozess der Spracherkennung mit zusätzlichen Wissen zu versorgen.

Spracherzeugung

Nach der Verarbeitung durch die Dialogkomponente (welche später in diesem Abschnitt beschrieben wird) muss eine Ausgabe für den Nutzer in natürlicher Sprache konstruiert werden.
Üblicherweise wird diese natürliche Sprache mittels vorher definierter Lückentexte erzeugt, deren Lücken mit den Parametern, die die Dialogkomponente liefert, befüllt werden.

Es gibt auch einige Forschungsansätze, die Spracherzeugung als Prozess zu betrachten, der die Ausgabe unter Einbeziehung zusätzlicher Informationen generiert. So könnte beispielsweise ein Benutzermodell berücksichtigt und das System dadurch an den Benutzer adaptiert werden.

Sprachsynthese

Die so genannte TTS (Text-to-Speech)-Komponente wandelt die zuvor generierten Texte in gesprochene Sprache um. Dafür bedient sie sich zuerst einer Textanalyse, welche die Texte in ihrer syntaktischen und semantischen Struktur analysiert. Mit diesem Wissen wird dann die Sprache und ihre Prosodie (die Betonung) generiert.

[5] Eine Ambiguität ist eine Mehrdeutigkeit (Aus Lateinischem ambiguitas: Zweideutigkeit, Doppelsinn), von der gesprochen werden kann, wenn ein Zeichen mehrere Bedeutungen hat.

Falls diese automatischen Verfahren Fehler produzieren, ermöglichen die meisten TTS-Systeme, die manuell anders zu betonende Stellen zu markieren und eine Aussprache vorzugeben.

Dialogmanager

Ein Dialogmanager bedient sich aller bisher erwähnten Komponenten, koordiniert deren Ansteuerung, die Anfrage an externe Informationsquellen und steuert den Dialogablauf. Im Idealfall sollte er prinzipiell alle möglichen Modalitäten ansteuern können, also in einem multimodalen Dialog alle möglichen Systemein- und -ausgaben verarbeiten können.
Der Ablauf innerhalb eines Dialoges wird in der Dialogbeschreibung festgelegt, welche dabei auf verschiedene Arten definiert werden kann:

zustandsbasiert
: Der Nutzer wird durch eine Reihe vorbestimmter Schritte oder Zustände geführt. Der Dialogablauf besteht aus einer Menge von Dialogzuständen mit Zustandsübergängen, die diverse Wege durch den Dialoggraph definieren.

Schablonen-basiert
: In diesen Systemen wird der Nutzer solange vom System gefragt, bis alle Einträge einer intern vorgehaltenen Schablone gefüllt sind. Die Reihenfolge der nächsten Dialogschritte ist dabei nicht festgelegt, sondern wird je nach dem aktuellen Zustand des Systems ausgewählt. Die Fragen sind im System nur mit Informationen über ihre Startbedingungen abgelegt.

Agentensysteme
: Als Schnittstelle zu problemlösenden Anwendungen wird bei diesen Systemen Kommunikation wird als Interaktion zwischen zwei Agenten begriffen, in der jeder über seine Aktionen und Vorstellungen im Klaren ist und versucht, dies auch bei seinem Gegenüber zu erreichen. Dies entspricht ungefähr auch den Anforderungen, die die bereits definierten erweiterten Funktionen benötigen würden.

Während sowohl Konzepte für zustandbasierte als auch Schablonen-basierte Formen verbreitet benutzt werden, machen Agentensysteme vom Aufwand her nur Sinn, wenn die Verwendung der anderen Möglichkeiten prinzipiell nicht oder lediglich unter nicht vertretbarem Aufwand möglich ist.

Für die Dialogbeschreibung werden meist deklarative Markupsprachen eingesetzt, wie beispielsweise VoiceXML [MBC+04]. Diese können entweder per Hand oder durch entsprechende Werkzeuge erzeugt werden. Ein Vergleich verschiedener verbreiteter Beschreibungsformate und ihrer Interpreter findet sich in der Belegarbeit des Autors dieser Arbeit [Sch04].

Die beschriebenen Möglichkeiten der Dialogspezifikation widmen sich jedoch vor allem einer systemorientierten Sicht. Grundsätzlich ist aber ein Dialog etwas zweiseitiges, eine Abfolge von Äußerungen und Reaktionen darauf[6].Je nachdem, welche Sicht ein Sprachdialogsystem bevorzugt, nennt man es:

[6] Eine genauere Diskussion zu allgemeinen Definition des Begriffes Dialog im Umfeld von Sprachdialogsystemen findet sich unter anderem bei Hamerich [Ham00] oder ausführlicher bei McTear [McT04]

systemorientiert: Hauptsächlich Äußerungen des Computers, Nutzer reagiert.

nutzerorientiert: Hauptsächlich Äußerungen des Nutzers, Computer reagiert.

Halten sich beide Möglichkeiten in einem Dialog die Waage, wird dies **gemischter Initiative** genannt. [7]

Weitergehende Überlegungen zu Sprachschnittstellen, wie die der Universal Speech Interfaces [ROR01], formulieren Ideen, wie nutzerorientierte Systeme sinnvoll aufgebaut werden könnten. In der praktischen Anwendung sind die meisten heutigen Systeme aber entweder systemorientiert oder benutzen gemischte Initiative.

Diese Konzepte mit Leben zu füllen ist die Aufgabe des Dialogdesigns. Dieses leitet sich einerseits aus dem klassischen Interaktionsdesign für die Mensch-Maschine-Schnittstelle ab, andererseits fließen auch Untersuchungen über das Wesen des Dialoges zwischen zwei Menschen ein.
Dabei ist eine wesentliche Grundlage eines sprachlichen Dialoges zwischen zwei Menschen das Prinzip der gemeinsamen Grundlage, also der Kontextabhängigkeit des Gespräches. Wissen beispielsweise zwei Gesprächspartner, was sie unter dem „tollen Flug“ verstehen wollen, ist eine gemeinsame Grundlage für die folgende Äußerung gegeben: „Ich buche morgen diesen tollen Flug.“ Anderenfalls muss sie hergestellt werden, meist wird das durch Nachfragen gelöst.
Bei Sprachdialogsystemen resultiert diese fehlende gemeinsame Grundlage beispielsweise aus nicht perfekten Ergebnissen der Spracherkennung, fehlendem Weltwissen oder Diskrepanzen zwischen den angeforderten Informationen und den vorhandenen Daten in der externen Wissensquelle. Die Wiederherstellung der gemeinsamen Grundlage kann nun auf zweierlei Weisen durchgeführt werden. Falls das System den Fehler erkannt hat (wenn keine/unvollständige/fehlerhafte Eingabe), dann kann es durch Nachfrage beim Nutzer die Unklarheit beseitigen.
Allerdings kann es vorkommen, dass das System fälschlicherweise etwas erkannt und verarbeitet hat, was dem System richtig erschien, aber etwas anderes ist, als das, was der Nutzer sagen oder erreichen wollte. Diesem Fehler kann nur durch Verifikation, der Bestätigung der Daten durch den Nutzer, begegnet werden. Diese kann entweder explizit geschehen oder implizit, indem zum Beispiel in die nächste Frage des Systems die Daten des letzten Dialogschrittes eingebunden werden. Dialog 2.2 zeigt dies in einem Beispiel.

Dialog 2.2: Implizite Verifikation innerhalb eines Dialoges (nach [McT04])

usr: Ich möchte von Belfast nach London.
sys: Zu welcher Zeit möchten Sie von Belfast nach London fliegen?
usr: Um 7 Uhr am Morgen.

Aus der Kommunikation mit externen Quellen, der bereits erwähnten weiteren Aufgabe des Dialogmanagers, ergeben sich ebenfalls einige Problemstellungen. Eventuell können

[7] Diese Prinzipien sind nicht nur in Sprachsystemen bekannt, sondern allgemeine Prinzipien des Interaktionsdesigns (wie zum Beispiel in Preim [Pre99] zusammenhängend dargestellt)

Abbildungsprobleme zwischen Datenbank und Grammatik auftreten, wenn sich diese in Art oder Aufbau unterscheiden (zum Beispiel ist die Sprachsteuerung in deutsch, die Datenbank in englisch). In so einem Fall muss der Dialogmanager die Umwandlung übernehmen oder einer speziellen Informationsmanagerkomponente übergeben.
Weiterhin soll hier ein Problem nicht unerwähnt bleiben, das so auch in normalen Datenbanken auftritt, das Problem von über- und unterspezifizierten Anfragen. Überspezifizierte Anfragen erzeugen durch zu viele Einschränkungen in der Anfrage keine Treffer, unterspezifizierte durch zu wenig Einschränkungen viele Treffer.
Verschärfend kommt bei Sprachsystemen hinzu, dass im ersten Fall klargemacht werden muss, dass die Anfrage syntaktisch richtig, aber inhaltlich nicht zielführend war. Im zweiten Fall kann dem Nutzer eine solch große Liste von Treffern nicht angemessen in sprachlicher Form präsentiert werden, da sich eine vernünftige Anzahl Treffer, die der Nutzer auch im Kurzzeitgedächtnis halten kann, um 7 (+-2) Treffer bewegt.[Mil56]
Lösungsansätze bieten sich hier über eventuelle automatische Ergänzung von Informationen aus dem Kontext, die der Dialogmanager bereitstellen könnte, oder ein automatisches Reduzieren der Einschränkungen, bis die Anfrage einen Treffer liefert. Diese Behandlung kann entweder im Dialogmanager ausgeführt werden oder direkt an der Datenquelle durchgeführt werden. Wie das im Idealfall aussehen könnte, wird genauer in Kapitel 3.3.3 vorgestellt.

Der Dialogmanager muss aber sein Wissen nicht immer nur aus externen Quellen und den Ergebnissen des Sprachverstehens ableiten, vielmehr gibt es Überlegungen in der Forschung, ihm auch eigene Wissensquellen als Entscheidungsgrundlage für den Dialogablauf verwalten zu lassen.
So könnte er einen Dialogverlauf verwalten, welcher beispielsweise eine einfaches Zurückspringen in einen Dialogschritt zuvor ermöglichen könnte („Zurück-Kommando“). Durch Bereitstellung von Weltwissen oder eines Modells des Diskursbereichs („Navigation ist nur zu einem Ziel möglich“) ist es denkbar, mögliche Ambiguitäten und Fehlinterpretionen zu reduzieren. Auch ein Nutzermodell, welches entweder statisch (Alter, Geschlecht, Herkunft) oder dynamisch (bisherige Vorgehensweise) Wissen über den Nutzer bereitstellt, kann in diesem Zusammenhang hilfreich sein. Die Umsetzung solcher Funktionen wäre jedoch mit großem Aufwand verbunden, weswegen ihr Einsatz sorgfältig abgewogen werden müsste.

Zusammenfassend hat der Entwickler eine Reihe von Entscheidungen für den Dialogmanager zu treffen:

Dialoginitiative
: systemorientiert, nutzerorientiert oder gemischte Initiative.

Dialogkontrollstrategie
: zustandsbasiert, Schablonen-basiert oder agentenbasiert.

Verfikationsstrategie
: explizit oder implizit.

Fehlerbehebungstrategie
: Anzahl Nachfragen bei Nicht-Verstehen. Zeitpunkt Abbruch. Möglichkeiten zur Fehlererkennung.

Benutzung Barge-In
„Reinsprechen“ in die noch laufende Systemausgabe:
nicht möglich/nach PTT/Kontinuierlich

2.4 Sprachdialogsysteme im Auto

Um im weiteren Verlauf dieser Arbeit Sprachdialogsysteme im Auto diskutieren zu können, bietet es sich zunächst an, Besonderheiten des Einsatzes und Aufbaus von Sprachdialogsystemen im Auto zu betrachten.

Der Einsatz von Sprachdialogsystemen im Auto erscheint folgerichtig, wenn sich die in Abschnitt 2.1 diskutierten Einsatzszenarien von Cameron [Cam00] noch einmal ins Gedächtnis gerufen werden. Einer dieser Punkte würde laut Cameron schon ausreichen, um die Verwendung von Sprache zu rechtfertigen bzw. nicht zu behindern, bei Sprachdialogsystemen im Auto sind aber gleich drei davon erfüllt.
Die stetig wachsende Komplexität der Informations- und Entertainmentprodukte (kurz: Infotainmentsysteme) erfordert beispielsweise im Auto ein Bedienkonzept, dass es ermöglicht, die immer komplexeren Bedienvorgänge schneller als bisher durchführen zu können. Durch den Einsatz von Sprache ist nicht nur das möglich, die sprachliche Interaktion erlaubt es auch, sich weiterhin voll auf die Fahraufgabe zu konzentrieren und die Vielfalt der Funktionen auch während der Fahrt sicher und komfortabel zu bedienen. Durch die Anwendung im Auto, das nach außen schallisolierend wirkt, erlaubt es auch die Erhaltung der Privatsphäre.

Folgerichtig gibt es eine wachsende Anzahl von Funktionen im Auto, für die heute eine Sprachbedienung verfügbar ist. Begonnen mit einfachen Freisprecheinrichtungen, die allmählich durch weitere Telefonfunktionen erweitert wurden, kamen später Angebote für Navigation, Radio, CD/DVD-Player und fast aller restlichen Infotainment-Funktionen des Autos per Sprache hinzu. Hierbei werden diese Systeme entweder in das Originalzubehör (OEM = Original Equipment Manufacturer) des Automobilherstellers nahtlos integriert oder als Nachrüstlösung (After Market) angeboten. Einen genaueren Überblick, welche Arten von Systeme momentan angeboten werden, bietet [Han04].
Jedoch sind durchaus weitere Anwendungszwecke denkbar, zum Beispiel zur Erhöhung der Sicherheit beim Fahren. White et al. [WRS04] diskutieren Konzepte, in denen das Auto bemerkt, wenn der Fahrer langsam ermüdet (über eventuell eingebaute biometrische Sensoren), und darauf hin ein allgemeines Gespräch initiiert oder ihn mit Spielen unterhält.

Alle diese Anwendungen müssen allerdings mit einigen spezifischen Einschränkungen des Einsatzes im Auto umgehen. So zeigen White et al [WRS04] eine Reihe von Problemfeldern auf:
Nutzer solcher Systeme verbringen meist viel Zeit in ihren Autos und sind dementsprechend sensibel, wenn das System nicht nach ihren Vorstellungen reagiert oder sie vielleicht sogar nervt. Verschärfend kommt hinzu, dass sich mit dem Einbau im Fahrzeug der Nutzerkreis auf verschiedenste Personen mit den unterschiedlichsten Hintergründen (Erfahrung, Dialekt, Bildung) gegenüber früheren Sprachanwendungen erweitert, die eher für eine schmale, eng-begrenzte Zielgruppe erstellt wurden. Und natürlich haben diese Nutzer hohe Erwartungen, geprägt durch Science Fiction Serien wie „Star Trek“ oder „Knight Rider“.

Dem stehen auf technischer Seite zahlreiche Einschränkungen gegenüber.
Durch den Einbau im Auto ergeben sich durch Geräusche des Motors und der Straße, Blinkertöne, Geräusche der Mitfahrer u.a. zahlreiche Störgeräusche. Diese senken das Signal-Rausch-Verhältnis, machen also das Signal, was der Spracherkenner verarbeiten soll, schwächer. Dieses Problem kann durch die Verwendung von nah am Sprecher angebrachte, direkte, störungsreduzierende Mikrophone reduziert werden. Aus Platz- und Kostengründen ist das nicht in jedem Auto möglich.
Aber es gibt auch Software-Lösungen für dieses Problem: So kann das Audiosignal vorverarbeitet werden um die Störungen herauszurechnen. Weiterhin kann ein erweitertes akustisches Modell verwendet oder mehr Aufwand in die Algorithmen der Spracherkennung gesteckt werden. Die eingeschränkte Leistung von Prozessor und Speicher von üblicherweise im Auto verfügbaren Systemen begrenzt jedoch die Nutzung dieser Lösungen. Diese Limitierung erfolgt aus den üblichen Kosten-, Größen- und Langlebigkeitsanforderungen, die Automobilhersteller an die verwendeten technischen Systeme stellen.
Die Kunst der Erstellung eines solchen Systems besteht nun darin, die Hardwareanforderungen und die Komplexität der Lösungen der oben diskutierten Probleme in eine vernünftige Balance zu bekommen.

Wie allerdings im Abschnitt 2.3 diskutiert, existieren auch andere Möglichkeiten, die Leistung des Sprachdialogsystems zu verbessern. So kann als weitere Möglichkeit versucht werden, Unzulänglichkeiten der Spracherkennung über gutes Dialogdesign zu vermeiden oder zu umgehen.[8]
Eine Möglichkeit wäre direkte, also zielorientierte Dialoge zu verwenden, welche den Kontext ihrer Ausführung geschickt mitnutzen. Ebenfalls erfolgversprechend ist, Rückmeldungen des Systems von Stärke des empfangenen Signals oder der Konfidenz der Spracherkennung abhängig zu machen. Dies entspricht dem menschlichen Dialog, in dem bei Unsicherheit nachgefragt wird. In diesem System führt es dazu, dass durch die genauen Fragen eine stärkere Fokussierung eintritt und somit Fehler vermieden werden können.
Und schließlich hilft es auch sehr, wenn die Dialoge für die Nutzer intuitiv sind, sie also ins Auto steigen und sofort mit der Bedienung loslegen können. Solche intuitiven Dialoge reduzieren Fehler, erhöhen Zufriedenheit und verzichten auf festgelegte Sätze, die sich der Nutzer gar noch einprägen müsste. Eine weitere Verbesserung könnte sich an dieser Stelle ergeben, wenn der Dialog darüber hinaus noch adaptiv ist, sich also nach Angaben in der History oder mithilfe von biometrischen Daten an den Nutzer und seine Strategien und Wording anpasst.

2.5 Abgrenzung des Diskursbereichs

Nachdem in den vorangegangen Abschnitten erläutert wurde, was Sprachdialogsysteme sind, wo sie eingesetzt werden, wie sie technisch aufgebaut sind und welche Besonderheiten sich im Automobilbereich ergeben, soll nun eine Abgrenzung des Diskursbereichs der Arbeit im Bereich der Sprachdialogsysteme vorgenommen werden.

In Abschnitt 2.1 wurden eine Reihe von möglichen Funktionen eines Sprachdialogsystems

[8] Grundsätze für gutes sprachliches Dialogdesign sowohl allgemeine als auch im Auto formuliert Abschnitt .

diskutiert. Für das Anwendungsbeispiel MP3-Player sind dabei insbesondere der Informationsabruf und die Gerätesteuerung interessant. Aufgrund der komplexeren Struktur muss ein vollwertiger Dialog verwendet werden.
Die Spezifikation dieses Dialogs soll dabei den Schwerpunkt dieser Arbeit bilden. Die in Abschnitt 2.3 vorgenommene Kapselung des Dialogmanagers von den restlichen Komponenten des Sprachdialogsystems ist dabei hilfreich. Die Komponenten der Spracherkennung oder -synthese sollten dabei nur benutzt, aber nicht verändert werden.

Allerdings ist es möglich, über einige der ebenfalls in Abschnitt 2.3 diskutierten Parameter Einfluss auf diese Komponenten auszuüben. In Anlehnung an Abbildung 2.2 veranschaulicht Abbildung 2.3 welche Werte durch das Dialogdesign weiterhin beeinflussbar (schwarz) und welche durch die Art der Anwendung bereits festgelegt sind (rot).

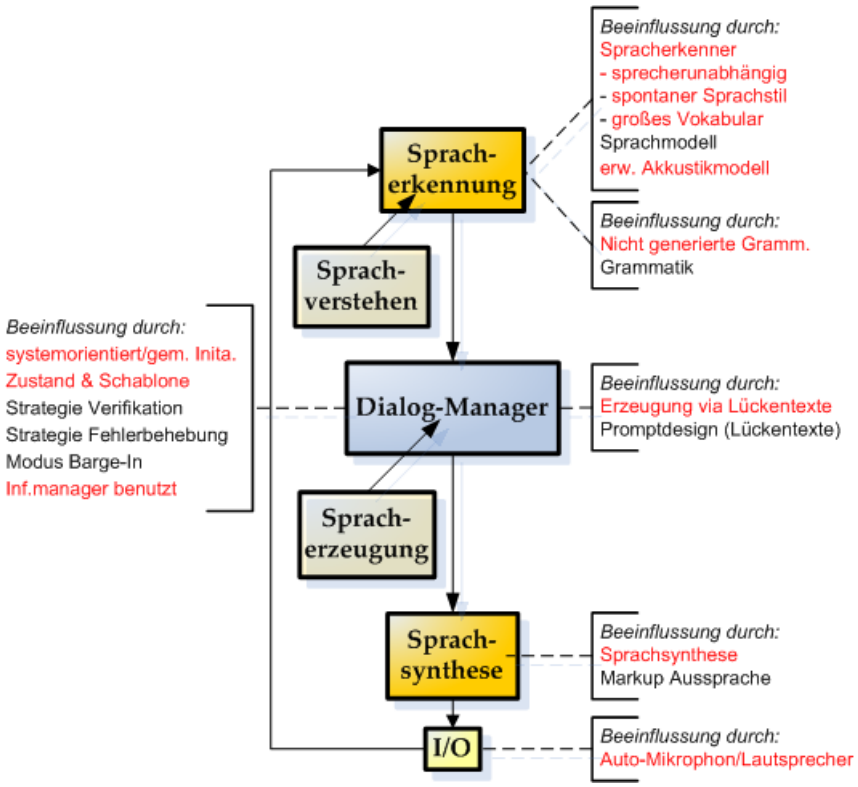

Abbildung 2.3: Konkrete Einflussgrößen für Dialogdesigner im Kontext der Arbeit

So brachte die Entwicklung für die Anwendung im Auto mit sich, dass ein specherunabhängiger Spracherkenner gewählt wurde, da im Auto keine Lernphase denkbar ist und auch die sprecherunabhängige Erkennung im Auto erwartet wird. Aus diesem Fakt folgte auch direkt, dass ein eher spontaner bzw. umgangssprachlicher Sprachstil angenommen und der Spracherkenner mit den entsprechenden Sprecherdaten trainiert wurde.
Das Ziel der Entwicklung des sprachgesteuerten MP3-Player bestand nun auch darin, Titel aus der MP3-Datenbank direkt auszuwählen. Dafür müssen die Einträge dieser Datenbank sprechbar sein, es wird also einen Spracherkenner benötigt, welcher mit großen bis sehr großen Vokabular umgehen kann. Und für den Einsatz im Auto waren auch erweiterte Akustikmodelle bereits vorhanden und nutzbar.

Dies alles sind Anforderungen unter den im Abschnitt 2.4 definierten technischen Einschränkungen, die Anforderungen an den Spracherkenner waren also sehr hoch.

Für die Komponenten des Sprachverstehens und der Spracherzeugung wurden jeweils keine erzeugenden Systeme verwendet, da dieser Aufwand nicht angemessen schien und auch so kaum kommerziell üblich ist.[9] So wurde die Funktionalität des Sprachverstehens in die Spracherkennung integriert, während die Spracherzeugung durch den Dialogmanager durchgeführt wurde. Deswegen ist im Weiteren, wenn von Spracherkennung die Rede ist, auch immer das Sprachverstehen gemeint, und mit einem Dialog soll auch immer Spracherzeugung gemeint sein.
Bei der Sprachsynthese wurde ein Paket benutzt, für welches auch eine den Hardware-Anforderungen im Auto prinzipiell entsprechende Version zur Verfügung stand und die Ein- und Ausgabegeräte im Fahrzeug waren auch im Wesentlichen vorgegeben.

Durch die Verwendung der firmeneigenen Dialogbeschreibungssprache GDML (mehr dazu in Abschnitt 7.2.2) war bei dem Dialogmanager eine zustandsbasierte oder Schablonenbasierte Modellierung vorgegeben, ebenso ermöglichte diese Sprache keine rein nutzerzentrierten Dialoge. Weiterhin konnte viel Logik des Zugriffs auf externe Informationsquellen durch Benutzung einer einfache Schnittstelle zu diesen (ebenfalls mehr in 7.2.2) ausgelagert werden.

Beeinflussbar für das Dialogdesign blieben also im Wesentlichen die Grammatik, die Systemausgaben (die so genannten Prompts und deren eventueller Markup) und der Dialogablauf, der auch durch Entscheidungen zur Verifikations- und Fehlerbehebungsstrategie und Barge-In-Benutzung beeinflusst wurde.

[9] Außerdem ist eine Anforderung der Automobilfirmen, dass an jeder Stelle des Dialogs klar ist, was konkret sprechbar ist. Dies wäre mit einem generativen System des Sprachverstehens so nicht möglich.

3
Musik

In diesem Kapitel soll zunächst versucht werden, Musik im Kontext dieser Arbeit einzuordnen und zu erläutern, wie diese beschrieben werden kann. Aufbauend darauf werden verschiedene Formen von Metadaten für Musik diskutiert, welche benutzt werden können, um Musik auszuwählen. Dafür werden Möglichkeiten diskutiert, auch unter den speziellen Umständen im Auto und bei der Benutzung von Sprache. Abschließend werden die Erkenntnisse des Kapitels noch einmal in Abschnitt 3.4 zusammengefasst.

3.1 Was ist Musik?

„Über Musik zu sprechen ist wie über Architektur zu tanzen.“
Steve Martin

Dieses Zitat des amerikanischen Schauspielers Steve Martin drückt das Grundproblem aus, das bei jedem Definitionsversuch von Musik entsteht. Egal ob physikalische, musiktheoretische, perzeptionelle, geschichtliche oder philosophische Ansätze gewählt werden, die Definition wird immer nur Teilbereiche beschreiben. Selbst die deutsche Ausgabe der freien Enzyklopädie Wikipedia scheitert an einem Definitionsversuch („Eine genaue Bestimmung, was Musik ist und was nicht, ist nicht möglich.“, [Wik05b]).
Das Grundproblem beginnt schon meist damit, Musik und Geräusche zu trennen. Jean-Jacques Nattiez stellt in seinem Werk „Toward a Semiology of Music“ [Nat90] fest: „Die Grenze zwischen Musik und Geräuschen ist immer kulturell definiert – was impliziert, dass diese Grenze sogar in einer einzigen Gesellschaft nicht an der selben Stelle gezogen wird; kurz gesagt, es gibt selten einen Konsens. Unter Berücksichtigung aller Quellen gibt es kein allein stehendes, interkulturelles und universelles Konzept, was definiert, was Musik sein möge.“
In Umgehung dieser Problematik soll Musik in Rahmen dieser Arbeit als Abfolge von Tönen mit besonderen Charakteristiken betrachtet werden, die intuitiv vom Hörer als Musik erkannt wird. Für eine tiefer gehende Betrachtung sollte das erwähnte Werk dem interessierten Leser einen Einstieg in die Thematik bieten.

Viel wichtiger im Kontext dieser Arbeit ist aber, wie sich Musik voneinander unterscheidet, wie sie geordnet werden kann und wie sie vom Einzelnen geordnet wird.
Um Musik unterscheiden zu können, muss der Mensch sie erst einmal wahrnehmen, also die physikalischen Schwingungen des Schalls in die Sinneswahrnehmung von Tönen, Melodien und Rhythmen umzuwandeln. Dabei werden die Nerven des Innenohres von den Schwingungen des Luftdrucks (die als Frequenz bzw. Amplitude beschrieben werden können) stimuliert. Im Verlauf dieses Prozesses verschmelzen bestimmte Frequenzen des Schalls zu Tönen, Melodien entstehen durch erst Maskierungseffekte verschiedener Töne und Ermüdungseffekte der Nervenzellen im Innenohr ermöglichen die Wahrnehmung von Rhythmen. Eine genaue Beschreibung dieses Vorgangs geht weit über den Rahmen dieser Arbeit hinaus, einen umfassenden und auch sehr anschaulich beschriebenen Einblick in die Thematik bietet z.B. Jourdain [Jou01] an.

Auf der physikalischen Ebene ist Musik für Menschen noch nicht fassbar, also auch nicht unterscheidbar. Nach der Wahrnehmung der Töne kann dann ein Mensch zwar Musik hören, aber meist immer noch nicht verbal fassen. Die undenkbar vielen Kombinationen von verschiedenen Melodien, Harmonien, Rhythmen und auch Inhalte der Stücke machen es schwer, für jedes Stück ein Wort zu finden. Das wäre aber nötig, um dem Hörer die Möglichkeit zu geben, zu beschreiben, was er mag, was er gerade hört oder was ein Musiker jetzt spielen soll. Eine Beschreibung auf dieser Ebene wäre immer sehr subjektiv und ungenau. Trotzdem hat der Mensch im Laufe der Jahren immer wieder versucht, dies über Kategorisierungen, Genre und ähnliches, also über die Ähnlichkeit, zumindestens Gruppen von Musikstücken zu beschreiben.[1] Ein Grundproblem dabei ist nun, dass jede Person in jeder Situation Ähnlichkeit anders begreifen kann. So ist es möglich Ähnlichkeit als Maß der Gleichheit der Instrumentierung, des Stil, des Inhalts, globaler Maße wie Melodie oder Rhythmus, emotionaler Reaktion oder schlicht als willkürliches Maß zu sehen[PBZA04].
Das Benutzen solcher unscharfer Kriterien soll im Weiteren unscharfe Musikkategorisierung genannt werden.
Diese funktioniert leidlich für Gruppen von Musikstücken, um einzelne Stücke zu beschreiben, wurden jedoch sehr bald Hilfsgrößen eingeführt.

So sind schon in der Antike Musikstücke durch einen Namen beschrieben wurden [Mic81]. Denn sollte Musik über Improvisationen hinaus gehen, musste ein Weg gefunden werden, diese Wiederholbarkeit nicht nur durch schriftliche Fixierung, sondern auch durch eine einfache Benennung zu ermöglichen.
Diese Benennung reichte für Volkslieder vollkommen aus. Dies ermöglichte die exakte Zuordnung von einem Namen zu einem Musikstück. Als es später nötig wurde, zu unterscheiden, wer ein Stück vorgetragen hatte, kam die Information über den Interpreten hinzu. Erst in neuerer Zeit entwickelte sich mit dem Aufkommen von Tonträgern (am Anfang also der Schallplatte) eine weitere Bescheibungsform, das Musikalbum. Dies bündelte mehrere Stücke eines Künstlers in einem oder mehreren Tonträgern, was als Gesamtheit wieder einen Namen bekam.
Auf dieser Ebene von Interpret - Album - Titel ist zwar die Beziehung zu Musik nicht

[1] Eine interessante philosophische Frage wäre hierbei, warum Musikstücke, welche „unbeschreiblich schön“ sind, für viele Menschen perfekte Musik auszeichnen (eine Google-Suche nach „Musik unbeschreiblich schön“ liefert beispielsweise allein 123.000 Treffer), aber Menschen trotzdem sofort immer wieder nach einer Beschreibungsmöglichkeit suchen.

abgeleitet aus der konkreten Musik, sondern eher aus den Begleitumständen, aber dafür genau und (in der großen Mehrzahl der Fälle) eindeutig. Solche Namen können ein Stück nur sehr unvollkommen beschreiben, da sie nicht im Ansatz die Informationsfülle widerspiegeln können, die in einem Musikstück enthalten ist. Daher ist das Finden des Namens für Musikstück und Album (und in neuster Zeit auch schon des Interpretennamens) ein künstlerischer Akt, der zu der Kunst des Musikschaffens dazugehört.
Ich will diese Beschreibungsform im Rahmen der Arbeit explizite Musikkategorisierung nennen.

In Abbildung 3.1 wird diese grundlegende Unterscheidung zwischen Musiksignal, Musikwahrnehmung und Musikbeschreibung visualisiert und menschliche Beschreibungsmöglichkeiten den einzelnen Ebenen zugeordnet.

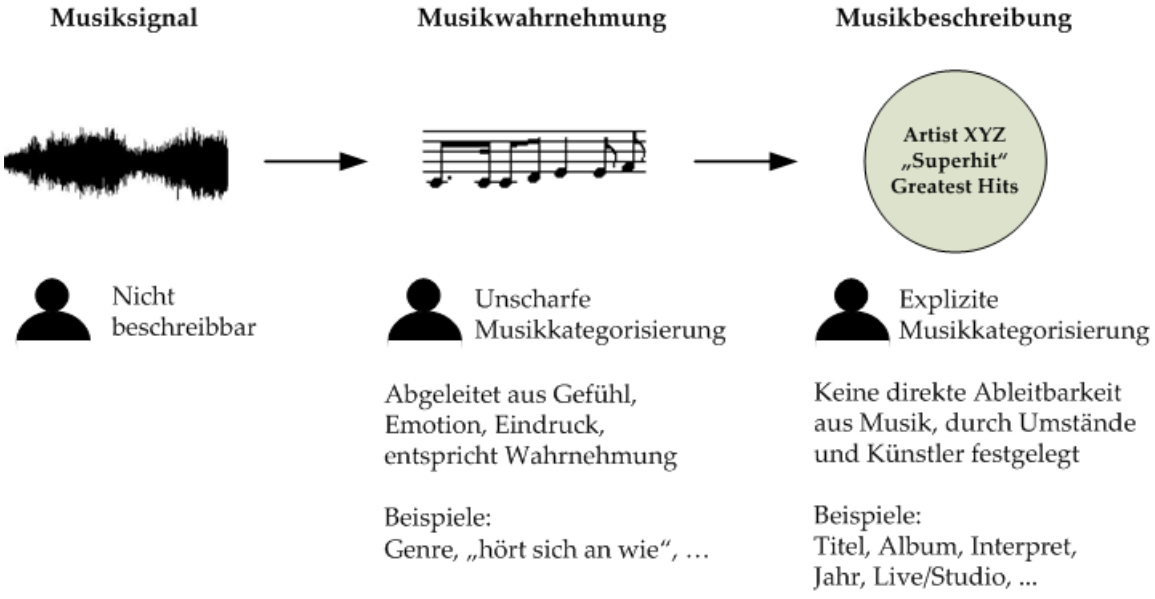

Abbildung 3.1: Unterscheidung und menschliche Beschreibungsmöglichkeiten Musiksignal, Musikwahrnehmung und Musikbeschreibung

Neben diesen prinzipiellen Problemen bei der Benennung und Kategorisierung von Musik gibt es allerdings auch noch Faktoren, die diese Grundprobleme verschärfen.
So hat sich in den letzten Jahren und Jahrzehnten der Zugang zu Musik grundlegend verändert. Hatten die Menschen am Anfang des 20. Jahrhundert meist nur die Möglichkeit, Musik entweder selbst zu machen oder bei seltenen Konzerten oder Auftritten zu erleben, veränderte die Speicherbarkeit von Musik dies vollständig. So wurde Musik immer verfügbarer, sowohl von der Menge (die durch den Verkauf der Tonträger und den damit entstehenden Verdienstmöglichkeiten um ein Vielfaches anstieg), als auch vom Ort (erst zuhause, dann auch unterwegs). Zwar gab es jetzt immer noch Momente von Livemusik, doch wurden diese für die diesen neuen Massenmedium Musik ausgesetzten Menschen immer mehr zu einer raren Ausnahme.
Mit der Masse und Ortsunabhängigkeit wuchs aber auch der Wunsch nach Orientierung in diesem unübersichtlichen Wust an Veröffentlichungen. Diese Orientierung boten wiederum andere Massenmedien, zuerst Radio und Musikmagazine, ab den 80er Jahren dann auch

das Musikfernsehen. Daneben gab es zwar immer auch andere Möglichkeiten der Information (Hinweis von Freunden, „Stöbern“ im Plattenladen), doch boten die niemals die Breite der Einordnung, die durch die Massenmedien angeboten wurde. Erstmal gekauft, wurden die Tonträger meist liebevoll zuhause in eine eigene Ordnung gebracht, was meist noch angesichts der überschaubaren Menge möglich war.

Erst das Aufkommen von MP3 um 1995 sollte dies grundlegend ändern. Dieses Kompressionsverfahren für Musik [Ans00] ermöglichte es, plötzlich Musik in annehmbarer Qualität auf einen Bruchteil der bisher benötigen Datenmenge zu verkleinern. Dies ermöglichte, zusammen mit sprunghaften Verbreitung des Internets gegen Ende der 90er Jahre, den massenhaften Tausch von Musikdaten, und innerhalb weniger Jahre den Aufbau von teilweise riesigen Musiksammlungen auf Geräten der Nutzern.
In diesen Sammlungen boten nun meist die klassischen Massenmedien kaum noch Orientierungshilfe, da die Daten nicht so gut strukturiert und geordnet waren, wie dies vorher von den im Laden gekauften Alben und Singles bekannt war. Für viele Nutzer endet der Rausch der Downloads in einem MP3-Datenfriedhof. Die Unübersichtlichkeit des Angebotes an Musik, die sie bisher vielleicht nur aus dem Plattenladen kannten, war bei den Musikhörern zuhause angekommen. Nur das jetzt auch noch die Beschriftungen fehlten.

3.2 Metadaten

Als Metadaten oder Metainformationen werden allgemein Daten bezeichnet, die Informationen über andere Daten enthalten [Tim97]. Musik-Metadaten im Speziellen enthalten meist Informationen über den Autor, Album und Titel, häufig aber sogar viel mehr Informationen wie beispielsweise Genre, Jahr oder Veröffentlichungsdatum. Diese Daten bilden quasi das Äquivalent zum traditionellen Label auf den Schallplatten bzw. dem Album-Cover, auf dem Informationen zu der Musik beschrieben wurden. Metadaten bieten so einen Ansatz zur Orientierung im Chaos der großen MP3-Sammlungen, wenngleich sie nicht die einzige Möglichkeit sind.

So kann auch eine Ordnung über Dateinamen und Verzeichnisse herstellt werden, wie dies auch heute noch durchaus eine gebräuchliche Praxis ist (wie teilweise die Erkenntnisse aus Abschnitt 6.1 und 6.2 bestätigen). Doch ist dieses Vorgehen auf wenige Informationen beschränkt (meist Autor-Album-Titel), zusätzliche Informationen können nicht hinzugefügt werden. Weiterhin ist eine Umsortierung in einer solchen hierarchischen Struktur sehr umständlich.

Solche Probleme haben die Metadaten nicht, da hier die Struktur des Speicherortes der Musikdateien keinen Einfluss auf die Informationen hat. Auch lassen sich leicht ganz Kategorien auswechseln, ohne das eine Struktur zerstört wird.
Als Beispiel für solche Metadaten sollen nun die so genannten ID3-Tags [Nil00] betrachtet werden, ID3 steht dabei für „Identify an MP3“. Grundsätzlich existiert ID3 in zwei Versionen:
ID3v1 definiert einen festen, 128 Byte großen Block am Ende einer MP3-Datei, der feste Felder für Informationen zu Titel, Album, Interpret, Jahr und Genre enthält, begrenzt auf 30 Zeichen pro Eintrag. Weiterhin können noch Kommentare anfügt werden.
ID3v2 löste diese starre Struktur auf, so werden die Zeichen-Begrenzungen für die Einzel-

felder deutlich angehoben, und eine Vielzahl von möglichen Informationen zugelassen, die jedoch bei weitem nicht von allen Programmen unterstützt werden. So bietet ID3v2 die Möglichkeit, fast beliebig genau Informationen zu einem Titel zu spezifizieren.
Allerdings, um zu dem Bild von weiter oben zurückzukehren, ebenso wie nicht alle Label und Cover wirklich aussagekräftig sind, hängt der Nutzwert von ID3-Tags weniger von der Struktur, als vielmehr davon ab, welche Inhalten in diesen Tags stehen und wie diese gefüllt werden.

So können diese Tags manuell gefüllt werden, doch ist dies sehr aufwendig, fehleranfällig und fast gänzlich zeitlich undurchführbar bei größeren Musikbeständen. Dort kommt nur eine automatisierte Erzeugung in Frage.
Für solche Zwecke existieren mit der CDDB [Gra06] und der freedb [fre06] weltweit verfügbare Datenbanken, die Metadaten zu fast allen veröffentlichten Alben bieten. Bei Zugriff wird mithilfe einer eindeutigen ID, die sich aus einer Prüfsumme der Startsektoren und Längen der Albumtitel zusammensetzt, der Eintrag in einer meist im Internet verfügbaren Datenbank ermittelt und die Informationen in die Metadaten des Albums übertragen. Der Vorteil hierbei ist, dass die Information zu diesem Album weltweit nur einmal manuell eingetragen werden muss.
Pachet et al. [PBZA04] zeigen allerdings auf, dass solche expliziten Metadaten noch sehr viel mehr enthalten können. So hinterfragen sie die Trennung von Interpret und Komponist in den klassischen Konzepten der ID3-Tags, und bieten dafür das Konzept der primären und sekundären Interpreten an. Primäre sind dabei solche Interpreten, die Nutzer spontan zuerst einem Musikstück zurechnen würden (bei klassischer Musik der Komponist, bei einer Coverversion der Coverinterpret), sekundäre dementsprechend was danach dem Stück zuordnet würde (Klassik: Interpret, Orchester; Cover: Originalinterpret). Fraglich bleibt, ob diese Unterscheidung in den Metadaten modelliert werden muss, oder ob sie nicht im Dialog eingebunden werden kann (durch Gleichsetzung von Interpret und Komponist und Suche in beiden Feldern).
Weiterhin schlagen Pachet et al. vor, Interpreten generell ausführlicher zu modellieren, in so genannten MHEs, Musical Human Entitys. In denen soll festgehalten werden, ob ein Interpret ein Einzelkünstler, Teil einer Band oder vielleicht Duett-Singer ist. Das würde zum Beispiel ermöglichen, bei einer Suche nach Phil Collins auch Stücke von Genesis zu finden. Ob dieser Nutzen den Aufwand wirklich lohnt, ist allerdings fraglich.

Neben dieser Einspeisung von Informationen für die explizite Musikkategorisierung gibt es auch Möglichkeiten, die Informationen über die Ähnlichkeit der Musik zu erfassen bzw. automatisiert zu erzeugen. Dabei geht es vor allem um die Erzeugung unscharfer Musikkategorisierung. Diese automatischen Verfahren sollen also meist helfen, entdeckendes Hören, ein „Spiele mir mehr wie das“-Feature oder eine Lieblingsmusik-Playliste zu ermöglichen.
Dabei unterscheiden Pachet et al. die Zuordnung auf Ebene der kulturellen Hintergründe und der Klangfarbe.

Auf kultureller Ebene kann eine solche unscharfe Kategorisierung beispielsweise durch das Prinzip des gemeinschaftlichen Filterns stattfinden [PR06]. Dieses Prinzip, welches prinzipiell wie Mundpropaganda funktioniert und auf der Einbeziehung einer Community beruht, bedeutet die Einbeziehung von Vorlieben anderer Nutzer zur Vorhersage eigener Vorlieben. Die Nutzer, die dafür einbezogen werden sollen, werden über Ähnlichkeit in den Vorlieben ermittelt. Neue Metadaten werden dabei also nicht erzeugt, sondern lediglich

vorhandene Daten besser genutzt. Voraussetzung für einen solchen Dienst ist aber natürlich eine Community, was zwingend eine Datenleitung zu anderen Nutzern voraussetzt. Ein Beispiel für die Anwendung dieses Prinzips in der Praxis stellt last.fm [Las06] dar, wo über ein solches gemeinschaftliches Filtern bestimmt wird, welches nächste Musikstück im so genannten „neighbour radio“ gespielt werden soll. So soll ermöglicht werden, neue Musik zu entdecken, die einem potentiell gefällt, aber eventuell noch nicht bekannt ist.
last.fm bietet darüber hinaus noch weitere Möglichkeiten. So ist es möglich, einem Lied direkt frei wählbare Begriffe (Tags) zuzuordnen, mit denen Klassen oder Genres ähnlicher Musik gebildet werden können. Das könnte eine Lösung für das Problem der prinzipiellen Unmöglichkeit allgemein akzeptierter Unterteilung in Musikrichtungen (siehe [AP03]) darstellen, zumindestens aber eine Minderung der Problematik bewirken. Zwar ist dies bei last.fm nur für die Datenbank der Musik möglich (direktes Abspielen ist wohl aus Lizenzgründen nicht möglich) und die Idee der Tags ganz offensichtlich von der Foto-Plattform Flickr [Yah06] entliehen, trotzdem zeigt es auf, welche Möglichkeiten der Anreicherung mit Metadaten in diesen Techniken stecken.

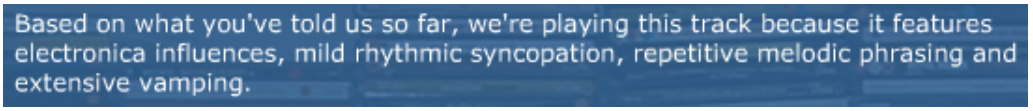

Abbildung 3.2: Beschreibung Selektionskriterium bei Pandora [PM06]

Pachet et al. nennen auch noch weitere Möglichkeiten zur Zuordnung auf kultureller Ebene. So wird ein Web-Mining-Ansatz (auf welcher Webseite stehen welche Namen zusammen) geschildert und die Möglichkeit einer Auswertung von Radioplaylisten, die dann ein eher metaphorisches Bild auf Ähnlichkeit eröffnen würden, beschrieben.

Auf Ebene der Klangfarbe, also der akustischen Eigenschaften kann nun auf jegliche Eingaben oder Analysen des Nutzers verzichtet werden.
So können Menschen aus reinem Schall keine Systematik ableiten, wie bereits in Abschnitt 3.1 diskutiert. Doch diese Möglichkeit besitzen Computer und vielfach wird eine solche objektive Kategorisierung als weiterer Ausweg aus dem bereits angesprochenen Dilemma der fehlenden allgemeine Kategorisierung von Musikrichtungen angesehen.
Das einfachste Vorgehen besteht darin, physikalische Faktoren des Musiksignals wie Mittelwerte und Varianz der Amplituden, Spektalanalysen und ähnliches durchzuführen. Mit MPEG7 [Mar04] steht auch ein technisches Format bereit, in das solche Information eingespeist werden könnten.
Doch sagen solche rein physikalischen Eigenschaften doch erschreckend wenig über die Ähnlichkeit von Musik aus, entsprechend gibt es verschiedene Anstrengungen, die menschliche Wahrnehmung in Teilbereichen nachzubilden, und nach den dort gewonnenen Parametern wie Rhythmus, Klangfarbe oder ähnlichem zu selektieren. So wird in [ZPDG02] ein so genannter „Rhytm Extractor“ beschrieben, welcher versucht die Rhythmusstruktur eines Stückes zu ermitteln. Eine weitere interessante Möglichkeit wäre die Bewertung wahrgenommener Energie, also der Unterschied der wahrgenommenen Intensität zwischen beispielsweise Heavy Metal und Volksmusik [PBZA04].
Eine andere Möglichkeit, wenn auch viel aufwendiger, ist die Klassifikation der Daten manuell bzw. halbautomatisch vorzunehmen, und spezielle Charakteristika wie Stimmlage, Rhythmus, typische Instrumente, Textinhalt für Musikstück in definierten Skalen zu

bewerten. Diesen Weg sind die Entwickler von Pandora [PM06] gegangen, nach eigenen Angaben steckten sie 5 Jahre Arbeit in dieses Projekt. Das Ergebnis in Form eines adaptiven Webradios, das nach der Eingabe von nur einem Lieblingslied fast immer ähnliche Musik abspielt, ist freilich beeindruckend. Nicht ein Genre oder Stil ist ausschlaggebend, sondern die charakteristischen Eigenschaften der Musik (Abbildung 3.2 zeigt solche Auswahlkritierien, wie sie vom Pandora-System angezeigt werden). Zu diskutieren wäre, ob eine solche Datenbank in 10 Jahren noch aktuell wäre, und wie auch Nutzer an der Weiterentwicklung und dem Feintuning des Systems beteiligt werden könnten.
Darüber hinaus könnte auch versucht werden, einem klassischen Problem Herr zu werden. Viele Menschen prägen sich zwar den Refrain eines Liedes ein, aber können sich nicht den Titel (der nicht zwangsweise etwas mit dem Refrain zu tun hat) merken. Eine Lösung für dieses Problem wird von Baumann und Klüter [BK02] vorgeschlagen, sie stellen ein System vor, welches aus dem Signal eines Musikstückes Liedtexte extrahiert, den Text des Refrain erkennt und in Form von Metadaten zur Verfügung stellt.

In Abbildung 3.3 wird nun nochmals zusammenfassend dargestellt, welche Arten von Musikmetadaten vorgestellt wurden und aus welchen Datenquellen sie ihre Informationen beziehen.

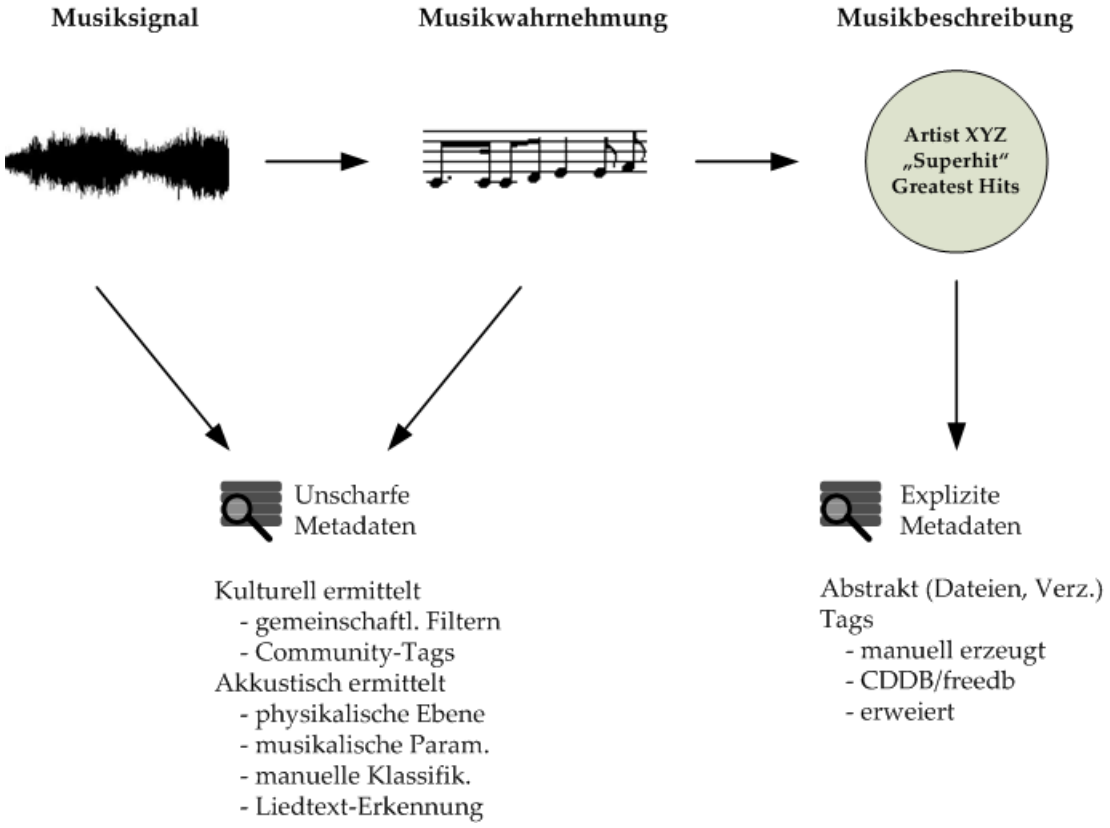

Abbildung 3.3: Arten von und Datenquellen für Musikmetadaten

3.3 Musikauswahl

Wollen Menschen beschreiben, welche Musik sie mögen, was sie gerade hören wollen oder über welche Musik sie gerade reden, müssen sie sich für eine Art der Beschreibung entscheiden. So könnte sie explizite oder unscharfe Beschreibungsformen dafür nutzen, beide Formen mischen oder gar ganz anderes an diese Aufgabe herangehen. Auch ist für viele Menschen Musik jeweils etwas völlig anderes, ihr Vorgehen eher aus ihren Erfahrungen und Vorlieben ab. Die Herausforderung besteht nun darin, eine Musikauswahl zu finden, die diesem Durcheinander von verschiedensten Musikrepräsentationen entspricht.
In diesem Abschnitt werden nun einige bisherige Arbeiten zum Thema Musikauswahl vorgestellt, bevor danach im folgenden Abschnitt Schlussfolgerungen für das Design formuliert werden.

3.3.1 Allgemein

Mit Musikauswahl aus digitalen Beständen haben viele Nutzer seit Jahren Erfahrung, schließlich machen sie mit ihren Programmen am Computer oder ihrem mobilen Gerät unterwegs nichts anders.

Zunächst lohnt ein Blick auf einige bereits auf dem Markt befindlichen Software-Lösungen und Geräte.
In einer Arbeit von Haegler und Ulmer [HU04] wurde eine solcher Marktüberblick vorgenommen und für vier gebräuchliche mobile MP3-Player der Funktionsumfang analysiert und folgende typische Funktionen identifiziert:

Wiedergabesteuerung
: Funktionen mit Steuerungscharakter der Musikwiedergabe, dazu zählen neben Wiedergabe, Pause und Stopp auch die Abspielmodi wie Zufallswiedergabe und Wiederholung.

Lautstärke- und Equalizersteuerung
: Manuelle und adaptive Steuerung. Gewisse Einstellungen können meist auch in so genannten Presets gespeichert werden.

Information
: Abfrage aller verfügbaren Informationen zu den aktuellen Titel sind jederzeit verfügbar, durch den großen Bildschirm kann eine große Informationsmenge angezeigt werden.

Auswahl über Metadaten
: Benutzung der ID3-Tags zur Auswahl von Titeln.

Dateinavigation
: Eine Möglichkeit auch auf nicht mit Metadaten versehene Daten zuzugreifen. Verwendet eine Ansicht von Verzeichnissen und Dateien.

Wiedergabelisten
: Werden von allen Playern abgespielt. Einige unterstützen auch das Erstellen dieser Listen.

Diese Funktionen werden auch von den meisten Software-Programmen unterstützt, wenngleich natürlich mit unterschiedlichen Akzentsetzungen.
Wichtiger als der reine Funktionsumfang ist sicher die Struktur der Auswahl. Dazu stellen Haegler und Ulmer als Gemeinsamkeit die Navigation über eine Art Cursorsteuerung fest, die es jeweils ermöglicht, aus Listen auszuwählen. Diese Listen entsprechen meist jeweils einer Hierarchiestufe (entweder Genre, Interpret, Album, Titel oder gewähltes Verzeichnis), aus der in eine weitere Stufe verzweigt, oder teilweise auch direkt abgespielt werden kann. Ich möchte dies an dieser Stelle hierarchische Auswahl nennen.[2]
Kontrastierend dazu besteht die Möglichkeit, auch Musik unabhängig von einer gegebenen Hierarchie durch Suchen zu finden. Diese Funktion, welche meist nur bei Software-Player besteht, wird meist über ein einfaches Suchfeld bereitgestellt (z.B. in der Winamp Libary [Win06]). Dabei wird einfach in den vorhandenen Daten gesucht, eine Eingrenzung auf gewisse Teilbereiche ist dabei meist auch möglich (zum Beispiel die Suche nur in den Interpreten). In der weiteren Diskussion möchte ich das explizite Suche nennen. Diese relativ freie Art der Musikauswahl verlangt vom Nutzer immer noch, sehr genau zu wissen, was er konkret hören möchte.

Mit den Anforderungen der Nutzer und der Benutzerfreundlichkeit beschäftigten sich Pachet at al. in einer grundsätzlichen Arbeit zum Thema Musikauswahl [PBZA04]. Darin unterscheiden sie im Wesentlichen zwei Typen von Musikhörern: Die Stöberer und die Bibliothekare. Während Erstere eher unbewusst durch ihre Sammlung streifen, neues Entdecken wollen, sind Letztere sich genau im Klaren, wie ihre Sammlung organisiert ist und möchten anhand dieser Struktur ganz genau ihre gewählte Musik auswählen. Später soll auf diese zwei Typen (die auch eine Person in verschiedenen Situationen darstellen kann) noch zurückgekommen werden.

In der Arbeit beschreiben Pachet et al. das Prinzip und den Aufbau eines EMD-System, ein System für „Electronic Music Distribution“ (also für Musikverkauf über das Netz). Jedoch sind viele ihrer Erkenntnisse auch auf die Musikauswahl für die eigene Musik-Sammlung interessant. So begreifen sie Musikauswahl als etwas anderes als einfachen Zugriff. Sie beschreiben Musikauswahl nicht als rationalen Prozess, denn nicht nur, dass die Nutzer nicht wissen, wie sie auf ihre Musik zugreifen sollen, oft wissen sie auch gar nicht, was sie wirklich suchen. Dies ist nicht nur im Bereich des elektronischen Musikvertriebs der Fall, sondern, wie Pachet et al. ausdrücklich hinweisen, auch in vielen privaten Sammlungen. In den Tausenden von Liedern ist es für Nutzer unmöglich, wirklich jedes Lied daraus zu kennen.
Weiterhin gehen sie davon aus, dass solche Personen wohl vor allem Musikliebhaber sind, denn wer keine größere Ansprüche hat, hört meist nur Radio. Weiterhin unterstellen sie, dass es das hauptsächliche Ziel dieser Nutzer ist, unbekannte, aber ihren musikalischen Vorlieben entsprechende Stücke zu finden. Ziel sollte also sein, „Aha“-Erlebnisse zu Produzieren, Freude am Entdecken zu vermitteln, und das unabhängig, wie ungenau, oder falsch die Angabe vielleicht war.
Einen Anhalt für diese These liefert eine Untersuchung über die Benutzung des iTunes Music Sharing [VGD+05], was seit Version 4.0 bei iTunes möglich ist. Dabei können Musikdateien der lokalen Datenbank freigegeben werden und von iTunes-Benutzern, die sich

[2] Für genauere Information finden sich in der erwähnten Arbeit [HU04] auch vollständige Dialogcharts der untersuchten MP3-Player.

im selben Teilnetz befinden, angehört werden. Bei der Untersuchung der Nutzung dieser Technologie durch Voida et al. ergaben sich einige Implikationen für den allgemeinen Zugriff auf Musik-Datenbanken. So gaben die Teilnehmer dieser Studie an, dass sie ganz bewußt das System benutzten, um neue Musik zu entdecken. Einer merkte dazu an, seine Motivation resultiere daraus, dass er „kein musikalischer Spießer werden wollte“. Es scheint also wirklich eine Tendenz zu geben, stöbernd auf Musiksammlungen zuzugreifen.
Und betrachtet im Licht des in Abschnitt 3.1 diskutierten Bedeutungsverlustes der Massenmedien und daraus resultierenden Unsicherheit beim Auffinden neuer interessanter Musik, ist dieses Argument schließlich nachvollziehbar.

Neben solchen Untersuchungen zu Zielen und Charakeristika von potentiellen Nutzern solcher Systeme lohnt natürlich auch der Blick auf Gedanken zum User Interface für diese. Wie Pachet et al. feststellen, sind dabei die meisten der heute verfügbaren User Interfaces in ihrer Auswahlfunktion entweder sehr genau oder sehr explorativ, wobei sicherlich die ersteren die große Mehrzahl darstellen. Nun könnte einfach jeweils ein Interface für explorative Musikauswahl für Stöberer und ein explizites für Bibliothekare aufgebaut werden (wie in Abbildung 3.4 gezeigt), doch besser sollte der Nutzer nicht zwischen zwei Geräten oder auch nur zwei Modi wählen können, vielmehr soll einfach die Menge an Information, die der Nutzer dem System mitteilt, darüber entscheiden, welcher Modus benutzt wird (Abbildung 3.5).

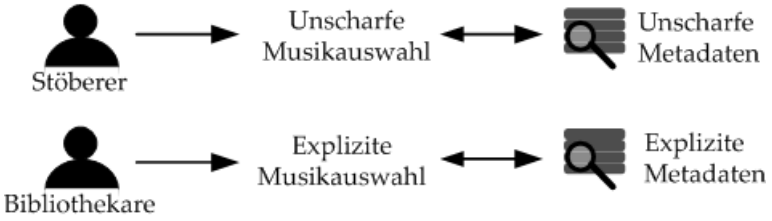

Abbildung 3.4: Im einfachsten Fall: Zusammenhang zwischen Benutzergruppen, Musikauswahlmethoden und Metadaten

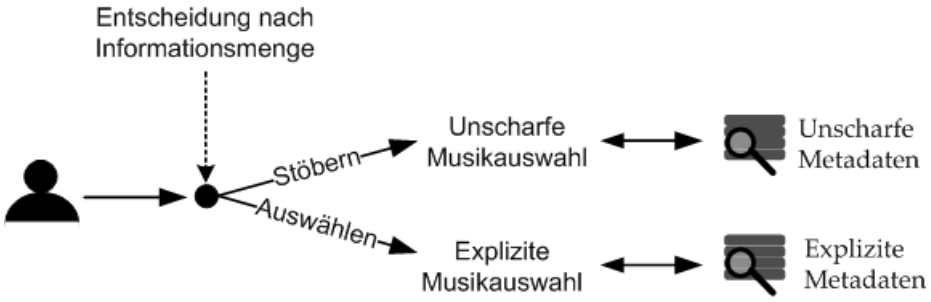

Abbildung 3.5: Bei automatischer Wahl: Zusammenhang des Modus der Musikauswahl zwischen Benutzergruppen, Musikauswahlmethoden und Metadaten

Weitere Ansätze zum „erlebbar machen“ von Musik bieten die beiden schon in Abschnitt 3.2 erwähnten Webradios last.fm [Las06] und Pandora [PM06]. Beide ermöglichen nicht nur durch ihr innovatives Metadaten-Konzept einen einfachen Zugriff auf Musik, auch kann der Nutzer durch ein Bewertungssystem („Titel gefällt mir/gefällt mir nicht“) die Auswahl der Musik beeinflussen. Denkbar wäre, dieses System in ein Nutzermodell zu in-

tegrieren und bei Unklarheiten in der Musikauswahl als mögliches Entscheidungskriterium mitzubenutzen.

3.3.2 Im Auto

Mobile Musikspieler gibt es schon seit Jahrzehnten, von den ersten tragbaren Plattenspielern über den Walkman bis zum Discman bildeten sie eine Grundlage auch für Entwicklungen für den Automobilbereich. Anfangs kam dort die Entwicklung immer etwas später in Gang, anscheinend genügten noch bis Mitte der neunziger Jahre Radio und Kassettenlaufwerk den Ansprüchen der meisten Autofahrer an Musik im Auto. Erst danach setzten sich auch nach und nach CD-Player durch.[Wik05a]
Mit dem Aufkommen mobiler MP3-Player nach 2000 entstand nun auch recht bald der Wunsch, diese Funktion ins Auto zu bringen, was sich auch relativ leicht in CD-Player integrieren ließ. Damit wurde es erstmals möglich, MP3s im Auto zu hören, wenngleich die Kapazität einer CD beschränkt war.
Mit dem Aufkommen von Festplatten-Playern (insbesondere des iPod) und der langsam beginnenden Verbreitung von MP3-Handys[3] ergeben sich jedoch auch umfangreichere weitere Möglichkeiten mobiler Musik. So ermöglicht die Speicherkapazität dieser mobilen Geräte, einen Großteil der eigenen MP3-Sammlung immer mobil dabei zu haben und darauf immer gleich mit dem entsprechenden Gerät zugreifen zu können. Musik wird mobil, überall zugriffsbereit und allverfügbar.
Die Möglichkeit, diese Alleskönner direkt im Auto anschließen zu können, wird dabei immer verlockender und zunehmend auch von den Herstellern oder Nachrüstern angeboten.

Die technischen Grundlagen für einen Zugriff auf MP3-Datenbanken im Auto sind also geschaffen, doch welche Einstellungen haben speziell Autofahrer zu Musik im Auto? Auf diese Frage konnte leider aus der Literatur keine ausreichende Antworten gefunden werden, Akesson und Nilsson halten in einer Untersuchung über Pendler [AN02] z.B. lediglich fest, das Autofahren je nach Ziel (zur Arbeit, nach Hause) andere Wünsche an Funktionen hervorbringt, z.B. Entertainmentfunktionen meist auf dem Weg nach Hause benutzt werden. Ein weiteres Puzzlestück ergibt sich durch ein Blick auf eine Webseite wie halfbakery.com [Hal06]. Dort werden halbfertige Ideen diskutiert, darunter auch Ideen für Musik im Auto. Viele Wünsche der Nutzer dort gingen stark in Richtung Personalisierung, so wünschten sich einige Nutzer Musik, die direkt zu ihrem Auto passt bzw. mit der sie ihrem Auto eine persönliche Note geben könnten.

Alles in allem war aber keine dieser Erkenntnisse aussagekräftig genug, um die weiter oben aufgeworfene Frage zu beantworten. Deswegen wurden eine Reihe von Fragen in die Umfrage im Rahmen dieser Arbeit integriert (siehe Abschnitt 6.1).

3.3.3 Sprachgesteuert

Sprachsteuerung für Musikauswahl wird zwar seit einigen Jahren in der Forschung thematisiert, bisher gibt es jedoch kaum Produkte, bei denen diese auch eingesetzt wird. Eine Konzeptstudie, wie ein solches Produkt aussehen könnte, veröffentlichte Infineon zwar

[3] So wünschen sich nach einer Umfrage der Zeitung 'connect' 52,3 Prozent aller Nutzer ein MP3-Player in ihrem nächsten Handy.[Pre06]

schon 2002 mit einem in eine Jacke integrierten MP3-Player mit Sprachsteuerung [IP02], bis heute sind aber alle Produkte von der mittlerweile ausgegliederten Firma Interactive Wear ohne eine solche Sprachsteuerung erschienen. Auch gibt es auf dem Handymarkt neben einigen Ankündigungen kaum MP3-Handys, die ihre MP3-Funktion per Sprache bedienen lassen.[4] Es stellt sich die Frage, warum Sprachtechnologie sich an dieser Stelle noch nicht durchgesetzt hat.

Ein Hauptgrund besteht sicher darin, dass Spracheingabe in vielen Situationen keinen Mehrwert bietet, da die Integration mit dem grafisch-haptischen Benutzerinterface nur halbherzig angegangen wurde. Wie bereits in Abschnitt 2.1 diskutiert, ist dieses aber eine der Grundvoraussetzungen, dass eine solche zusätzliche Modalität Mehrwert bringt.
Ein Problem, welches bei der Umsetzung einer solchen Sprachauswahl meistens bewältigt werden muss, ist das riesige Vokabular, das nötig ist, um die Auswahl über die Metadaten der Musikdateien möglich zu machen. Da es durch die Flüchtigkeit der Sprache schwierig ist, Listen darzustellen, muss dafür entweder eine geschickte multimodale Kombination gefunden oder versucht werden, die Anzeige von Listen generell zu vermeiden.

Natürlich hat die Verwendung von Sprache auch Vorteile, so kann mit Sprache die Genauigkeit der Anfrage an das System stufenlos reguliert werden. Diese Problematik wurde bereits im Abschnitt 3.3.1 thematisiert, angewandt auf Musik bedeutet dies, dass je nach Menge der spezifizierten Informationen über einen Titel, ein Album oder ähnliches, entweder eine direkte Auswahl ausgeführt wird oder passende Nachfragen zu fehlender Information gestellt werden. In einem grafisch-haptischen System ist sowas zwar heute meist auch möglich, natürlicher erscheint es aber mit Sprache.
Eine weiterer Vorteil ist, dass nicht nur verbale Äußerungen zur Musikauswahl genutzt werden müssen. So gibt es umfangreiche Forschungsanstrengungen zu Query-by-Humming, der Auswahl von Musik durch Summen der Melodie. Schon 1995 wurde ein solches System angedacht [GLCS95], mittlerweile gibt es auch erste Produkte dazu, so die Melodiesuche von musicline.de [mus06b], die vom Fraunhofer Institut für Digitale Medientechnologie entwickelt wurde.[5] Dabei wird die Aufnahme des Summens vorverarbeitet und dann mit relevanten Teilen der gespeicherten Musikdaten verglichen. Da gerade in diesem Zusammenhang die Definition von Relevanz äußerst schwer fällt, wird ein Großteil der Güte eines solchen Systems dadurch bestimmt, wie gut die Musikanalyse der Datenbank funktioniert. Doch haben Nutzer auch prinzipielle Schwierigkeiten, Rhythmus überhaupt adäquat wiederzugeben, so das eine Charakeristik ableitbar wäre [PBZA04].

Ein weiteres wichtiges Gebiet, in dem die Verwendung von Sprache für die Musikauswahl angemessen erscheint, ist die Nutzung im Auto. Diese wird im nun folgenden Abschnitt beschrieben.

[4] Lediglich eine Zusatzsoftware von Microsoft für PocketPc-Handys, VoiceCommand [Mic06], macht dies zur Zeit möglich.

[5] Ähnliche Dienste stellen der O_2 MusicSpy [Pre04] und Vodafone-MusicFinder [VK05] dar, die allerdings auf die Verarbeitung eines Signals aus Lautsprechern optimiert sind, zum Beispiel zum Herausfinden des Namens eines aktuellen Titels auf einer Party. Insofern sind diese Dienste für Sprachsysteme nicht direkt relevant.

3.3.4 Sprachgesteuert im Auto

Musik im Auto zu hören ist heute normal für die meisten Leute [AN02], trotzdem bedeutet dies noch nicht, dass sie von der Musikauswahl nicht abgelenkt werden. Wenn Unfälle durch Ablenkung verursacht werden, ist nach einer Untersuchung von Stutts et al. in 11% der Fälle die Bedienung des Audiosystems im Auto die Ursache dafür [SRSR01]. Seit dieser Studie im Jahr 2001 hat sich der Anteil potentiell ablenkender Audiotechnik im Auto nocheinmal deutlich erhöht, und auch die zusätzliche Komplexität der Navigation und Selektion aus großen Musiksammlungen während des Fahren geben Anlass zur Sorge, wie Forlines et al. festhalten [FSNR+05].
Wie schon in Abschnitt 2.4 diskutiert, bietet sich hier Sprache als Lösung an. So stellen Forlines et al. insbesondere den Vorteil des Direktzugriffes auf große Datenmengen im Kontext von Musikauswahl heraus. Diese Suche wäre bei der grafisch-haptischen Ausführung entweder mit einer großen Hierarchie bzw. langen Listen oder einer Texteingabe verbunden. Während ersteres meist sehr umständlich ist (da das Vorlesen langer Listen aufgrund von Flüchtigkeit der Sprache und der Begrenztheit des Kurzzeitgedächtnisses nicht sinnvoll erscheint), wirft letzteres die Frage auf, wie eine Eingabe geschehen sollte. Da eine Tastatur im Auto nicht angemessen erscheint, ist die Eingabe von Text mit Hilfe von vorhandenen Bedienelementen offensichtlich schwierig.
Sprache ermöglicht hier eine freiere Eingaben, wenngleich meist immer noch eine Syntax von Kommandowörter gelernt werden muss. Alternativ wäre hier natürlich auch wieder eine Hierarchie denkbar, deren Struktur ganz einfach die möglichen Äußerungen einschränken würde, was aber wieder den eigentlichen Vorteilen von Sprache von Natürlichkeit und Kontextfreiheit zuwiderlaufen würde. Mit diesem Paradoxum beschäftigen sich einige der nachfolgenden Forschungsarbeiten.

So werden bei McGlaun et al. [MAR+01] ein hierarchie- und ein (eher freierer) kommandobasierter Ansatz zur Musikauswahl im Auto verglichen, allerdings unter verschärften technischen Einschränkungen einer Entwicklung für ein Niedrigpreissystem (so war es nur möglich 30-50 Worte gleichzeitig zu erkennen, Synonyme waren so kaum möglich). Dabei waren aber in dem kommandobasierten System kein direktes Sprechen der Titel u.ä. möglich, im Nutzertest ergab sich dann eine klare Präferenz für das hierarchie-basierende System.
Im Kontrast dazu steht der „Speech In, List Out“-Ansatz von Forlines et al. [FSNR+05], denen ein eher „Google-artiges“ Interface vorschwebte. So wollten sie auch die Struktur der Interpretation der Nutzeräußerung auflösen. Dies erreichten sie, in dem sie die Trennung zwischen Dialog und Spracherkennung aufgaben, und nicht mit schon interpretierten Werten die Datenbank abfragten, sondern stattdessen mit einem Bündel von gewichteten Wörtern (mit jeweils einem Wahrscheinlichkeitsvektor) die Anfrage an die Datenbank stellten. Ergebnis war immer jeweils eine Liste, die nach Wahrscheinlichkeiten sortiert war, aus welcher der Nutzer auswählen konnte. Diese Liste wurde immer präsentiert, egal wie falsch die Angabe war, so dass die Ausgabe immer von der Struktur immer gleich war (wenn auch die Sortierung der Liste manchmal komplett falsch war)[6]. Weiterhin wurde die Liste nur grafisch präsentiert, so dass jede weitere Initiative immer beim Nutzer lag. In einer Evaluation (unter Einbeziehung einer Fahrsimulation als Ablenkung) gegenüber einem üblichen, Hierarchie-basierten grafischen Interface zeigte sich dieser Dialog als über-

[6] Eine anderen Arbeit, die dieses Prinzip in einem anderen Anwendungsgebiet (Finden von Mietwohnungen) auf die Spitze treibt, findet sich bei [OP02]

legen, die kognitive Belastung sank.

Ist das nun ein Widerspruch? Beide Arbeiten scheinen grundsätzlich gegenteilige Entwicklungen zu befürworten, Nutzertests bestätigen jeweils beide Ansätze. Zwar ist bei McGlaun et al. [MAR+01] sicher auch die technische Beschränkung der Spracherkennung und damit die fehlenden Synonyme bzw. die schon diskutierte fehlende direkte Auswahl über Tags Gründe für die Entscheidung für das Hierarchie-basierte System, doch gibt es sicher auch andere Gründe:
So ist der Ansatz von Forlines et al. [FSNR+05] sehr gut geeignet, um schnell, aber auch ungenau einen einfachen Zugriff auf die Musiksammlung zu geben[7]. Aber manche Menschen haben vielleicht doch lieber eine bekannte Struktur, um schnell und effektiv zu ihrer Musik zu finden. Bei der Benutzung einer Hierarchie sind die Vorteile und Nachteile dann jeweils vertauscht.[8] Es ist zu vermuten, dass eine Synthese aus beiden Ansätzen Erfolg versprechen könnte.
Dieses wurde im Rahmen der Erstellung eines Konzept-Autos für Ford [PDB+03] versucht, wo dem Nutzer drei Arten des Dialogs zur Verfügung standen.
So existiert hier ein systemgeführter Dialog, welcher auf einer festen Hierarchie basiert. Doch erfahrenere Nutzer können auch die Informationen in einem Kommando dem System mitteilen, so also eine gewisse Art von Suche anstoßen. Weiterhin kann auch mit natürlicher Sprache mit dem System interagiert werden. Die Erstellung eines solchen Systems ist allerdings sehr aufwendig, umfangreiche Nutzertests müssen vorher durchgeführt werden, um diese drei Methoden eng zu verzahnen und Widersprüche zu vermeiden, wie Forliners et al. einwenden. Weiterhin ist zu bedenken, dass dieses System so nicht im Auto zu realisieren wäre, da Hardwareanforderungen hier weitestgehend ausgeblendet wurden.
Ein solches auch an Hardware-Anforderungen orientiertes System diskutieren Wang et al. [WHHS05], es ist auch auf üblicherweise vorhandener Hardware im Auto lauffähig. Der Dialog, dessen Design hier aus Ergebnissen einer Nutzerbefragung und allgemeinen Dialogkritierien abgeleitet wurde, unterschied hierbei zwischen Play- und einem Browse-Kommandos. Ausgehend von einer angenommenen natürlichen Hierarchie (Genre → Interpret → Album → Titel) ermöglicht ein Browse-Kommando, aus den Unterkategorien einer gewählten Kategorie auszuwählen, während ein Play-Kommando direkt die gewählte Auswahl abspielt. Mit diesen Modi soll verschiedenen Benutzergruppen Genüge getan werden, eine Evaluation von angenommener Struktur und Unterscheidbarkeit der Modi stand allerdings noch aus, ebenso eine Evaluation des Gesamtsystems.

Somit lässt sich eine klare Aussage zur Struktur noch nicht abschließend formulieren, es gibt aber Anhaltspunkte, bei einem MP3-Dialog im Auto die Bedürfnisse verschiedene Nutzergruppen zu bedenken und diese Ansprüche in einem System konsistent und bedienbar zu vereinen. Dies kann trotz des hohen Aufwands nur durch eine starke Nutzerorientierung im Entwicklungsprozess geschehen, entsprechende Methoden werden im Abschnitt 5.3 diskutiert und im weiteren Verlauf der Arbeit auch angewendet (siehe Abschnitt 6). Dabei sollen Erwartungen und Ideen der Nutzer für die Benutzung einer MP3-Anwendung erfasst

[7] Wenngleich die Vermischung von Dialog und Spracherkennung sich problematisch darstellt, da die Austauschbarkeit von Komponenten nicht mehr gegeben ist. Außerdem funktioniert das System nur für Anfragen, einen Dialog kann damit nicht erstellt werden.

[8] Diese Unterteilung in Menschen, die schnell, ohne Anstrengung, aber nicht übermäßig genau Musik auswählen und solche, die immer exakt das haben wollen, entspricht ungefähr der Einteilung in Stöberer und Bibliothekar, wie sie in Abschnitt 3.3.1 vorgenommen wurde. Denn auch hier geht es um die Frage, ob ein Titel ganz genau ausgewählt wird, oder die Auswahl eher nur ungefähr richtig ist.

werden. Dabei wird auch versucht zu bestimmen, in welchen Anteilen typische Benutzergruppen solche Systeme bedienen und ob bereits diskutierte Unterteilungen überhaupt sinnvoll sind.

Abgesehen von der allgemeinen Struktur lassen sich aber auch Erkenntnisse zu einzelnen Punkten formulieren. So werden in allen hier diskutierten Systemen ausschließlich MP3-Tags als Basis der Auswahl benutzt, eine Auswahl über das Dateisystem ist nicht vorgesehen. Daraus folgt, dass aber auch keine unscharfen Metadaten benutzt werden, wie sie in Abschnitt 3.2 diskutiert wurden. So wird zwar in allen Systemen bisher der Gedanke der unscharfen Auswahl thematisiert, der Schritt zum entdeckenden Zugriff aber, wie in Abschnitt 3.3.1 diskutiert, nicht gegangen.
Ebenfalls nicht diskutiert wird die Verwendung von Community-Funktionen, wie sie ebenfalls in Abschnitt 3.3.1 erwähnt wurden. Das hat jedoch unter anderen praktische Gründe, da zur Nutzung dieser Funktionen eine ständige Netzverbindung nötig ist, die bei heutigen Autos noch nicht verfügbar ist.
Dagegen widmen sich die vorgestellten Arbeiten zur Musikauswahl im Auto breit der Tatsache, dass Automobilsysteme gewöhnlich multimodale Systeme sind. So betonen sie, dass alles, was ein Nutzer auf dem Bildschirm sieht, auch sprechbar ist[9] und ein Wechsel der Modalität jederzeit möglich sein sollte. Wie auch Forliners et al. betonen Pieraccini et al. die Ansicht, dass der Dialog vor allem nutzerorientiert sein soll, also keine Aufforderung zum Sprechen vom System kommen soll.

Insbesondere Wang et al. diskutieren auch Probleme, die speziell bei Musikauswahl mit Sprache auftreten. So stellt die Kontextfreiheit von Titel-, Album- und teilweise Interpretnamen ein großes Problem dar. Einerseits können dort (ungebräuchliche) Abkürzungen oder Slang vorkommen („Femme like U“), viel schwerer wiegt aber, dass (selbst für das Beispiel des englischen Dialogs) auch fremdsprachliche Titel gebräuchlich sind, zum Beispiel spanische oder französische Titelnamen. Verschärfend zu dem Problem, dass der Spracherkenner plötzlich zwei oder mehr Sprachen zu verstehen hat, kommt die Tatsache hinzu, dass viele Nutzer sicher nicht einmal selbst wissen, wie ein solcher Titel richtig ausgesprochen wird und dabei Fehler produzieren werden. Diesen Problemen widmen sich eine Reihe von Forschungsanstrengungen ([Com01],[GMN04]) Eine grundsätzliche Lösung ist aber momentan nicht abzusehen.
Verschärfend kommt das Problem der fehlenden Teilstring-Erkennung hinzu. So haben Titel oft Zusätze wie „All Cried Out (Unplugged)“ oder „Ka-Ching! (Red Disc)“ oder nicht funktionale wie „The Shoop Shoop Song (It´s In His Kiss)”. Diese wird ein Nutzer nicht in jedem Fall aussprechen, aber eine entsprechende Erkennung erwarten. Auch Interpreten wie „Herbert Grönemeyer“, die allgemein auch mal als nur „Grönemeyer“ bekannt sind, führen zu solchen Situation. Zuverlässige Erkennung würde hier nur eine echte Teilstring-Erkennung bieten, die allerdings auch noch nicht zur Verfügung steht.

In einigen Arbeiten werden auch grundlegend andere, weitergehende Visionen formuliert. Rist [Ris04] diskutiert beispielsweise die Möglichkeit, die Musikauswahl an Affekt und physiologischer Verfassung auszurichten. Dadurch möchte er zum Beispiel erreichen, dass bei Schläfrigkeit des Fahrers belebende Musik gespielt wird. Ein solches Vorgehen wäre auch etwas völlig anderes als der bewußte Akt der Musikauswahl, wie er bisher hier betrachtet

[9] Wie grafische Bildschirmausgaben so anpasst werden können, dass auch die semantische Struktur für Spracheingaben implizit vorgegeben ist, wird in [Wan03] diskutiert.

wurde. Viele Probleme wie plötzliche Musik zur Unzeit, die Erzeugung von geeigneten Metadaten oder die Kombination mit expliziter Musikauswahl wären hierbei zu lösen. Trotzdem zeigt es auf, wohin die Entwicklung laufen könnte.
Die Gedanken in eine ganz andere Richtung zu lenken ermöglicht ein Projekt an der TU-Dresden [JSF05]. Dort wurde ein System geschaffen, welches es ermöglicht, durch „Reinrufen“ den weiteren Ablauf eines Films im Kino zu beeinflussen. Diese Abstimmungssituation gibt es auch häufig im Auto, wenn unklar ist, auf welches Stück sich die Insassen einigen wollen, so dass sich eine Kombination dieses Prinzips mit der Musikauswahl im Auto als Versuch anbieten würde. Die technischen Voraussetzungen dafür (Sprechererididentifizierung, gleichzeitige Erkennung von mehreren Phrasen) müssten allerdings erst geschaffen werden.

3.4 Schlussfolgerungen

Nachdem in diesem Kapitel nun umfassend Musik, Metadaten und Musikauswahl diskutiert wurden, soll nun versucht werden, Schlussfolgerungen für die Dialogentwicklung des sprachgesteuerten MP3-Players im Auto zu formulieren.
So rückt zunächst die Struktur des Auswahldialogs als kontrovers diskutiertes Thema in den Mittelpunkt. In Abschnitt 3.3.1 wurden anhand bereits im Markt befindlicher Geräte und Software hierarchie-basierte und explizite Suche unterschieden. Zusätzlich wurde ein Ansatz von Pachet et al. [PBZA04] vorgestellt, auch ohne Wissen über konkrete explizite Daten quasi entdeckend Musik auszuwählen, dies soll als implizite Suche bezeichnet werden. Beispiele dafür boten die weiteren Abschnitte, in denen z.B. Query-By-Humming oder die automatische Auswahl in den personalisierten Webradios vorgestellt wurde. Wie in Abbildung 3.6 im Kontrast zu der in Abschnitt 3.3.1 diskutierten vereinfachten Abbildung 3.4 deutlich wird, ist diese Form auch die einzigste, die explizit unscharfe Metadaten erfordert.
Solche unscharfen Metadaten standen jedoch im Rahmen der Arbeit nicht zur Verfügung, so dass die implizite Suche, trotz hier diskutierter viel versprechender Ansätze, nicht umgesetzt werden konnte.

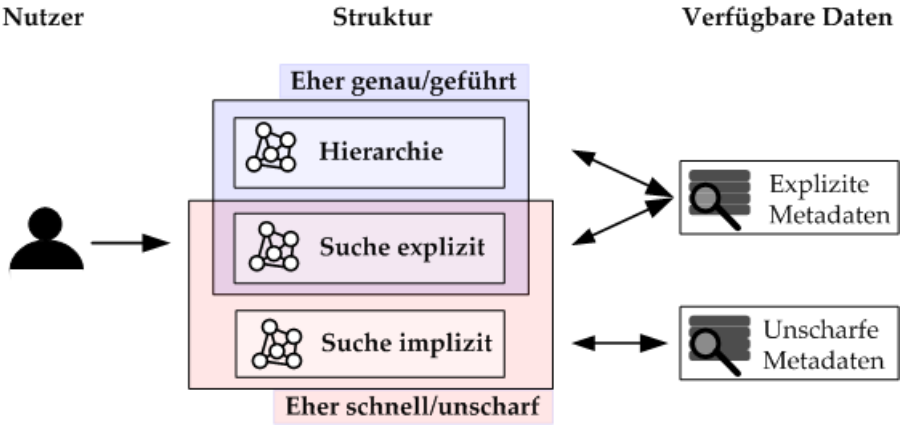

Abbildung 3.6: Die möglichen Grundformen der Dialogstruktur für Musikauswahl

Aus der Abbildung wird ersichtlich, dass eine Suche über explizite Metadaten sowohl einer genauen und geführten Auswahl entsprechen kann, aber auch benutzt werden kann, um

eher entdeckend eine Suche durchzuführen.
Der Nachteil hierarchischer Auswahl (vor allem aus der Verwendung von Listen) bei Sprachbedienung wird in Abschnitt 3.3.4 diskutiert. Trotzdem folgt aus der Diskussion um verschiedene Nutzungsstrategien, dass auch ein solch strukturierter Ansatz verfügbar sein sollte.
Überhaupt muss eine mögliche Struktur die verschiedenen Benutzergruppen berücksichtigen. Dazu wird nicht, wie in einigen Ansätzen, eine bestimmte Benutzergruppe angesprochen bzw. angenommen, sondern mit freien Nutzertests mögliche Strategien beobachtet und dann als Voraussetzung für das Systemdesign benutzt. Dabei soll der Übergang zwischen einzelnen Modi für den Benutzer möglichst sanft sein, im Idealfall sollte er davon gar nichts merken. Diese Vorgehensweise entspricht auch dem Anwendungszweck im Auto, bei dem der Dialog auch allen Strategien der Benutzern entsprechen und trotzdem konsistente Bedienung ermöglichen muss.

Als weitere Festlegung lässt sich relativ einfach ableiten, dass ein solcher MP3-Player keine Dateiansicht unterstützen muss, schon allein, wenn bedacht wird, welche Schwierigkeit es bereiten würde, durch Verzeichnisse und Dateien mit Sprache zu navigieren. Die Verwendung von ID3-Tags wird auch von den meisten Ansätzen als ausreichend angesehen, eine weiterführende Beschäftigung mit weiteren Metadaten müsste sich allerdings nicht auf standardisierte Methoden stützen und selbst entwickelt werden. Da Metadaten-Erzeugung nicht dem Fokus dieser Arbeit darstellen, wurde davon Abstand genommen. Dadurch war auch klar, dass eine implizite Suche so nicht möglich sein würde (siehe Abbildung 3.6).

Auf Basis der Arbeit von Wang et al. [WHHS05], auf der diese Arbeit beruht, wurden nun Überlegungen angestellt, wie ein Dialog aussehen könnte, der diesen Anforderungen entspricht. Die diskutierten Probleme, die Wang et al. aufzeigten, sollen jedoch nicht weiter im Rahmen dieser Arbeit betrachtet werden, da diese eher mit Spracherkennung zu tun haben.

4

Non-verbale Interaktionselemente

Nach einer Einordnung und Definition soll in diesem Kapitel aufgezeigt werden, welche Formen non-verbaler Interaktionselemente unterschieden werden können. Anschließend wird diskutiert, inwiefern die Verwendung bestimmter Formen bei der Anwendung im Auto und spezieller bei der Musikauswahl im Auto angemessen scheint.

4.1 Einordnung & Definition

> *„Der Mensch ist ein auf vielen Ebenen kommunizierendes Wesen, das manchmal auch spricht.“* Ray L. Birdwhistell

In den letzten zwei Kapiteln wurden zwei unterschiedliche Möglichkeiten der Nutzung der auditiven Wahrnehmung des Menschen dargestellt, die Sprache und die Musik. Doch gibt es eine einfache Form von Kommunikationsunterstützung, welche aber schon viel länger zur Kommunikation von Maschinen mit Menschen eingesetzt wird: Der Ton.

Mithilfe von Tönen können Informationen auf sehr unterschiedliche Art Nutzern präsentiert werden, zur klassischen Benutzung von grafischen Benutzerschnittstellen kommen heute auditive und haptische Möglichkeiten hinzu.
Da alle Möglichkeiten spezifische Vor- und Nachteile in bestimmten Situationen haben, ist ebenfalls eine Kombination dieser Möglichkeiten in multimodalen Interfaces denkbar, wie dies schon in Kapitel 2 diskutiert wurde. Diese Form der Kombination verschiedener Modalitäten ist aber nicht neu, sie wurde sie bereits eingesetzt, bevor die eher in neueren Zeiten hinzugekommenen Modalitäten Sprache und Haptik zur Verfügung standen. Hierbei wurden GUIs durch Töne erweitert, um dem Nutzer zusätzliche Informationen zur Verfügung zu stellen, die nicht die Konzentration auf die aktuelle Hauptanwendung stören [BD92].
Wie Brewster [Bre03] feststellt, wurden Töne deswegen anfangs vor allem für Warnungen und Alarmzustände sowie zur Überwachung von Statuszuständen eingesetzt. Erst mit der

Merkmal	Sprache	Töne
visuelle Analogie	Text	Icons
Universalität	sprachabhängig	nur kulturabhängig
Präsentation	lang, seriell	kurz, parallel möglich
Wahrnehmung		
Intuitivität	sofort verständlich	muss evtl. gelernt werden
Bedeutungszuordnung	semantische Analyse	direkte Repräs. ohne Verbalität
Prägnanz	evtl. „verschwimmen“	Hervorhebung
Nebenläufigkeit	drängt in Vordergrund	Ausblendbar
Verwendung	absolute Werte, Instruktionen	Feedback auf Aktionen, hochstrukt./kontinuierl. Daten

Tabelle 4.1: Vergleich Sprache/Töne nach [Bre03] und [RLL+04]

Arbeit von Buxton [Bux89] reifte die Idee, vielleicht auch noch komplexere Information mit Hilfe von Tönen darzustellen.
Brewster motiviert die Verwendung solche Töne weiter aus dem Fakt, dass non-verbale Elemente wie Effekte und Musik schon lange eingesetzt werden, meist um klar zu machen, was gerade vor sich geht und eine Stimmung zu erzeugen. Der logischer nächste Schritt wäre nun, damit Informationen zu präsentieren, die der Nutzer sonst nicht bemerken würde.
Brewster illustriert diese Informationsübermittlung mit einem Beispiel aus den Anfangstagen der Computer. Damals gab es Lautsprecher, die immer ein „klick“ von sich gaben, wenn der Programmzähler verändert wurde. Mit der Zeit lernten die Programmierer die Muster und den Rhythmus der Töne zu deuten und konnten daraus ableiten, was gerade geschah. Computernutzer heutiger Tage kennen dieses Phänomen von ihrer Festplatte. Viele Nutzer können aus den Geräuschen, welche die Festplatte von sich gibt, ableiten, ob ein Speicher- oder Kopiervorgang schon beendet wurde oder noch andauert. Dies ermöglicht es ihnen, ihr Handeln danach auszurichten, also beispielsweise keine rechenintensiven Prozesse zu starten, bevor die Festplatte zur Ruhe gekommen ist.

Es ist also anzunehmen, dass auch prinzipiell die Verwendung akustischer Ausgaben Vorteile haben kann. Doch wann speziell bieten sie einen Mehrwert gegenüber der Verwendung von Sprache (wann die auditive Modalität generell benutzt werden sollte, wurde bereits in Kapitel 2.1 definiert)?

Tabelle 4.1 bietet dazu einen Überblick über die Merkmale von Sprache und non-verbalen Elementen, die nachfolgend erläutert werden sollen.
Brewster versucht Unterschiede zunächst über eine Analogie zur visuellen Modalität klar zu machen. Sprache entspricht dabei ungefähr dem angezeigtem Text in GUIs, non-verbale Elemente dagegen eher Icons, die durchaus viele Bedeutungen in einem kleinen Symbol verschlüsseln können. Dieser Analogie folgend, stellt sich als Vorteil non-verbaler Elemente die Universalität heraus. Während sprachliche Äußerungen nur verstanden werden, wenn die Sprache beherrscht wird, bleiben Töne fast überall gleich gut verständlich. Eine Ausnahme bilden hier nur kulturelle Grenzen, da ohne ein gewisses kulturelles Vorwissen manche Töne nicht mehr eindeutig sind (siehe dazu auch das Beispiel des Handyklingeltons in Abschnitt 4.2).
Dabei sind sprachliche Äußerungen meist lang und seriell, viel „Drumherum“ ist nötig, um

einfache Informationen zu befördern. Dagegen schaffen es Töne kurz und eventuell sogar parallel Informationen zu befördern. Dies verringert aber die Intuitivität, während Sprache sofort verständlich ist, muss die Bedeutung von Tönen eventuell erst gelernt werden. Eine Abhilfe könnten hier die Verwendung von Metaphern zu schon bekannten Tönen schaffen. Sind die Töne einmal erlernt, ermöglichen sie jedoch die direkte Assoziation mit einer Bedeutung, ohne dass sich der Nutzer überhaupt die Worte zu der Bedeutung überlegen muss. Das ermöglicht auch eine direktere und schnellere Reaktion als bei Sprache, welche erst nach der semantischer Analyse des gesamten Satzes verstanden werden kann. Dieser Effekt wird verstärkt durch eine Eigenart der Sinneszellen im menschlichen Ohr, welche bei längeren, gleich bleibenden Töne eher ermüden und so eine Dämpfung des Geräusches eintritt, aber bei kurzen Geräuschen eher stärker reagieren. Bei Sprache führt das gelegentlich zum Verschwimmen der Information, bei Tönen zu einer Verstärkung [Jou01].
Trotz dieser Eigenschaft sind Menschen fähig, Töne auszublenden, wenn sie bekannt und aktuell unwichtig sind. Bei Sprache will das nicht gelingen, sie drängt sich immer in den Vordergrund, beansprucht kognitive Verarbeitungskapazität [RLL+04].
Zusammenfassend stellt Brewster [Bre03] fest, dass Sprache eher für die Kommunikation von absoluten Werten und der Vermittlung von Instruktionen geeignet sind, Töne dagegen für schnelles Feedback auf Aktionen, hoch strukturierte oder kontinuierliche Daten. Eine Kombination scheint in vielen Fällen sinnvoll.

Nun gilt dies nicht nur für Töne, allgemeiner kann auch von non-verbalen Interaktionselementen gesprochen werden. Diese erfassen auch die Möglichkeit, mehrere Töne zu einer Interaktionseinheit zusammenzufassen:

Definition 4.1 ***Non-verbale Interaktionselemente*** *sind systemseitige, auditive Elemente innerhalb der Interaktion zwischen Mensch und Maschine, die keine sprachlichen Informationen enthalten und als Interaktionseinheit begriffen werden. Sie können aus einem oder mehreren Tönen bestehen, aber auch aus Geräuschen.*

Aus Definition 4.1 folgt auch, dass non-verbale Eingaben wie Query-By-Humming (siehe Abschnitt 3.3.3) explizit keine non-verbalen Interaktionselemente sind, da diese nur Elemente der Ausgabe des Computers bezeichnen.
Weiterhin möchte ich diese Definition strikt von der der auditiven Interaktionsobjekten (audIO) abgrenzen, welche Klante [Kla03b] vornimmt. Dort wird eine Definition für sehr konkrete Soundobjekte vorgenommen, welche in ihrer Art dann sehr eingeschränkt sind. Non-verbale Interaktionselemente nach der hier vorgenommenen Definition bestimmen jedoch allgemeine Prinzipien und es wird auch keine Aussage darüber getroffen, wie diese implementiert werden.

4.2 Kategorisierung

Im Weiteren sollen verschiedene Arten von non-verbalen Interaktionsobjekten unterschieden werden.

Grundsätzlich existieren drei Ebenen der Bedeutungszuordnung von Tönen zu denen von ihnen übermittelten Informationen, Suied at al. [SSM+05] leiten diese aus semiotischen Betrachtungen ab. So unterscheiden sie ikonische, indizierende und symbolische Repräsentationen zwischen dem Zeichen und dem Objekt.

Dabei drückt eine ikonische Repräsentation eine direkte Beziehung zwischen Objekt und Zeichen aus. Ein ikonischer Ton muss also in einer unmittelbaren Beziehung zu dem Ereignis stehen, das er ausdrückt. Üblicherweise ist das also der Ton, welcher durch dieses Ereignis erzeugt wird. Der Ton von zerknülltem Papier für das Werfen eines Dokumentes in den Papierkorb des Computers wäre ein Beispiel für diese Art von Repräsentation.
Ein indizierender oder (wie Gaver ihn nennt) metaphorischer Ton hingegen drückt eine Verbindung mit Eigenschaften des beschriebenen Objektes oder Vorgangs aus. Ein Beispiel wäre das Senden einer Email durch die einen ansteigenden Ton zu repräsentieren, die beschriebene Eigenschaft wäre die Fortentwicklung in der Zeit.
Eine symbolische Repräsentation ist dagegen kontextfrei, hier ist die Zuordnung nur durch den Kontext oder Vorwissen gegeben. Beispielsweise hat ein Handyklingelton nichts mit dem eigentlichen Ereignis (Anrufsignalisierung) zu tun, durch persönliche und kulturelle Zuordnung ist aber klar, was damit gemeint ist.

Während ikonische und indizierende Repräsentationen zwar leicht und ohne lernen verständlich, aber unter Umständen schwierig zu finden sind (nicht für alle Objekte und Vorgänge lassen sich direkt oder indirekt Töne zuordnen) haben symbolische Repräsentationen fast immer den Nachteil, dass sie gelernt werden müssen. Dafür können bei diesen auditive Entsprechungen leichter gefunden und diese auch in ein einheitliches System gebracht werden.
Zum praktischen Einsatz werden für die Umsetzung dieser Repräsentation so genannte Auditory Icons verwendet. Diese sollen hier in Anlehnung an eine in diesem Zusammenhang zu spezielle Definition von Gaver [Gav94] folgendermaßen definiert werden:

Definition 4.2 ***Auditory Icons*** *sind non-verbale Interaktionselemente, welche in einer Analogie zu einem repräsentierten Objekt oder Ereignis Informationen über diese an den Nutzer übermitteln.*

Somit wird die Einbindung von Tönen und Geräuschen in ihre natürliche Umgebung und nicht ihre Eigenschaften benutzt, um Informationen zu übermitteln. Durch diese natürliche Art der Bedeutungsrepräsentation sind Auditory Icons relativ einfach zu erlernen und auch wieder aus der Erinnerung abzurufen.
Auch können mehrdimensionale Informationen enthalten sein, so vermittelt der Ton einer zuschlagenden Tür nicht nur Informationen über Größe und Material der Tür, sondern auch Informationen über z.B. die benutzte Kraft oder die Größe des Raums, der zu der Tür gehört [Bre03]. Allerdings sind diese Dimensionen nicht frei wählbar, wie dies bei den weiter unter definierten Earcons der Fall ist.
Weiter bleibt festzuhalten, dass je symbolhafter eine solche Repräsentation wird, seine Kultur- und Zeitgeist-Abhängigkeit steigt. So ist in der westlichen Welt heute des kurze Piepen eines Handys wie natürlich mit dem Empfang einer Kurznachricht verbunden. Doch noch vor wenigen Jahren hätte niemand etwas mit diesem Ton anfangen können, und auch heute gibt es (abgeschottete) Kulturen auf dieser Welt, die dem keine Bedeutung zuordnen können, da sie keine Handys benutzen [Wit03].

Erstellt werden Auditory Icons unabhängig von der Art ihrer Repräsentation meist durch Sampling der entsprechenden Töne oder Geräusche, Brewster [Bre03] formuliert darüber hinaus einige weitere Richtlinien für die Erstellung von Auditory Icons:

1. Verwendung von kurzen Tönen, welche eine große Bandbreite haben, aber in Länge,

Intensität und Tonqualität ungefähr übereinstimmen.

2. Evaluation der Identifizierbarkeit von auditiven Hinweise durch Untersuchung mit offene Fragen.

3. Evaluation der Lernbarkeit von auditiven Hinweisen, welche nicht identifiziert werden können.

4. Test der möglichen konzeptionellen Zuordnungen in einem Versuchsaufbau, in dem das Konzept, welches der auditive Hinweis repräsentiert, die unabhängige Variable ist.

5. Evaluation möglicher Mengen von Auditory Icons zur Feststellung potentieller Probleme mit Maskierungseffekten, Unterscheidbarkeit und widersprüchlichen Zuordnungen.

6. Durchführung von Usability-Experimenten mit Systemen, die die Auditory Icons nutzen.

Im Gegensatz zu den bisher dargestellten Auditory Icons definiert Blattner [BSG89] eine andere Art von non-verbalen Interaktionsobjekten, die so genannten Earcons. Auch Blattners Definition erweist sich im Rahmen dieser Arbeit als nicht ausreichend, anlehnend an seine Definition sollen Earcons folgendermaßen definiert werden:

Definition 4.3 ***Earcons*** *sind non-verbale Interaktionselemente, deren Veränderung einzelner musikalischer Eigenschaften Informationen über ein repräsentiertes Objekt oder Ereignis an den Nutzer übermitteln.*

Somit ist jetzt nicht mehr die Gesamtheit des Objektes entscheidend, sondern in welchem Maße sich bestimmte Eigenschaften Objektes im Bezug zur Zeit verändern. Earcons drücken also eine relative Beziehung zwischen einander aus, und verschlüsseln damit Informationen über die Objekte oder Ereignisse.
Dabei werden Earcons aus simplen Blöcken konstruiert, die Motive genannt werden. Auch wenn dies eine musikalische Bezeichnung ist, müssen damit nicht immer musikalische Motive gemeint sein.

Durch ihre eher künstliche bzw. musikalische Zuordnung ist die Repräsentation nicht intuitiv und muss jeweils gelernt werden, durch die sehr freie Zuordnung kann jedoch eine durchgängige Struktur und damit die Übertragbarkeit von Wissen ermöglicht werden; sogar die Erstellung von sich in die Anwendung nahtlos einfügenden Soundlandschaften ist möglich. Zur besseren Veranschaulichung können Earcons im übertragenden Sinn, der Gleichsetzung von Auditory Icons und Icons folgend, als die Farben und Formen einer grafischen Benutzungsoberfläche betrachtet werden.
Somit ist auch klar, dass eine ikonische Zuordnung bei Earcons nicht möglich ist. Zwar stellen Suied et al. [SSM+05] fest, dass einfache Earcons auch indizierende Repräsentationen ermöglichen können, wahrscheinlicher ist jedoch, dass eher symbolische Referenzen durch Earcons ausgedrückt werden.
Erzeugt werden Earcons meist über algorithmische Verfahren zur Laufzeit unter Verwendung der MIDI-Schnittstelle [Bre03], die ermöglichen, beliebige Variationen von Motiven

gegeneinander durchzuführen. Damit bei diesen Kombinationen und auch den möglichen Varianzen innerhalb der benutzten Motive keine Dissonanzen auftreten, sollte bei der Erstellung ein Musiker involviert werden, wie Alty [ARV05] feststellt. Und natürlich beeinflusst auch die musikalische Begabung des Hörers die Möglichkeit, Informationen aus den Earcons zu erkennen.
Doch welche Motive der Earcons können nun beeinflusst werden? Blattner [BSG89] unterscheidet dabei grundsätzlich die Beeinflussung in:

- Klangfarbe,
- Rhythmus,
- Tonhöhe,
- Oktave,
- Lautstärke.

Dabei sind die ersten Einträge diese Aufzählung die ausdrucksstärksten, und die letzten die ausdrucksschwächsten Motive. Bei der Unterscheidung in Oktaven tritt das Problem auf, dass das menschliche Gehör diese schwer unterscheiden kann, da die jeweiligen Obertöne jeweils ebenfalls in diesem Abstand erklingen (zu näheren Details siehe [Jou01]). Weiterhin können solche Veränderungen in der Lautstärke von Nutzern leicht als störend empfunden werden, wie Brewster [Bre03] feststellt.

Sowohl Earcons als auch die zuvor behandelten Auditory Icons erweisen sich für die Kommunikation von Information via auditiver Möglichkeiten in bestimmten Situationen als sinnvoll. Ein Vergleich beider Formen ist in Tabelle 4.2 dargestellt.

Merkmal	Auditory Icons	Earcons
visuelle Analogie	Icons	Farben & Formen
Intuitivtät	semant. Link meist klar bzw. einfach zu lernen/erinnern	muss immer gelernt werden, event. implizites Lernen möglich
Abstraktion	eher gering	sehr mächtig
Einheitlichkeit	schwer zu gewährleisten	möglich
Verwendung	eher Einzeltöne, Anw. für „Laufpublikum“	eher komplexe Anwendungen, keine realen Entsprech. nötig

Tabelle 4.2: Vergleich Auditory Icons/Earcons nach [Bre03]

Daraus ist ersichtlich, dass bei Auditory Icons in den meisten Fällen kein Lernen stattfinden muss. Insbesondere bei den ikonischen Auditory Icons ist die semantische Verbindung oft intuitiv klar, wogegen bei Earcons immer ein Lernvorgang stattfinden muss. Brewster [Bre03] weist aber darauf hin, dass es bereits Arbeiten gibt, die darauf hinarbeiten, implizites Lernen von Earcons während der Benutzung der Anwendung zu ermöglichen. Dieser klare Vorteil von Auditory Icons geht allerdings verloren, wenn ihre Bedeutung nicht einfach ableitbar ist. Insofern muss beim Design darauf sorgfältig geachtet werden, dass die Auditory Icons möglichst prägnant sind.
Der Vorteil von Earcons zeigt sich bei abstrakten, nicht in der Realität mit einer Entsprechung versehenen Anwendungsszenarien. Auch eine Vereinheitlichung des gesamten

auditiven Auftritts fällt mit Earcons leichter als mit Auditory Icons.
Als Fazit bleibt die Empfehlung von Auditory Icons für Anwendungen mit viel „Laufpublikum“, welches schnell und einfach verstehen soll, was die Töne bedeuten. Auch sollten die Anwendungen nicht zu komplex sein, da sich anderenfalls an dieser Stelle der Einsatz von Earcons lohnen würde, ebenso wie auch in sehr abstrakten Themengebieten.

Neben Auditory Icons und Earcons wurden in der Literatur auch Hearcons diskutiert [DKG02]. Diese erweitern Earcons in den dreidimensionalen Raum, Bölke [Böl97] definiert sie als akustische Objekte, die in einem akustischen 3-dimensionalen Raum positioniert werden. Dies macht allerdings eine Möglichkeit zur Benutzung von Raumklang im Audiosystem notwendig, welche im Rahmen der Arbeit nicht zur Verfügung stand.[1]

Einen weiteren interessanten Ansatz bietet Conversy [Con98] an, der Samples als Grundlage von Auditory Icons analysiert, Parameter auf der Wahrnehmungsebene analysiert (wie z.B. Rhythmus, Melodie, etc., siehe Abschnitt 3.1) und diese ad-hoc verändern kann. Damit wird ein Weg aufgezeigt, in Zukunft die strikte Grenze zwischen Auditory Icons und Earcons zu überwinden.

Letztendlich sind die meisten bisher veröffentlichten Kategorisierungen vor allem Strukturbetrachtungen. Durch sie ist es möglich abzuleiten, welche Formen von non-verbalen Interaktionselementen idealerweise im jeweiligen Kontext benutzt werden sollen.
Doch bestehen Schwierigkeiten, Regeln aufzustellen, wie solche non-verbalen Interaktionselemente aussehen sollen, es fehlen klare Gestaltungsregeln. So stellt Gärdenfors [Gär02] fest, dass die westliche Kultur zwar über eine reichhaltige visuelle Ikonographie verfügt, es aber keine gängige auditiven Entsprechung gibt.[2] Eine Inspirationsmöglichkeit besteht zwar in der musikalischen Sprache von Film, Fernsehen und Radio, doch ist diese in vielen Fällen auch nicht hilfreich. In den Zukunft interessant in diesem Zusammenhang wäre eine Diskussion der Verwendung non-verbale Interaktionselemente in Computerspielen und wie sich dort entwickelte Prinzipien vielleicht für die Verwendung im Kontext von Sprachdialogsystemen eignen würden.

4.3 Automobile Anwendung

Sollen non-verbale Interaktionselemente im Auto benutzt werden, muss zunächst geklärt werden, welche Voraussetzungen dort vorgefunden werden. Zunächst ist dabei zu beachten, dass im Auto nicht nur Rückmeldungen eines eventuellen Dialogs zwischen Fahrer und Auto kommuniziert werden, sondern auch sicherheitskritische Warntöne oder beispielsweise das Geräusch eines Fahrtrichtungsanzeigers. In ihrer Gesamtheit lassen sich diese akustischen Signale an den Fahrer (so werden die Rückmeldungen in der ISO-Norm genannt) in drei Klassen einteilen [ISO04]:

kurzfristige Nachrichten
: Erfordert unmittelbares Handeln des Fahrers. Für kritische Ereignisse.

[1] Was bedauerlich war, denn gerade im Auto kann die feste und bekannte Anordnung der Lautsprecher eine exakte Positionierung möglich machen. Ein solches System könnte aber nur in enger Zusammenarbeit mit Automobilbauern entwickelt werden, was im Rahmen der Arbeit nicht möglich war.

[2] Das führt dazu, dass noch häufiger Töne erklärt werden müssen (natürlich mit Sprache), was an vielen Stellen die Vorteile der Verwendung solche Töne wieder entwertet.

Beispiele für Töne:
Auffällige Tonfolgen, wechselnde Tonhöhen, schneller Rhythmus, Dissonanzen.

mittelfristige Nachrichten
Reaktion sollte in kurzer Zeit erfolgen (10-20s). Für aktuelle Ereignisse (Routenhinweis Navigationssystem). Beispiele für Töne:
Muster mit konstanter Tonhöhe, mindestens 0,3s lang.

langfristige Nachrichten
Zukünftiges Verhalten wird erwartet. Ankündigungen und Statusmitteilungen. Beispiele für Töne:
Zweimaliger Gong, nicht-periodische Hoch-Tief-Töne, meist gefolgt von Sprachmitteilung oder grafischer Anzeige.

Für diese Arten von Signalen gibt es im Wesentlichen zwei Funktionen: Sie können Aufmerksamkeit lenken, aber auch konkrete Information vermitteln.
Die meisten klassischen Warntöne dienen eher der Aufmerksamkeitslenkung, während das schon erwähnte Blinkergeräusch eine konkrete Information übermittelt. Im Kontext der Arbeit sind kurzfristige Töne nicht interessant, da ein Entertainment-System keine systemkritischen Nachrichten vermitteln muss. Aber sowohl Aufmerksamkeitslenkung als auch Informationsübermittlung ist für den MP3-Player prinzipiell denkbar.
Die ISO-Norm weist weiter darauf hin, nur wenige Töne zu verwenden, um Unterscheidbarkeit zu gewährleisten. In den Fälle, in denen Töne gelernt werden müssen, sollte gewährleistet sein, dass die Nutzer ihnen regelmäßig ausgesetzt sind, um ihre Bedeutung im Zweifelsfall klar und eindeutig abrufen zu können. Für die Anzahl von Warntönen geben die Design Richtlinien des US Verkehrsministeriums [GLPS95] sogar eine quantitative Aussage, nicht mehr als drei oder vier solcher Töne sollen es demnach sein. Daraus leitet sich ab, dass für die Entertainment-Funktionen nicht mehr viel übrig bleibt, da ein weiterer Ton mit der PTT-Signalisierung belegt ist.
Weiterhin enthalten diese Design Richtlinien neben Angaben zu typischer Lautstärke und Frequenz für nicht kritische Töne Empfehlungen, eher von der Tönhöhe her niedrige Töne zu verwenden und eine Dauer von 100 bis 150 ms zu wählen.

Doch neben diesen allgemeinen physikalisch-technischen Empfehlungen muss natürlich auch betrachtet werden, was speziell bei der Anwendung der non-verbalen Interaktionselemente im Auto zu beachten ist. In einer Arbeit von Vilimek und Hempel [VH05] wird beispielsweise darauf hingewiesen, dass die Nutzung im Auto nicht, wie beispielsweise bei der Erweiterung von GUIs, aus Optimierungsgesichtspunkten folgt, sondern eine Möglichkeit bietet, nicht-ablenkende Hinweise über das Systemverhalten zu integrieren, die sonst so nicht möglich wären.
Weiterhin muss allerdings das System im Auto auch weiter intuitiv bedienbar bleiben, wie dies bereits im Abschnitt 3.3.3 diskutiert wurde.
Das spricht für die Verwendung möglichst ikonischer Auditory Icons.[3] Solche sehr auffälligen Töne unterliegen aber der Gefahr, mit den ähnlich prägnanten Warntönen verwechselt zu werden. Statt nicht-ablenkender Hinweise zu integrieren, könnte so mehr Ablenkung

[3] Vilimek und Hempel [VH05] empfehlen in diesem Umfeld die Verwendung von kurzen Schüsselwörtern und prägnanter Auditory Icons.

für den Fahrer entstehen. Auch besteht die Gefahr, dass bei zu ikonischen Tönen die Nervigkeit für die Nutzer stark zunimmt, wenn sie oft ertönen, wie sowohl Bussemakers el al. [BH00] als auch McKeown [McK05] festhalten. Hier besteht der wesentliche Unterschied zwischen Warntönen und non-verbalen Interaktionselementen, welche nicht um jeden Preis auffallen, aber schon erkannt werden müssen.
Folglich muss an dieser Stelle ein Kompromiss zwischen der Ausdrucksstärke der ikonischen Auditory Icons einerseits und symbolischer Abbildung andererseits gefunden werden. Die verwendeten non-verbalen Interaktionselemente müssen auffällig genug sein, um als Hervorhebung wahrgenommen zu werden, aber unauffällig genug, um die restliche Wahrnehmung nicht zu stören. Abbildung 4.1 verdeutlicht zusammenfassend noch einmal den Unterschied der verschiedenen Abbildungsformen und ihrer hauptsächlichen Eigenschaften.

Abbildung 4.1: Eigenschaften ikonischer und symbolischer Abbildung

Brewster [Bre03] stellt schließlich noch zwei Systeme vor, die aufzeigen, wohin die Nutzung von non-verbalen Interaktionselementen im mobilen Einsatz führen kann und wie einige der erwähnten Widersprüche aufgelöst werden könnten.
So erwähnt er das Forschungsprojekt Normadic Radio [SS00], welches neben der Verwendung von dreidimensional angeordneten Tönen auch Präsentationsebenen adressiert. Normadic Radio besitzt sieben Ebenen auditiver Präsentation, von Stille, über ambiente und normale Töne bis zu vier abgestuften Ebenen von Sprachausgabe. Gesteuert wird die Auswahl der Modi über eine Erkennung der Hintergrundgeräusche, je nach Geräuschpegel um den Benutzer wird die Detailebene eingestellt. Dies könnte auch im Auto eine Lösung darstellen, nur das dort andere Sensoren benutzt werden müssten. Je nach Verkehrssituation könnten dort Töne ausdrucksvoller oder weniger ablenkend eingestellt werden.
Ein anderes Projekt ist Audio Aura [MBW+98]. Dieses betont vor allem das bereits erwähnte Ziel, möglichst wenig Aufmerksamkeit aktiv zu binden, aber trotzdem Hintergrundinformationen anzubieten. Die Autoren definieren dafür die „sonic ecologies“, das sind Gruppen von Tönen oder Geräuschen, die zusammenpassen als ein stimmiges Ganzes. Dies wird beispielsweise dadurch erreicht, dass ankommende Emails durch Möwenschreie auralisiert werden, und dies für fünf oder zehn Emails einfach mehr Schreie werden. Die anderen Funktionen sind auch durch Tierlaute bestimmt, so das hier wirklich ein zusammenhängendes Ganzes entsteht. Auch zeigt es auf, wie die Stärken von Auditory Icons (intuitiv, kurz, ausdrucksstark) und Earcons (mächtig, nicht zu auffällig) kombiniert werden können.

4.4 Benutzung für sprachgesteuerte Musikauswahl im Auto

Nachdem die Verwendung von non-verbalen Interaktionselementen im Auto allgemein diskutiert wurde, stellt sich nun die Frage, wie diese grundsätzlich im Kontext sprachgesteuerter Musikauswahl im Auto eingesetzt werden könnten. Dazu muss zunächst einmal betrachtet werden, was die Umgebung von Musik für non-verbale Interaktionselemente bedeutet.
Denn wie bereits diskutiert, können beispielsweise Earcons auch aus musikalischen Elementen bestehen. In Anbetracht dieser Tatsache ist der Mix von Musik und zu musikalischen non-verbalen Elementen, problematisch, da Verwechslungsgefahr besteht. Im Umfeld experimenteller Musik können sogar Einzeltöne oder Geräusche als Teil der Musik verstanden werden. Somit ist die wesentliche Herausforderung bei der Musikauswahl, klare Unterscheidbarkeit von Inhalten und Interface sicherzustellen.
Dem gegenüber steht die schon diskutierte mögliche Nervigkeit von zu ikonischen Tönen, die bei der Musikauswahl, welche eine häufige Tätigkeit im Auto darstellt (siehe Abschnitt 3.3.2), zudem noch oft auftreten würden.

Auch strukturelle Betrachtungen helfen bei der Einordnung möglicher Anwendungsmöglichkeiten von non-verbalen Interaktionselementen. So wurden im Abschnitt 3.4 hauptsächlich zwei verschiedene Möglichkeiten zur Musikauswahl identifiziert, eine eher explizite Auswahl im Hierarchie-Modus und eine unschärfere Möglichkeit über eine Art von Suchfunktion. In Abbildung 4.2 wird nun dargestellt, welche Arten von non-verbalen Interaktionselementen bei diesen beiden Möglichkeiten potentiell Vorteile bieten.

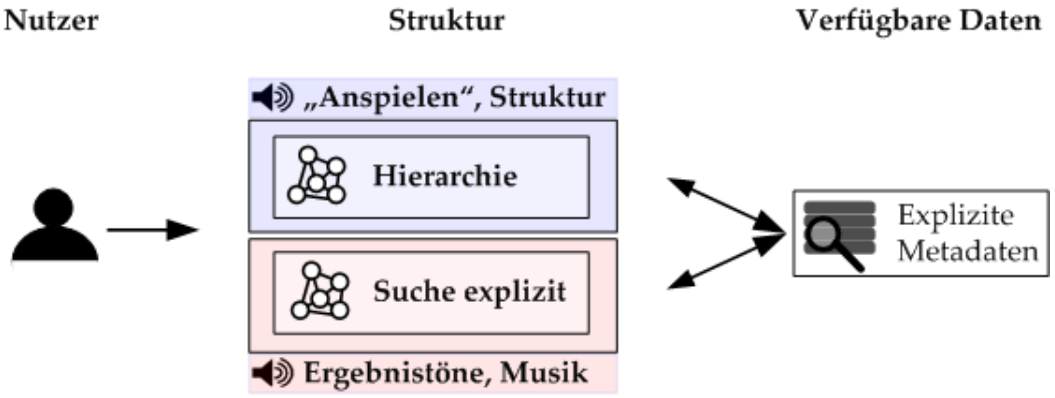

Abbildung 4.2: Einsatzmöglichkeiten non-verbaler Interaktionselemente bei Musikauswahl

Im Hierarchie-Modus wird dabei Musik in definierten Schritten ausgewählt, dem Nutzer ist dabei die Struktur klar und ersichtlich. Der Nutzer navigiert dabei hauptsächlich durch Listen, wobei die Informationen werden entweder über konkrete Einträge oder die Struktur übermittelt.
Bei ersterem wird häufig der Name des Eintrags benutzt, im Themenkreis Musik ist hier natürlich auch denkbar die Musik selbst, zum Beispiel über eine „Anspielen“-Funktionalität, zu benutzten. Bei der Struktur ist eine Auralisation von Strukturinformationen (Tiefe des aktuellen Menüs, Art der ausgewählten Liste usw.) denkbar. Die Bestätigung von Befehlen mit Hilfe von non-verbalen Elementen erscheint dagegen eher unpraktisch, weil so die verbale Verifikation des Kommandos verhindert würde.

Wenn dagegen hierbei eine Suchfunktion mit non-verbalen Interaktionselementen unterstützt werden soll, stellen sich andere Herausforderungen. Der Nutzer soll hierbei dabei unterstützt werden, herauszufinden, wie nah sein Suchergebnis seinem eigentlichen Suchziel gekommen ist, um sein weiteres Vorgehen danach auszurichten.
Dafür wären Ergebnistöne denkbar (Anzahl Treffer, Genauigkeit der zurückgelieferten Treffer, Art von Ergebnis), aber auch wiederum die Verwendung von Musik direkt als Indikator für Qualität des Suchergebnisses (wenn die Musik ungefähr dem entspricht, was erwartet wurde, war die Suche gut). Generell wäre hier die ambiente Unterstützung vorzuziehen, da diese Informationen nur Zusatzinformtionen enthalten, denn letztendlich geht es bei einer Suche um direktes Finden. Nur wenn dieses nicht direkt zum Ziel führt, sind für den Nutzer auch Zusatzinformationen von Interesse.

Wie bereits in Abschnitt 3.4 diskutiert, muss eine Musikauswahl für das Auto beide hier diskutierten Möglichkeiten der Musikauswahl unterstützen. Das erfordert eine Synthese der Ansätze für die Verwendung non-verbaler Interaktionselemente. Dabei wurde die Kombination von Struktur- und Ergebnistönen betrachtet, die sich in ein konkretes System einheitlich und konsistent einfügen mussten. Musik als Interaktionselement zu verwenden wurde in beiden Ansätze betont, deswegen erschien die Verwendung im konkreten System sinnvoll.

Weiterhin boten sich Anknüpfungspunkte bei einer Arbeit von Fröhlich und Hammer [FH05] an.
Sie diskutierten die Anwendung der Gestaltgesetze [EK00] für Hervorhebungen in Emails unter Nutzung non-verbaler Interaktionselemente und versuchten so Möglichkeiten zu finden, Gruppierung von Text zu ermöglichen. Aus ihrer Diskussion von Gestaltgesetzen wie Ähnlichkeit, Nähe und gemeinsamen Schicksal leiten sie ab, wie Töne zur Abgrenzung (in Aufzählungen), Zusammenführung (für Zitate) und Hervorhebung von Elementen (Überschriften) eingesetzt werden können.
Bei der Begrenzung von Elementen erwies sich im durchgeführten Nutzertest ein kurzer Ton („click“) als deutlich nützlicher als die bloße Benutzung von Pausen. Für die Zusammenführung von Elementen bewährten sich Hintergrundtöne in Form von sanften Akkorden, eine Variante mit Pausen wurde auch hier mehrheitlich abgelehnt. Und auch die Hervorhebung, benutzt wurde hier ein anderer Hintergrundakkord, wurde mehrheitlich so akzeptiert.[4]
Fröhlich und Hammer merken im Ausblick ihrer Arbeit an, dass eine allgemeine Anwendung dieser Prinzipien in Sprachdialogsystemen angedacht werden sollte. Möglichkeiten dafür zeigt Tabelle 4.3 in einer Aufstellung der Prinzipien von Fröhlich und Hammer und ihre Umsetzung in Tönen in Bezug auf mögliche Anwendungen im Bereich der Musikauswahl auf.
Auch dort treten Elemente auf, die getrennt bzw. hervorgehoben werden müssen, zum Beispiel die Elemente einer Ergebnisliste. Wenn diese Liste angespielt werden soll, wie oben bereits als Möglichkeit diskutiert, müsste eine Trennung zwischen den Elementen

[4] Hamer-Hodges et al. [HHLC05] diskutierten ebenfalls Hervorhebung, aber am Beispiel von Hyperlinks. Sie arbeiten dabei im Gegensatz zu Fröhlich und Hammer [FH05] mit Vortönen zur Hervorhebung des Links und nutzten dabei non-verbale Möglichkeiten nur zur Erweiterung der schon vorhandenen sprachliche Auszeichnung (Wechsel der Stimme). Wahrscheinlich auch deswegen konnten sie in ihrem Nutzertest keinen Mehrwert der Verwendung non-verbaler Elemente belegen.

Prinzip	Töne	Anwendung bei Musikauswahl
Abgrenzung	kurzer Ton („klick“)	Trennung Elemente Ergebnisliste
Zusammenführung	Hintergrundtöne (Akkord)	Auswahlmodi (Hierarchie/Suche)
Hervorhebung	Hintergrundtöne (Akkord)	ausgewählter Eintrag

Tabelle 4.3: Prinzipien von Fröhlich und Hammer [FH05] und ihre eventuelle Anwendung für Musikauswahl

gefunden werden. Und sollten sie vorgelesen werden, wäre es vielleicht interessant, wenn die Systemmeldung von den Namen der Einträge auch akustisch getrennt wäre. Ebenso böte sich hier die Möglichkeit, unterschiedliche Auswahlmodi eventuell mit Hintergrundtönen zu verbinden. Und schließlich wäre auch die akustische Markierung eines „Cursors“ denkbar, wofür das Prinzip der Hervorhebung sehr gut geeignet wäre.

Die konkreten Festlegungen über die Art und den Umfang der Integration non-verbaler Interaktionselemente konnten erst nach dem Festlegen des endgültigen Systemdesigns getroffen werden und werden deswegen in Abschnitt 7.2 dargestellt. Dies folgt der Empfehlung von Klante [Kla03a], der die Verwendung getrennter Entwicklungszyklen für Dialog und Töne vorschlägt, um unabhängige Evaluation der Einzelkomponenten zu ermöglichen.

5
Usability Engineering

In diesem Kapitel soll zunächst eine Begriffsklärung bezüglich Usability bzw. Usability-Engineering erfolgen. Ebenso soll diskutiert werden, welche Konsequenzen eine strikte Orientierung der Entwicklung an diesen Prinzipien bedeutet. Schließlich sollen Besonderheiten im Auto und bei Sprache beleuchtet und schließlich Evaluationsmethoden vorgestellt werden.

5.1 Einordnung & Definition

> *„Alles sollte so einfach wie möglich gemacht werden, aber nicht einfacher."*
> Albert Einstein

Bisher wurden in dieser Arbeit im Wesentlichen Funktionen und technische Rahmenbedingungen diskutiert. Aber die tatsächliche Nutzung von Funktionen eines Systems hängt auch davon ab, ob der Nutzer in der Interaktion mit dem System seine Ziele auch umsetzen kann. Mit diesen Fragestellungen beschäftigt sich das Gebiet der Usability, in welchem der Nutzer in den Mittelpunkt der Betrachtung gestellt wird. Witte [Wit03] gibt im Rahmen seiner Arbeit einen Einblick in das Gebiet der Usability, die weitere Diskussion lehnt sich an seine Darstellung an.
Erste systematische Forschungsansätze zu Usability gab es schon in den 70er Jahren. Die Entwicklung der grafischen Benutzerschnittstelle ist eines der bekanntesten Ergebnisse solcher Forschung. Im Laufe der Jahre entwickelte sich neben einer umfangreichen Methodik auch eine genaue, abgrenzte Definition der Begrifflichkeiten rund um das Gebiet Usability, welche schließlich im Laufe der 90er Jahre genormt wurden.
Laut ISO-Norm 9241-11 [ISO98] wird Gebrauchstauglichkeit[1] (die deutsche Bezeichnung für Usability) so definiert:

[1] Im Weiteren wird hierfür aber weiterhin der Begriff Usability verwendet.

Definition 5.1 ***Gebrauchstauglichkeit*** *ist das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.*

Dabei bezeichnet Effektivität die Genauigkeit und Vollständigkeit, mit der ein Benutzer ein bestimmtes Ziel erreicht, Effizienz das Verhältnis von Effektivität zum aufgewendeten Aufwand und Zufriedenstellung die Freiheit von Beeinträchtigungen und positive Einstellungen gegenüber der Nutzung.[ISO98]
Diese drei Ziele können aber nur innerhalb eines bestimmten Nutzungskontextes und für eine bestimmte Benutzergruppe getroffen werden. Es ist demnach nicht allein die Eigenschaft eines Produktes, sondern das Attribut einer Interaktion eines Benutzers mit einem Produkt innerhalb eines bestimmten Kontextes. Also spielen sowohl Eigenschaften des Nutzers, beispielsweise Vorwissen, Erfahrungen und persönliche Eigenschaften, als auch der Kontext des Einsatzes eine gewichtige Rolle. Das bedeutet aber auch, dass Erkenntnisse über die Usability weder einfach von Nutzer zu Nutzer noch von Kontext zu Kontext übertragen werden können.

Aus diesen Anforderungen an die Gebrauchstauglichkeit können nun Grundsätze für die Dialoggestaltung formuliert werden, was in Teil 10 der bereits angesprochenen ISO-Norm 9241 [ISO96] getan wurde [2]. Diese Grundsätze umschreiben konkrete Eigenschaften, die ein Dialog nach der Entwicklung besitzen sollte (z.B. Selbstbeschreibungsfähigkeit, Erwartungskonformität, Lernförderlichkeit).

Doch wie kann sichergestellt werden, dass diese Ziele am Ende auch eingehalten werden? Wie kann auch die Beachtung dieser Prinzipien direkt in den Entwicklungsprozess von Systemen eingebunden werden?
Am Anfang, bevor sich im Rahmen von Softwareentwicklungen überhaupt vertieft mit Usability auseinander gesetzt wurde, geschah dies meist am Ende des Entwicklungsprozesses der Software. Checklisten und kleine Nutzertests sollten ermöglichen, kleinere Optimierungen am fertigen Produkt durchzuführen. Usability war zu der Zeit nicht mehr als ein Aufsatz auf den gewöhnlichen Entwicklungsprozess.
Sinnvoller ist es jedoch, den gesamten Entwicklungsprozess konsequent auf die Erreichung von Usability-Zielen auszurichten. An dieser Stelle setzt Usability Engineering an, das Nielsen [Nie93] als systematische Integration von Methoden und Techniken der Usability in den Entwicklungsprozess beschreibt. Hierbei wird Usability in den Mittelpunkt der Betrachtung gestellt und alle Entwicklungsschritte der Erzielung von Usability-Zielen untergeordnet. Dies kann durch Anwendung der Dialogprinzipien in vorgefertigten Dialogbausteinen oder konkreter anwendungsspezifischer Style Guides geschehen, aber auch durch die Integration verschiedener Formen der Nutzerbefragung und -evaluation.
In verschiedenen Arbeiten (so von Nielsen [Nie93] oder Mayhew [May99]) wird dabei jeweils zentral die Notwendigkeit von iterativen Zyklen in der Entwicklung betont, in denen oft die Nutzer einbezogen werden, um Verbesserungspotential für einen nächsten Schritt festzustellen. Dabei können die bereits diskutierten Grundsätze der Dialoggestaltung als Startpunkt (neben Aufgabenanalyse und Analyse technischer Einschränkungen) dienen, ihre konkrete Anwendung sollte jedoch in Tests überprüft werden.
Ausgehend von dieser Diskussion soll Usability Engineering wie folgt definiert werden:

[2] Eine übersichtliche und ausführliche Diskussion dieser Grundsätze findet sich beispielsweise in [Wir06]

Definition 5.2 ***Usability Engineering*** *ist ein iterativer Entwicklungsprozess, der die Erreichung von Usability-Zielen in den Mittelpunkt der Betrachtung stellt und dafür entsprechende Methoden und Techniken in den Entwicklungsprozess integriert. Dabei dienen meist allgemeine Grundsätze der Dialoggestaltung als Ausgangspunkt, Nutzer werden aber umfassend am Prozess beteiligt.*

Diese Beteiligung findet mittels verschiedener Untersuchungsmethoden statt, die aus verschiedensten Motiven durchgeführt werden können. So können solche Untersuchungen, etwa durch Befragungen, eine Hilfe in der Analysephase darstellen, um so mehr über den Nutzer, seine Anforderungen und Strategien zu erfahren. Weiterhin können Nutzer in verschiedenen Phasen der Entwicklung durch Evaluation von Konzepten und Prototypen eingebunden werden. Und selbst nach der Auslieferung eines Systems kann der Beobachtung der Benutzung der Systeme Auskunft darüber geben, wie ein System weiter verbessert werden kann.[3].

Soviel Beteiligung des Nutzers hat meist seinen Preis, die Benutzung angemessen ausgestatteter Usability Labore kostet Geld und die Zeit, die für Vorbereitung, Probandensuche, Durchführung und Auswertung solcher Tests aufgewendet wird, kann beträchtlich sein und letztendlich in keinem Verhältnis zum erzielten Nutzen stehen. Preim [Pre99] diskutiert deswegen Methoden zur Kostensenkung solcher Experimente, welche meist erst umfangreiche Beteiligung von Nutzern ermöglichen:

Testziel sorgfältig auswählen
: Meist sind nicht alle Aspekte des Systems von Interesse. Eine Verkürzung des Tests auf wesentliche Punkte verkürzt Vorbereitung, Durchführung und Auswertung erheblich.

Aufwand Prototypenerstellung begrenzen
: Der Test eines zu 100% funktionstüchtigen Systems ist nur selten notwendig. Oft reicht es aus, Papier und Bleistift, Zeichnungen, Präsentationsprogramme oder Prototyping-Werkzeuge zu verwenden und diese zu nutzen, um im Zweifel in eine Diskussion mit den Testteilnehmern einzusteigen.

Testumgebung
: Nicht jeder Test braucht ein voll ausgerüstetes Labor, oft reicht auch einfache Standardtechnik, um die erforderlichen Daten zu erheben. Die Umgebung muss nur natürlich genug sein, um die Testbedingungen nicht zu beeinflussen.

Anzahl der Testpersonen
: Laut Nielsen [Nie89] sind nur 4 Personen nötig, um 85% aller Probleme in einem System zu erkennen. Dies ermöglicht mehrere kleine Tests statt weniger großer. Nur bei statistischen Aussagen („ein gewisser Prozentsatz löst diese Aufgabe unter diesen Bedingungen“) ist eine größere Anzahl von Probanden (Richtwert 15) erforderlich.

Auswahl der Testpersonen
: Die Wichtigste Bedingung ist, dass Personen ausgewählt werden, die unvoreingenommen sind und nicht direkt an der Entwicklung beteiligt sind. Neue Mitarbeiter sind beispielsweise immer natürliche Testpersonen, die auch leicht verfügbar sind.

[3]Die konkreten, im Rahmen dieser Arbeit benutzten Methoden werden in Abschnitt 5.4 vorgestellt.

Im Weiteren soll nun dargestellt werden, welche speziellen Usability-Probleme sich bei Systemerstellung für das Auto und bei der Benutzung von Sprache stellen und wie diese entsprechend im Entwicklungsprozess adressiert werden können. In diesem Rahmen werden dann auch jeweils Methoden diskutiert, um die Kosten der Untersuchungen möglichst möglichst niedrig zu halten. Ausgehend davon, nach einer Vorstellung verschiedener Usability-Untersuchungsmethoden, soll dann unter Beachtung von ähnlichen Entwicklungen ein Vorgehensmodell für diese Arbeit entwickelt werden.

5.2 Usability im Auto

Usability im Auto wurde lange als nahezu perfekt angesehen, Preim [Pre99] benutzt es in seinem Buch sogar als Vorbild für sinnvolle Umsetzung von Usability-Zielen. Das beeindruckt umso mehr, da bei der Festlegung der meisten Parameter der Bedienung eines Autos (Lenkrad, Schaltung, Gas, Bremse) Usability-Ziele noch nicht bekannt waren.
Standardisierung und Normung zentraler Bestandteile und der Zwang zu einer Schulung vor der Benutzung des Fahrzeugs (Fahrschule) sind sicher Hauptgründe dafür, warum heute viele Menschen die Bedienung eines Autos selbstverständlich und kinderleicht erscheint. Dies gilt jedoch nur für die Fahraufgabe, denn heutige Autos integrieren eine immer größer werdende Anzahl von Komfortfunktionen, deren Bedienung zuletzt die Anzahl der dafür notwendigen Knöpfe im Auto förmlich explodieren ließ. Zudem mussten immer mehr komplexe Funktionen gesteuert werden, die sich nicht mehr mit einfachen Knöpfen bedienen lassen konnten. So ähnelte die Bedienung im Auto immer mehr der am Computer.

Dies führte zum Aufeinandertreffen zweier bisher getrennter Design-Ansätzen, wie Akesson und Nilsson [AN02] festhalten. So haben klassische Automobildesigner vor allem das Ziel, das Auto und das Fahren so sicher wie möglich zu machen. Dagegen geht es klassischen Usability-Designern eher um eine effektive Nutzung von Funktionen, eine Steigerung des Erlebnisfaktors und die Förderung von Verständlichkeit solcher Systeme. Die Verbindung beider Ziele sollte das Endziel jeder Entwicklung im Auto sein.
In letzter Zeit wurden Anstrengungen zur Vereinfachung und Zentralisierung der Bedienung in einem oder wenigen Steuergeräten unternommen (beispielsweise BMWs iDrive). Jedoch konnten diese ambitionierten Versuche das Versprechen der Einhaltung beider Designziele nicht vollständig erfüllen. Sprachsysteme, wie bereits in Abschnitt 2.4 diskutiert, bieten hier einen Ausweg an.
Usability-Untersuchungsmethoden für das Auto müssen immer beide Ziele berücksichtigen. So muss die Auswirkung der Bedienung auf die Sicherheit des Fahrens immer berücksichtigt werden, aber auch, ob die Benutzung der Systeme effektiv, effizient und zufriedenstellend für den Nutzer möglich ist.
Um die Sicherheit zu berücksichtigen, sollte ein Test solcher Systeme mit der Bedienung einer Fahrsimulation gekoppelt werden, um die Bedienung in Verhältnis zu der verursachten Ablenkung zu testen. Dafür bietet sich als kostensenkende Methode der Lane Change Test von DaimlerChrysler [Mat03] an, einer Fahrsimulationssoftware für Standard-PCs, die den Nutzer mit einer Fahraufgabe zu Spurwechseln ablenkt. Diese lässt sich über preiswerte Computerspiel-Lenkrädern steuern und ermöglicht so im Vergleich ähnlich realistische Tests wie in einem voll ausgestatteten Fahrsimulator, aber zu einem vielfach geringeren Preis und Aufwand.

5.3 Usability bei Sprache

Während für die Gestaltung klassischer haptisch-grafischer Benutzerschnittstellen umfangreiche Style-Guides, Dialogmodule und Erfahrungen aus früheren Projekten am Anfang einer Entwicklung eines neuen Systems zur Verfügung stehen, gibt es bei Sprache wenige solche ausgereiften Vorarbeiten. Zwar gelten prinzipiell die allgemeinen Grundsätze für die Dialoggestaltung laut ISO 9241 Teil 10 (Abschnitt 5.1) auch bei Sprache, jedoch sind hier spezielle Eigenheiten zu beachten. So weist Larsen [Lar03] darauf hin, dass unter Anderem durch die Nicht-Persistenz von Sprache oder die höhere Komplexität und Fehleranfälligkeit von Sprache ganz andere Themen an Bedeutung gewinnen, als dieses bei traditionellen Interfaces der Fall ist. Suhm [Suh03] ergänzt, dass auch Eigenschaften wie Umgebung, menschliche Kognition und unerfüllbare Erwartungen der Nutzer zu Usabilityproblemen führen können.
Als Antwort auf diese Herausforderungen ist eine allgemein akzeptierte Definition von Usability-Zielen an verschiedenen Stellen bereits versucht, eine allgemein akzeptierte oder genormte Definition jedoch noch nicht gefunden worden.
Unter Beachtung auch der Besonderheiten des Einsatzes im Auto soll hier aber der Konsens von verschiedenen Empfehlungen wiedergegeben werden.

Am Anfang jeder Entwicklung steht eine gründliche Analyse, die neben der fachlichen Seite auch Aspekte der Usability umfassen sollte. Dabei sollte versucht werden, von der größtmöglichen Bandbreite möglicher Nutzer und Nutzungskontexte auszugehen [RLL+04]. Beispielsweise sollten umfassend Informationen über Motivationen für eine eventuelle Nutzung, Erfahrungen und kulturelle Hintergründe der Nutzer, aber auch über physische Attribute (Alter, Hören,..) einholt werden.
Dafür können an dieser Stelle auch Wizard-of-Oz-Methoden (siehe Abschnitt 5.4.3) benutzt werden, um nicht nur herauszufinden, mit welchen Strategien Nutzer etwas versuchen, sondern auch mit welchen Worten sie dies tun.
Dabei sollte immer im Hinterkopf behalten werden, dass der Nutzer das entscheidende Korrektiv ist, und nicht überheblich manche Möglichkeiten als nicht angemessen ausgeblendet werden. Will der Nutzer beispielsweise Slang benutzen, soll er dies können, möchte er eher alles in drei Schritten machen statt in einem, soll das auch ermöglicht werden [LAB+05]. Ebenfalls sollte beobachtet werden, ob die Nutzer eher ganze Sätze oder Kommandos sprechen wollen. [WHHS05]

Aufbauend auf diesen Informationen sollte nun ein nutzerzentrierter Entwurf beruhen, der so prägnant und ausführlich wie nötig ist, aber trotzdem nach dem KISS-Prinzip („Keep it small and simple“) unnötige Formulierungen vermeidet [LAB+05]. Die Wortwahl der Systemäußerungen (das „Wording“) sollte so beschaffen sein, dass daraus dem Nutzer klar ersichtlich wird, welche Möglichkeiten er hat [Suh03]. Dadurch kann einfach die Struktur erfasst und schnell das Gefühl entwickelt werden, die Anwendung unter Kontrolle zu haben [Lar03]. Eine Optimierung auch der Prosodie durch TTS-Markup erscheint ebenfalls sinnvoll [Suh03].
Weiterhin sollte es möglich sein, die Struktur durch intuitive Shortcuts abzukürzen und so erfahrenen Nutzer schnellen Zugriff zu ermöglichen [WHHS05]. Alle diese Funktionen sollten aber trotzdem in eine Struktur eingebettet sein, die ähnliche Funktionen möglichst konsistent abbildet [WHHS05].

Wie bereits angedeutet, macht jedoch Fehlerbehandlung einen Großteil eines Dialogs aus [Lar03]. Um Fehler zu vermeiden, sollten zunächst ausdrucksstarke Bestätigungen gewählt werden, die dem Nutzer einen klaren Eindruck davon vermitteln, was sie im letzten Schritt ausgewählt haben. Im Fall eines Fehlers ist so den Nutzern sehr schnell klar, dass ein solcher aufgetreten ist [WHHS05]. Darüber hinaus sollte auch eine Hilfe angeboten werden, die möglichst kontextsensitiv den Nutzern (i.A. unter der Verwendung von Beispielen) klarmacht, welche Äußerungen an dieser Stelle erlaubt sind [Suh03].
Ist nun ein Fehler aufgetreten, sollte möglichst das System die Schuld dafür auf sich nehmen („Ich konnte sie leider nicht verstehen." statt „Ihre Äußerung war fehlerhaft."), damit sich der Nutzer auf die Fehlerbehebung konzentrieren kann [Rol05]. Weiterhin sollte nach der Fehlermeldung neben fokussierten Nachfragen (am besten mit Ja/Nein-Fragen, da diese sehr robuste Erkennung ermöglichen [Suh03]) auch die Möglichkeit zum rückgängig machen der Aktion („Undo") oder zum Übergang in eine alternative Struktur oder Modalität möglich sein [RLL+04].

So ein Modalitätsübergang stellt aber nicht die einzige Möglichkeit dar, Usability auch für multimodale Schnittstellen zu betrachten. Da im Auto meist eine zweite Modalität in Form der üblichen haptisch-grafischen Schnittstelle zur Verfügung steht, soll hier auch eine Arbeit zu Usability-Richtlinien dafür von Reeves et al. [RLL+04] betrachtet werden.
Multimodalität eröffnet beispielsweise bei der diskutierten Fehlerbehandlung ganz neue Möglichkeiten. Denkbar wäre prinzipiell mehrere Eingabewege zu ermöglichen, die im Fehlerfall, aber auch nach Präferenz, jederzeit gewechselt werden können. Nutzer können so schon während der normalen Benutzung eine Alternative kennen lernen, die sie benutzten können, wenn sie mit einem Fehler konfrontiert werden.[4] Daraus resultiert auch, dass ein Nutzer seine Modalität stets frei wählen können sollte. Weiter bedeutet dies aber auch, dass jeder Vorgang in der einen Modalität seine Entsprechung in der anderen Modalität haben muss. Wenn dies nicht der Fall ist, muss der Nutzer klar und eindeutig darauf hingewiesen werden, damit er kein falsches Bild von der Situation entwickelt.
Für die Informationsdarstellung empfehlen Reeves et al. adaptive Verfahren, um die zu präsentierende Informationsmenge dem Nutzern und der Situation anzupassen. Falls solche Verfahren nicht eingesetzt werden, sollten aber zumindestens nicht Informationen in beiden Modalitäten präsentiert werden, wenn der Nutzer in einer gewissen Situation sowieso beide Modalitäten nutzen muss. Dadurch können die jeweiligen Stärken der Modalitäten genutzt werden, beispielsweise eignet sich die haptisch-grafische Modalität vor allem für räumlich und parallele Informationen, während die auditive eher für Statusinformationen, Vorgänge serieller Bearbeitung, Alarmierung und Kommandointeraktion geeignet ist. Außerdem kann eine Modalität genutzt werden, die Bedienung der anderen Modalität zu erklären (z.B. ein grafisches Menü, was sprechbare Kommandos zeigt)[5].
Trotz dieser teils unterschiedlichen Nutzungsmöglichkeiten muss jedoch vor allem Konsistenz gewahrt werden. So muss eine Eingabe mit gleichen Werten über jede Modalität die gleichen Ergebnisse liefern. Da Spracherkennung immer fehlerbehaftet ist (siehe Abschnitt 2.3), kann dieses Ziel immer nur in Annäherung erreicht werden. Um so wichtiger ist es, dass Befehle überall, sogar über Anwendungsgrenzen hinweg, einheitlich gebraucht werden. Weiterhin muss sich das System in genauer zeitlicher Synchronität befinden, das heißt

[4] Dies ist zum Beispiel besonders hilfreich, wenn ein Nutzer keine Sprache benutzen möchte, um seine Privatsphäre zu wahren (wenn ein Dialog geheime Informationen enthält u.ä.).
[5] Neben der Hilfe wird dieses „What You See Is What You Can Speak"-Prinzip auch allgemeiner benutzt. So sollten alle auf dem Bildschirm sichtbaren Kommandos auch sprechbar sein [WHHS05].

beispielsweise, dass Nutzereingaben immer gleichzeitig in allen Modalitäten bestätigt sein müssen.

5.4 Untersuchungsmethoden

Für die Untersuchung von Usability-Kriterien existiert eine Vielzahl von Untersuchungsmethoden. So unterscheidet Preim [Pre99] grundsätzlich in

- Formale,
- Heuristische und
- Empirische Untersuchungsmethoden.

Während formale Untersuchungen auf formalen, regelbasierten Auswertungen basieren (und damit sehr genau, aber auch sehr beschränkt einsetzbar sind) und heuristische Evaluationen Expertenwissen einsetzen, um anhand von Checklisten und Erfahrung Usability-Probleme zu erkennen, nutzen empirische Methoden die Befragungen und Tests von potentiellen Nutzern solcher Systeme.
Unter dem Gesichtspunkt der bereits diskutierten fehlenden allgemeinen Richtlinien und Erfahrungen im Bereich der Sprachdialogsysteme versprechen nur empirische Methoden wirklichen Mehrwert. Im Verlauf der Arbeit wurde einige Methoden eingesetzt, um auf verschiedene Weisen Informationen über die Nutzer, ihre Erfahrungen und Wünsche, aber auch über ihre konkrete Benutzung von sprachgesteuerten MP3-Playern zu erhalten.
Nachfolgend sollen nun die einzelnen Methoden vorgestellt werden, und darauf eingegangen werden, was sie als geeignet für die Benutzung im Rahmen dieser Arbeit erscheinen ließ.

5.4.1 Befragung

Generell können mit Befragungen sehr gut subjektive Einschätzungen, Vorlieben, aber auch objektiv feststehende Daten (Alter, Geschlecht, Beruf u.ä.) abgefragt werden, da Befragungen die subjektiven Eindrücke des Befragten einfangen können. Objektive Eindrücke oder Leistungsparameter lassen sich dagegen schlecht damit erheben[Kir00] .
Die wohl bekannteste Form einer Befragung ist ein Fragebogen, doch kann ein Übergang zwischen schriftlicher Befragung und mündlichen Interview manchmal fließend sein. Trotzdem lassen sich für Befragungen folgende Typen unterscheiden [Opp92]:

individuell auszufüllende Fragebögen
: Dabei wird der Fragebogen durch den Interviewer präsentiert und eine kurze Einführung gegeben, danach bearbeiten die Teilnehmer die schriftlich fixierten Fragen allein. Dieses Vorgehen ist aufwendig, dafür ist die Stichprobe sehr genau beeinflussbar und es entsteht nur wenig Beeinflussung durch den Interviewer.

Gruppenbefragung mit Fragebögen
: Ähnlich der ersten Methode wird hier aber gleichzeitig eine (große) Gruppe (Schulklassen, Vorlesungen, Veranstaltungen) von Teilnehmer befragt, die Einführung kann so für alle gleichzeitig erfolgen und auch die Bearbeitung parallel. Gegenseitige Beeinflussung (Abschreiben, gegenseitiges Befragen) ist dabei aber möglich.

postalische Befragung mit Fragebögen
Dieses Vorgehen bietet die Möglichkeit zur Befragung von sehr vielen Personen mittels Fragebögen und ist dabei nicht ortsgebunden. Auch Internetbefragungen zählen in diese Kategorie. Dabei entsteht kaum Beeinflussung, aber meist kann nur mit einer geringen Rücklaufquote gerechnet werden. Missverständnisse bei Fragen haben wegen der fehlenden Nachfragemöglichkeit große Auswirkungen.

standardisierte Interviews
Diese entweder per Telefon oder persönlich durchgeführten Befragungen ermöglichen auch Diskussionen und sehr freie Fragen, da jederzeit ein Nachfragen möglich ist. Fragen müssen dabei nicht immer vorher 100%ig ausformuliert sein. Durch die sehr freie Form der Befragung besteht eine große Gefahr der Beeinflussung, gleichzeitig ist die Vorbereitung und Durchführung mit einem hohen Aufwand verbunden.

Mit Ausnahme der postalischen Befragungen[6] wurden im Laufe der Arbeit alle anderen Methoden eingesetzt.
Dabei eröffnen Gruppenbefragungen die einfachste Möglichkeit, schnell und effektiv große Mengen an Daten zu sammeln, bieten aber auch die größten Unsicherheitsaspekte und durch die eingeschränkte Möglichkeit der Nachfragen wirken sich kleine Fehler im Befragungsentwurf massiv auf das Ergebnis aus [Gri06].
Die eher persönlicheren Methoden von individuell auszufüllenden Fragebögen oder standardisierten Interviews eignen sich dagegen mehr, zielgenau Daten aus bestimmten Stichproben zu erheben und subjektive Einschätzungen zu etwaigen Evaluationen (siehe Abschnitt 5.4.2) zu erhalten.

Beim Aufbau einer Befragung empfiehlt Oppenheim [Opp92] die Bearbeitung von fünf Fragekomplexen:

Typ der Datensammlung
Entscheidung für einen der bereits diskutierten Typen von Befragungen.

Akquisemethoden für Teilnehmer
Methoden zur Gewinnung der Teilnehmer (Nutzung von Vorlesungen, Orte und Zeitpunkt der Aushänge, Design der Aushänge), Methoden zur Steigerung der Rücklaufquote (Zusicherung von Geheimhaltung und Anonymität, Professionalität des Layouts, Auswahl direkt interessierter Teilnehmer und Vollständigkeitserinnerungen).

Ordnung der Frageblöcke
Ordnen nach der internen Logik und der möglichen Reaktion der Teilnehmer, Balance der Frage-Typen.

Ordnung der Fragen innerhalb der Blöcke
Eher eingrenzend (allgemein zu speziell) oder offen. Vermeidung Vorwegnahme von Antworten durch vorherige Fragen.

Typ der Fragen
Offene (keine Antwortvorgaben) oder geschlossene (Multiple Choice) Fragen. Wann Zusatzfragen möglich, eventuelles zusätzliches Protokoll von Versuchsleiter.

[6] Ein Einsatz dieser Methoden war nicht nötig, da für die anderen Methoden genügend Teilnehmer und Testpersonal zur Verfügung stand.

Wichtig ist dabei, dass stets beachtet werden sollte, was am Ende mit der Befragung herausgefunden werden soll (Fokus auf „Need to Know“ statt auf „Nice to know“ [Res06a]).

Aus dieser Darstellung wird ersichtlich, dass die Erstellung von Umfragen recht aufwendig sein kann. Dieser Aufwand kann entweder durch eine konsequente Orientierung an Richtlinien, beispielsweise zur Fragebogen-Erstellung[7], oder aber durch eine Benutzung von allgemeinen, standardisierten Fragebögen erreicht werden. Der nachvollziehbare Nachteil solcher standardisierter Fragebögen besteht darin, dass sie eben keine spezifischen Informationen liefern, sondern nur allgemeine Kriterien ermitteln können. Aber gerade für allgemeine Usability Maße, über deren Kriterien Konsens herrscht (Abschnitt 5.1), erscheint ein Einsatz solcher Standard-Fragebögen sinnvoll.
Eine Möglichkeit eines solchen Bogens bietet sich mit dem Standard Usability Scale (SUS) von Brooke [Bro96] an, welcher schnell und effektiv eine Usability-Einordnung in der Industrie ermöglichen sollte. Er ist sehr robust und wurde schon oft benutzt und adaptiert. Mit seinen zehn einfachen Multiple Choice Fragen ist er auch für Teilnehmer äußerst einfach und unkompliziert auszufüllen. In einem Überblick über Standard-Fragebögen von UsabilityNet [Usa03], eines EU-geförderten Projekts um Usability und benutzerzentriertes Design zu promoten, wird deswegen eine starke Empfehlung für die Benutzung dieses Fragebogens, auch gegenüber standardisierten Methoden, gegeben.
Deswegen wurde der SUS-Bogen im Rahmen dieser Arbeit herangezogen, wenn allgemein Usability eingeschätzt werden sollte. In anderen Fällen wurden unter Beachtung der vorgestellten Kriterien eigene Befragungen entwickelt.

5.4.2 Nutzertest

Neben Befragungen stellen auch Nutzertests eine wichtige Säule für Untersuchungsmethoden für Usability dar.[8] Zwar sind auch Befragungen als Teil von Nutzertests denkbar, und wurden auch im Rahmen der Arbeit so eingesetzt, in der folgenden Darstellung sollen allerdings die Charakteristika der Tests selbst abgeleitet werden. Die Darstellung orientiert sich im Wesentlichen an den Ausführungen von Preim [Pre99].

Für einen solchen Test wird zunächst einen Testgegenstand benötigt, im Allgemeinen ist dies ein im Verlauf des Entwicklungsprozess entstandener Prototyp. Dieser kann verschiedenen Umfang und unterschiedliche Detailierungsstufen haben, da Prototypen nie die gesamte Funktionalität umsetzen, sondern immer nur soweit implementiert werden müssen, wie dies zu Testzwecken notwendig ist. Je nachdem an welcher Stelle diese Einschränkungen im Wesentlichen vorgenommen werden, können zwei Typen von Prototypen unterschieden werden:

Horizontale Prototypen reduzieren die Tiefe, in der die Funktionalität implementiert wird. Dadurch wird die gesamte Oberfläche umgesetzt, aber meist nur ein Teil oder

[7] Beispiele hierfür sind Wilsons Arbeit [Wil03] oder die White Papers von DSS Research ([Res06b], [Res06a]).

[8] Vielfach wird auch der Begriff Evaluation als Synonym für Nutzertests gebraucht. Allerdings ist die Abgrenzung des Begriffes unklar, unterschiedliche Definitionen erläutern ihn jeweils anders. Im Rahmen der Arbeit möchte ich im Weiteren den Begriff als bedeutungsgleiche Entsprechung für Nutzertests verwenden.

gar nichts der darunter liegenden Funktionalität. Hierbei ist eher ein Gesamteindruck gewinnbar.

Vertikale Prototypen reduzieren den Umfang der Funktionalität. Einige Funktionen werden komplett umgesetzt, so dass in diesen Bereichen ein realistischer Eindruck des Systems möglich wird.

Die beiden Varianten sind auch in Abbildung 5.1 dargestellt.

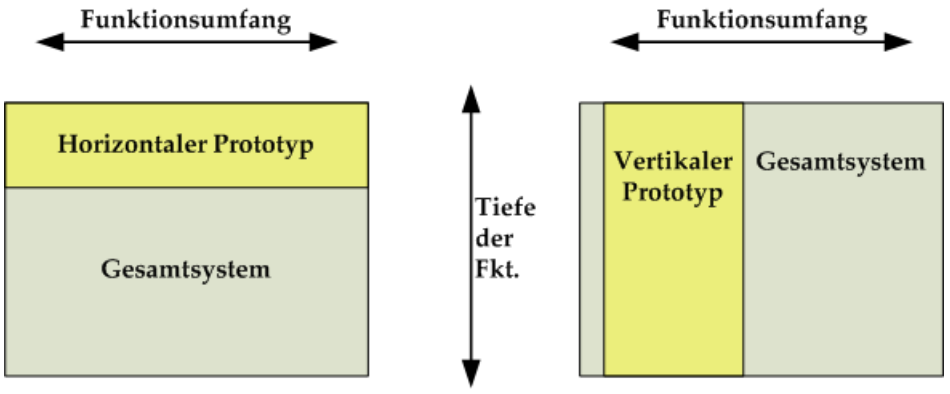

Abbildung 5.1: Horizontale und Vertikale Prototypen [Pre99]

Trotz der Verwendung von in ihrem Charakter eher horizontalen Prototypen in dieser Arbeit (da nicht die Integration und Funktion im Auto getestet werden, sondern der Gesamteindruck des Dialogs bewertet werden sollte), war trotzdem auch ein großer Anteil an Funktionalität umzusetzen bzw. zu integrieren, was auch an den Grundaufbau von Sprachdialogsystemen lag (siehe Kapitel 2). Eine Reduktion des Aufwands konnte entweder durch Verwendung der Wizard-of-Oz-Methode (siehe Abschnitt 5.4.3), aber auch durch die Beachtung einiger Methoden und Tricks von Nielsen [Nie89] erreicht werden.
So empfiehlt Nielsen unter anderem für Tests die Verwendung von besserer Hardware, den Einsatz von Pseudodaten (in Umfang und Komplexität reduzierte Datensätze), das Ignorieren von Sonderfällen und das Verwenden auch fehlerbehaftetem Code, wenn die Fehler nicht den Gesamteindruck zerstören.

Konkret wurde für den ersten Nutzertest die Wizard-of-Oz-Methode benutzt, für den zweiten Test aber die eben erwähnten Grundsätze umgesetzt. So wurde der Prototyp auf PC-Hardware entwickelt und getestet, die MP3-Datenbank reduziert und die Metadaten teilweise den Ansprüchen angepasst, nicht alle Sonderfälle ausprogrammiert und auch nicht alle Fehler beseitigt, solange diese nicht den eigentlichen Ablauf der Anwendung störten.

Neben dem Testgegenstand hat jeder Nutzertest Ziele. Häufig geht es darum, Schwachstellen eines Systems festzustellen oder mehrere Varianten miteinander zu vergleichen. Schwerpunkte ranken sich dabei vor allem um folgende Punkte:

Usability-Kriterien
Hilft das System tatsächlich Abläufe effektiver, effizienter und zufriedenstellender zu gestalten? Ist eine anderweitige qualitative Verbesserung möglich?

Akzeptanz
Wird das System von den Nutzern angenommen?

Integrationsfähigkeit
Kann die Benutzung des Systems mit anderen Systemen aus dem Umfeld der Benutzer integriert werden. Fügt es sich in Arbeits- oder Lebensabläufe nahtlos ein?

Fragestellungen im Rahmen der Arbeit legten gerade auf diese drei Punkte einen Schwerpunkt der Untersuchung. So musste festgestellt werden, ob das System allgemein benutzbar ist, aber auch, ob die Nutzer ein solches System (Sprachbedienung für einen MP3-Player im Auto) überhaupt haben wollen, und ob sie sich nicht doch weiter das gewohnte haptisch-grafische Interface bedienen wollen. Schließlich stellte sich die Frage, ob die Bedienphilosophie so überhaupt der Erwartungshaltung und inneren Haltung zur Musik entsprach. Diese Problematik wurde schon umfassend in Kapitel 3 diskutiert, war aber natürlich auch Hintergedanke bei den Untersuchungen im Rahmen der Arbeit.

Schließlich wurden neben Testgegenstand und Zielen Testkriterien definiert, deren Auswertung Rückschlüsse über die Untersuchungsgegenstände zulassen. Diese lassen sich in objektive und subjektive Kriterien untergliedern, wobei objektive messbare oder beobachtbare Kriterien umfassen (Durchführungsdauer, Fehlerrate, Fahrablenkung, Arbeitsergebnisse), während subjektive Kriterien durch Befragung der Nutzer erhoben werden.
In einer Untersuchung sollten möglichst viele objektive Kriterien erhoben werden, da diese leichte Vergleichbarkeit auch mit anderen Experimenten ermöglichen. Die Erhebung solcher Kriterien wird durch die Benutzung eines so genannten Usability-Labors erleichtert, da hier meist umfangreiche Technik auch zur erweiterten Auswertung (Video, Eyetracker u.a.) zur Verfügung stehen.
Im Rahmen der Arbeit wurden für jede Untersuchung verfügbare objektive Kriterien erfasst und ihre Benutzung kritisch bewertet, da meist eine konsequente Benutzung dieser zu Einschränkungen im Umfang der Untersuchungen geführt hätte. Darüber hinaus wurden die Tests mit Video dokumentiert, und später die Äußerungen der Nutzer umfangreich ausgewertet, so dass so eine umfangreiche, auch objektiv auszuwertende, Datenbasis entstand. Zusätzlich wurden Befragungen der Nutzer durchgeführt.

Um erfolgreich einen Test durchzuführen, müssen nun noch Tester, Teilnehmer und eine geeignete Testumgebung betrachtet werden. Tester sind Personen, die den Test vorbereiten, durchführen und auswerten. Meist wird in diesem Zusammenhang auch von Versuchsleitern (VL) gesprochen. Diese Personen müssen möglichst neutral und unbeeinflusst den Test durchführen
Testpersonen sind Personen, die den Test als Teilnehmer ausführen sollen. Mit deren Akquise beschäftigte sich bereits Abschnitt 5.1, zusätzlich muss die Anzahl der Versuchspersonen festgelegt und ein Zeitplan erstellt werden.
Die Testumgebung sollte so weit wie möglich dem realen Einsatzumfeld entsprechen, was das für den Automobilbereich bedeutet, wurde bereits in Abschnitt 5.2 diskutiert.

Nachdem dann alle Voraussetzungen erfüllt sind, sollte noch einmal überprüft werden, ob der Nutzertest bestimmte Qualitätskriterien einhält:

Glaubwürdigkeit
Ist es durch die Auswahl der Testpersonen (ähnlich der Zielgruppe, unbeeinflusst) und die Durchführung (realistische Aufgaben) plausibel, dass die Ergebnisse zustande kamen und Aussagen über die Ziele der Studie ermöglichen? Sind die zu untersuchenden Sachverhalte nicht von anderen Faktoren beeinflusst worden?

Übertragbarkeit
Ist es möglich, die Ergebnisse auf andere Sachverhalte zu übertragen?

Zuverlässigkeit
Steht zu erwarten, dass mit mehr Testpersonen und mehr Zeit keine anderen Ergebnisse zu erwarten wären?

Bei der Erstellung der Testkonzepte wurde stets darauf geachtet, inwieweit das skizzierte Konzept diesen Kriterien entspricht. Die mögliche Übertragbarkeit der Ergebnisse über die Grenzen des Musikauswahl hinaus wird in Abschnitt 8.4 noch einmal ausführlicher diskutiert.

5.4.3 Wizard-of-Oz-Test

Bei der Entwicklung von Sprachdialogsystemen stellte sich, angesichts fehlender Verfügbarkeit von allgemein gültigen Dialogprinzipen (siehe Abschnitt 5.3), die Frage, wonach die Entwicklung des Dialogs ausgerichtet werden soll. Früh wurde erkannt, dass die Mensch-Mensch-Kommunikation nicht die maßgebliche Größe sein konnte, da Menschen immer situationsangepasst sprechen. So können sich Menschen auch verständigen, wenn sie die Sprache nicht richtig beherrschen (Kinder, Ausländer), indem sie sich den Kommunikationsbedingungen anpassen. Dies ist auch bei der Kommunikation zwischen Mensch und Maschine der Fall. Menschen benutzen bei dieser Kommunikationsart zumeist eine einfache Sprache, verbunden mit mehr kommandosprachlichen Äußerungen.[Pet04]
Trotz dieser Tendenz ist aber nicht klar, welche Formulierungen im konkreten Einzelfall ideal wären. Um dies herauszufinden, müssen Testpersonen in eine möglichst realistische Situation versetzt werden, in der intuitives Verhalten beobachtbar wäre. Dafür hat sich im Laufe der Jahre eine Methode mit dem Namen Wizard-of-Oz-Tests (oder kurz: WOZ-Tests) herausgebildet.

Der Name leitet sich dabei von der Figur des Zauberers (engl. wizard) aus dem klassischen Kinderbuch „The Wonderful Wizard of Oz“ von Lyman Frank Baum [Bau00] ab. Der Zauberer ist dort ein gewöhnlicher Mann, der hinter einem Vorhang eine Maschine bedient. Von außen sieht es jedoch so aus, als wäre er ein mächtiger Zauberer.

Genau dieses Prinzip liegt auch den WOZ-Tests zugrunde, bei denen den Versuchspersonen vorgegaukelt wird, sie interagierten mit einem voll funktionstüchtigen Sprachdialogsystem. In Wirklichkeit sitzt in einem anderen Raum („hinter dem Vorhang“) eine weitere Person (der „Wizard“), welche die Spracherkennung statt einem Computer durchführt, und je nach Äußerung eine entsprechende Systemausgabe abschickt. Dieser prinzipielle Aufbau wird in Abbildung 5.2 verdeutlicht.
Die Äußerungen des Nutzers werden also nicht von seinem lokalen System verarbeitet,

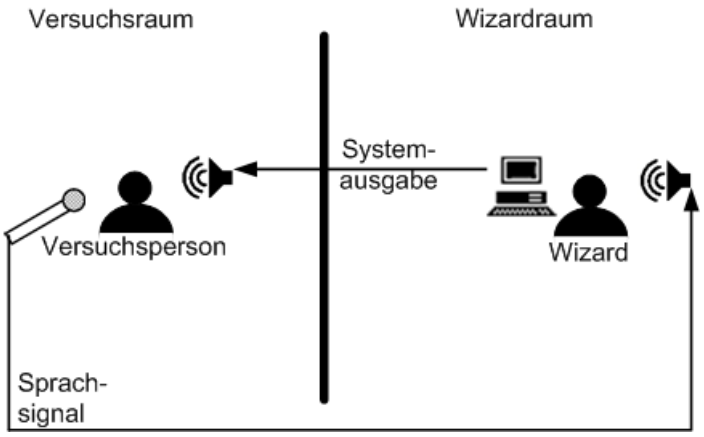

Abbildung 5.2: Wizard-of-Oz Methode (nach [Pet04])

sondern von dem Wizard. Für den Nutzer wirkt dies, Schnelligkeit und Korrektheit vorausgesetzt, wie die lokale Reaktion seines Systems.

Petrik diskutiert in seiner Arbeit über Wizard-of-Oz-Systeme für Sprachdialogsysteme [Pet04] Vorteile, Nachteile und Verwendungsmöglichkeiten dieser Methode. Dieser Darstellung soll im Folgenden gefolgt werden.
Wesentliche Vorteile ergeben sich vor allem in folgenden Punkten:

Authentizität der Daten
: Die realistische Simulation provoziert realistische Nutzerreaktionen, wie sie auch bei der Interaktion mit einem fertig implementierten System auftreten würden.

Einfache Durchführbarkeit
: Bei der Verwendung einer geeigneten WOZ-Simulationsumgebung können Hardware-Anforderungen gering gehalten werden, da Teile des Systems simuliert werden können.

Einfache Erstellung von Prototypen
: Durch das Fehlen komplexer Systemkomponenten ist die Erstellung von Prototypen meist vergleichbar einfach.

Ermöglichung frühzeitiger Tests
: Durch die einfache Erstellung von Prototypen und die einfache Durchführbarkeit ist es möglich, schon in frühesten Projektphasen Prototypen zu erstellen.

Neben diesen Vorteilen gibt es es allerdings auch eine Reihe von Probleme bei der Verwendung der Wizard-of-Oz-Methodik:

Wizard-Role sehr beanspruchend
: Der Wizard hat drei Aufgaben gleichzeitig zu bewältigen: Aufnahme der Sprache, Interpretation und Erzeugung der passenden Systemausgabe. Dies hat weiterhin in möglichst geringer Zeit und konsistent zu geschehen, um die Illusion beim Nutzer zu erhalten. Eine geeignete WOZ-Simulationsumgebung kann diese Probleme reduzieren helfen.

Wizard nicht fehlerfrei
Da der Wizard ein Mensch ist, macht er auch Fehler, die Abweichungen von der Konsistenz des Gesamtsystems bedeuten. Da aber auch eine gewisse Fehlerquote der Spracherkennung erwartet wird, ist dies bei einzelnen Fehler nicht zu bedeutend.

Ethische Bedenken
Da die Testperson bewußt fehlinformiert wird über den wahren Aufbau des Experiments, ist dies zu mindestens ethisch diskussionswürdig. Wenn die Testpersonen nach dem Experiment nachträglich informiert werden, kann dies zu mindestens abgemildert werden.

Eingangs wurde über ein Anwendungsbeispiel die Verwendung der Wizard-Of-Oz-Methode motiviert, dies ist jedoch nicht die einzigste Möglichkeit, Vorteile aus der Verwendung dieser zu ziehen. Nachfolgend sind einige denkbare Einsatzmöglichkeiten aufgezeigt:

Nutzerakzeptanz testen
Frühe Designvorschläge lassen sich so einfach und effektiv testen, bevor weiterer Aufwand in das Projekt investiert wird.

exploratives Dialogdesign
Auch ohne eine fest vorher festgelegte Struktur kann ein WOZ-Test eingesetzt werden, um explorativ ein bestimmtes Dialogdesign zu ermitteln. Dafür wird den Versuchspersonen innerhalb des Tests viel Freiheit (durch ein gewisses Maß an Varianz der Systemrückmeldungen) gelassen und die Beobachtung der Personen im Vordergrund gestellt.

Sammeln von Sprachdaten
Das Sprachdaten verschiedener Personen für die Spracherkennung wichtig sind, wurde bereits in Abschnitt 2.3 diskutiert. Doch auch für das Dialogdesign kann das Sammeln von Sprachdaten interessant sein, um beispielsweise typische Wortwahl zu ermitteln.

Testen bereits existierender Dialoge
Eine weitere Möglichkeit wäre der Nachbau fertiger Systeme, um unter Ausblendung der Spracherkennung belastbare Aussagen zum Dialogdesign zu erhalten.

Voraussetzung für die Verwendung der Wizard-of-Oz-Methode ist der Einsatz einer geeigneten WOZ-Simulationsumgebung, die einerseits eine schnelle und effektive Reaktion im Test ermöglicht (um die Illusion zu wahren), gleichzeitig aber offen ist für alle hier genannten Anwendungszwecke und die Entwicklung eines Tests trotzdem sehr einfach hält. Gleichzeitig sollte sie auch für den Test multimodaler Systeme geeignet sein.
Zu diesem Zwecke existieren eine Reihe von Systemen wie SUEDE [KSC+00] oder NEIMO [SC93], im Rahmen dieser Arbeit wurde jedoch das im Verlauf der Arbeit von Petrik entstandene WOZ-Tool benutzt, da es die genannten Anforderungen am Besten vereinigen konnte. Insbesondere die Abkehr von einem strikten, vorgegeben Dialogmodell hin zu einer hierarchischen Anordnung von inhaltlich zusammengehörigen Prompts machte ein sehr exploratives Design des Wizard-of-Oz-Tests möglich.

5.5 Vorgehensmodell

Die im vorletzten Abschnitt vorgestellten Prinzipien bilden die Richtlinien, an denen sich eine nutzerzentrierte Entwicklung von Sprachdialogsystemen orientieren sollte. Doch wie sollte ein konkretes Vorgehensmodell (welches die im letzten Abschnitt vorgestellten Untersuchungsmethoden benutzt) aussehen, um diese Anforderungen zu integrieren?

Drei grundlegende Strategien für solche Modelle wurden schon 1985 von Gould und Lewis [GL85] zusammengefasst:

- Frühzeitiger Fokus auf Nutzer und Aufgaben
- Benutzung Empirischer Methoden
- Iteratives Design

Diese Strategien bilden eine Grundlage zu Vorgehensmodellen im Bereich von Usability Engineering, konkreter zeigen jedoch Button und Dourish [BD96] für die Erstellung eines solchen Vorgehensmodells drei gangbare Wege auf:

1. Hinzufügen von Kognitionswissenschaftlern zu Entwicklungsteams, um deren Vorschläge im Prozess mit zu berücksichtigen
2. Einfügen von Methoden und Techniken des Usability Engineering in bestehende Entwicklungsprozesse
3. Neuentwicklung des gesamten Entwicklungsprozesses rund um Wissen, Methoden und Techniken des Usability Engineering

Nachfolgend sollen nun die beiden letzteren Möglichkeiten auf Eignung für das Vorgehensmodell dieser Arbeit untersucht werden, unter Beachtung des Gesichtspunkts, dass für Sprachdialogsysteme noch kein softwaretechnologisches Standardvorgehen (welches angepasst werden könnte) zur Verfügung steht.

Zumindestens einen Vorschlag für ein solches Modell präsentierten Bernsen et al. [BDD97] in ihrem Vorgehensmodell für Sprachsysteme, welches sie aus allgemeinen Software-Vorgehensmodellen entwickelten und welches in Abbildung 5.3 dargestellt ist.
In diesem Modell bilden entweder Forschungsideen oder kommerzielle Anforderungen die Basis für eine umfassende Bestandsaufnahme, bei der neben strategischen Zielen, technischer Machbarkeit und Ressourenplanung auch Befragungen zukünftiger Nutzer einbezogen werden. Aus diesem Prozess leitet sich das Pflichtenheft ab und gleichzeitig sollten gleich Evaluationskritierien formuliert werden, die das endgültige System erfüllen sollte.
Anschließend kann in eine erste Phase von Analyse und Entwurf eingetreten werden, die sich aber nach jedem Evaluationszyklus wiederholen sollte[9]. Von dort kann direkt in die Implementation eingetreten werden, oder eine Simulation von Teilen oder der Gesamtheit des Systems durchgeführt werden und damit weitere Ergebnisse zur Verbesserung des Systems gewonnen werden.

[9] Dies ist so aus der Grafik von Bernsen et al. nicht ersichtlich, wird aber in der Arbeit erläutert

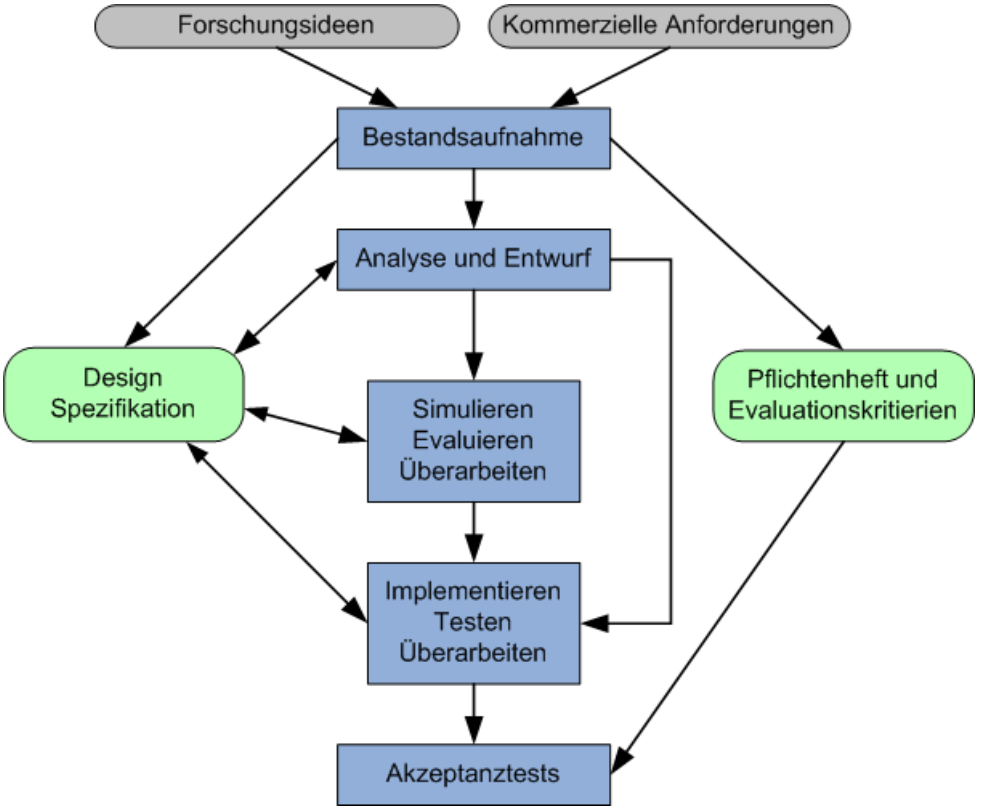

Abbildung 5.3: Software-Vorgehensmodell für die Entwicklung und Evaluation von interaktiven Sprachsystemen [BDD97]

Während aller Phasen wird ständig eine Design Spezifikation mit Ideen und Ergebnissen aus Evaluation und Entwurf ergänzt, so dass diese für die finale Implementation alle relevanten Daten enthält.

Dieses Vorgehen lehnt sich also an den zweiten Ansatz von Button und Dourish an, ein vorhandenes Entwicklungsmodell anzupassen, um Usability-Methoden und Techniken zu integrieren. Natürlich wäre auch denkbar, ein Vorgehensmodell des Usability Engineering zu benutzen (dritter Ansatz von Button und Dourish). Dieses müsste dann aber noch an die speziellen Voraussetzungen bei Sprachdialogsystemen angepasst werden.
Ein solche Möglichkeit würde sich mit dem Vorgehensmodell des Usability Engineering Lifecycle von Mayhew [May99] anbieten. In diesem sehr komplexen Modell wird detailliert beschrieben, wie eine Anforderungsanalyse, Zyklen von Konzeption, Test sowie weiterer Entwicklung und sogar die Verbesserung nach der Installation aussehen könnten (siehe Abbildung 5.4).

Dabei macht Mayhew in jeder Stufe der Entwicklung klar, welche Schritte erforderlich sind, um in die nächste Stufe der Entwicklung eintreten zu müssen. Mit jeder Entwicklungsstufe nimmt dabei der Detaillierungsgrad stetig zu, vom konzeptionellen Modell über das Screen Design bis zum gesamten Interface.
Diese Phasen können zwar, abgekürzt werden, jedoch bleibt auch dann die Komplexität dieses Modells ein Problem, da eine Schablonen-artige Anwendung wegen der Akzentsetzung Richtung grafisch-haptischer Interfaces nicht möglich ist. Zwar diskutiert Klante

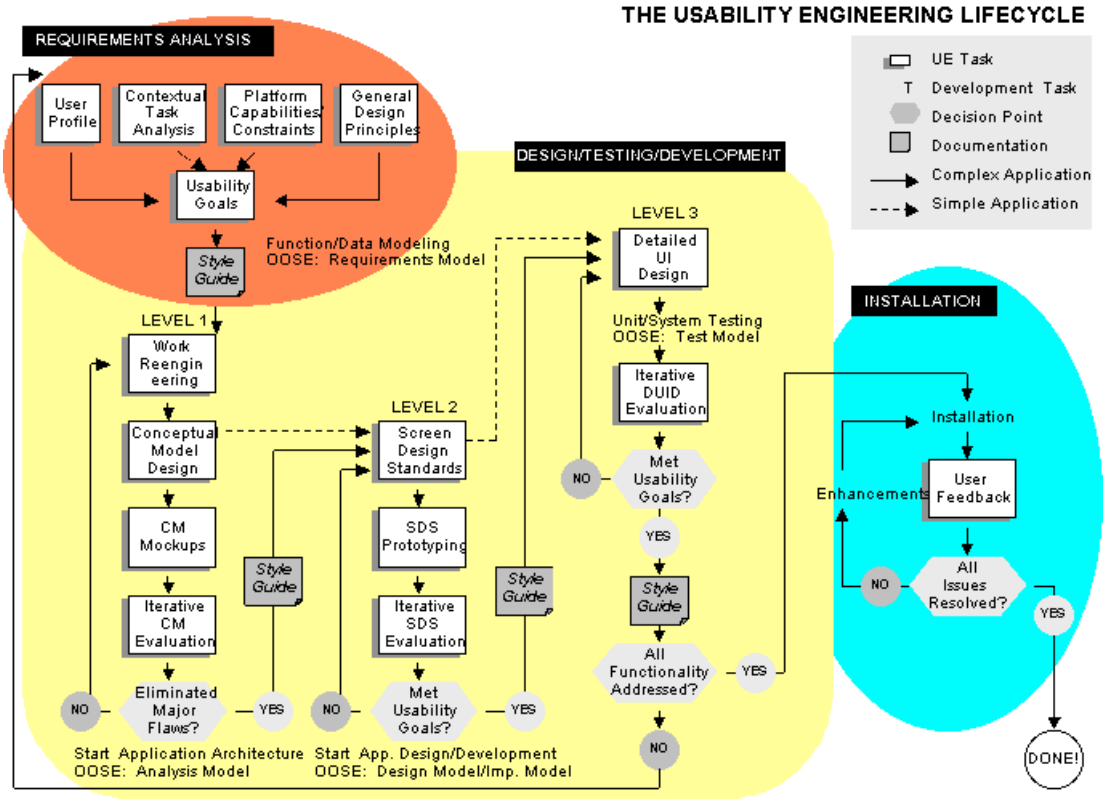

Abbildung 5.4: The Usability Engineering Lifecycle [May04]

[Kla03a] eine Anwendung dieses Modells auf die Entwicklung auditiver Benutzeroberflächen, allerdings ist diese Betrachtung sehr fixiert auf die Verwendung von Hearcons.[10] Infolgedessen müsste also trotzdem Mayhews Modell umfangreich betrachtet und angepasst werden, um eine Anwendbarkeit auf allgemeine Sprachdialogsysteme sicherzustellen.

Diese Überlegungen führten dazu, von der Verwendung dieses Modells Abstand zu nehmen, und stattdessen das vorher erläuterte Modell von Bernsen zu verwenden und für die Verwendung im Rahmen der Arbeit zu modifizieren.

Zuvor war es jedoch nötig, Voraussetzungen, Festlegungen und vorgegebene Entwicklungsschritte, die aus der Aufgabenstellung der Arbeit abgeleitet werden konnten, zu diskutieren. So stand am Anfang der Entwicklung lediglich eine vorhandene technische Basis [WHHS05], der kommerziellen Anspruch der Firma, einen möglichst intuitiven MP3-Dialog im Rahmen der Arbeit zu erhalten und einige Ideen aus der Forschung, wie effektiver sprachlicher Zugriff auf große Datenmengen aussehen könnte. Im Rahmen der Aufgabenspezifikation wurde dann festgelegt, dass zur Analyse eine Befragung durchgeführt werden sollte. Danach sollten in zwei Stufen Nutzertests durchgeführt werden. Zuerst sollten einzelne Konzepte mit Nutzern getestet werden.. Anschließend sollte aus den Ergebnissen ein Gesamtkonzept abgeleitet und prototypisch umgesetzt werden. In einem abschließenden Test sollte dann die erreichte Qualität der Umsetzung überprüft werden.

[10] Dieses Vorgehensmodell gab jedoch Anregungen, an welcher Stelle und wie die Verwendung von nonverbalen Interaktionselementen zu bedenken wäre, wie dies bereits am Ende von Abschnitt 4.4 diskutiert wurde.

Dieser Aufbau entsprach in seiner Gliederung der drei Unterstufen ziemlich genau dem Vorgehen, welches Bernsen et al. in ihrem Vorgehensmodell skizziert hatten. Wird nun die Anwendung des Modells auf das Vergehen in dieser Diplomarbeit betrachtet, entsteht ein Vorgehensmodell, wie es in Abbildung 5.5 dargestellt ist.[11]

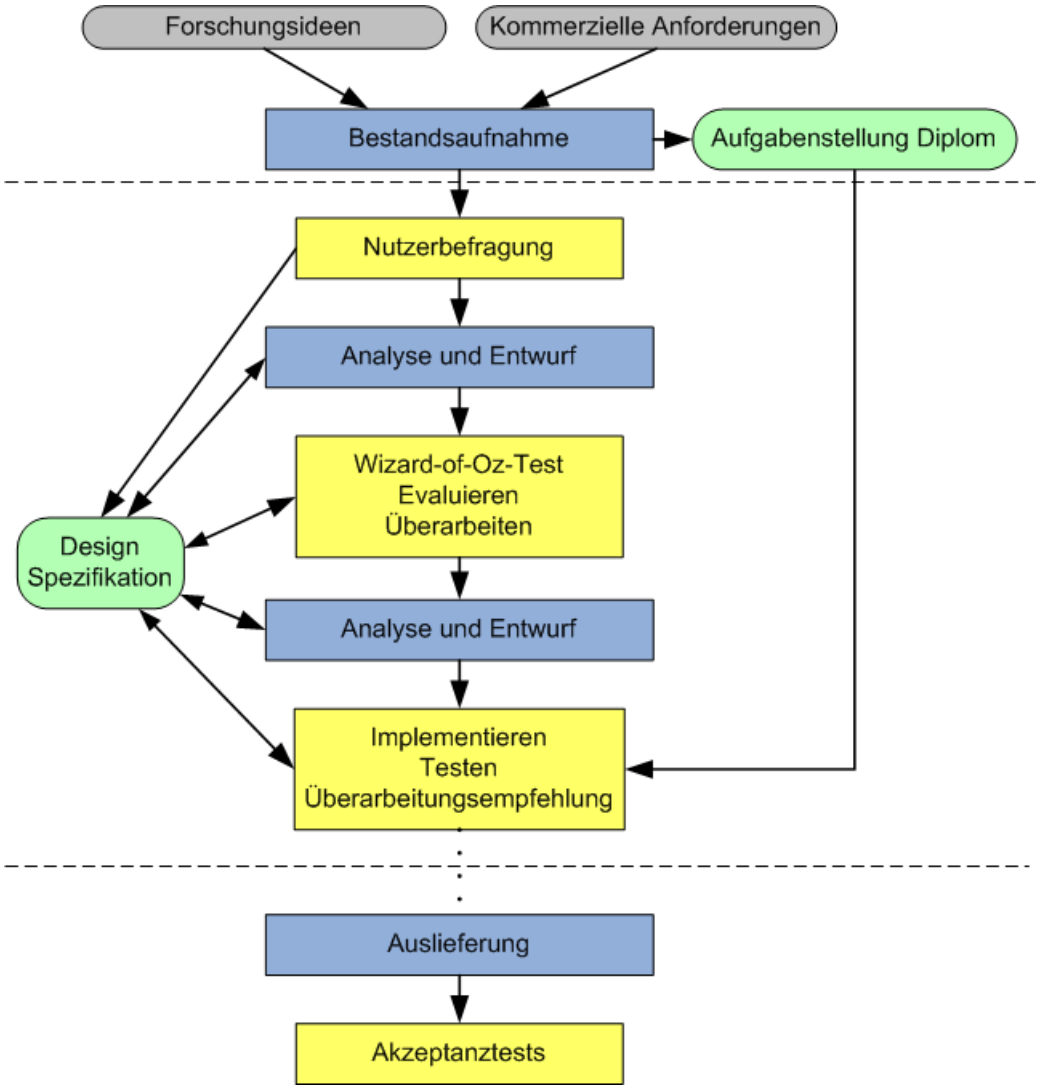

Abbildung 5.5: Vorgehensmodell dieser Arbeit für die Entwicklung und Evaluation eines Sprachdialogsystems - Idealvorstellung

So wurde aus den verschiedenen Anforderungen der Arbeit schon vor dem Beginn die Aufgabenstellung abgeleitet, die als Pflichtenheft die Zielstellung der Arbeit repräsentierte. Ausgelagert aus der Bestandsaufnahme sollte jedoch eine Nutzerbefragung durchgeführt werden, die als Basis für den Entwurf des Dialogs für die Simulation in Form eines Wizard-of-Oz-Tests (siehe Abschnitt 5.4.3) und die prototypische Umsetzung für den abschließenden Test dienen sollte. Ein Akzeptanztest wurde im Rahmen der Arbeit nicht durchgeführt, da er eine Auslieferung des Prototypen vorausgesetzt hätte, die in absehbarer Zeit jedoch

[11] Dabei kennzeichnet der Bereich zwischen der oberen und der unteren gestrichelten Linie die im Rahmen der Diplomarbeit durchgeführten Phasen des Modells.

nicht zu erwarten war.

Jedoch konnte aus Zeitgründen diese Idealvorstellung nicht eingehalten werden und so mussten Nutzerbefragung und Wizard-of-Oz-Test teilweise parallel durchgeführt werden, was dazu führte, dass die Erkenntnisse des Fragebogens bei der Vorbereitung des Wizard-of-Oz-Tests nicht vollständig zur Verfügung standen und somit nicht verwendet werden konnten, wie dies Abbildung 5.6 veranschaulicht.

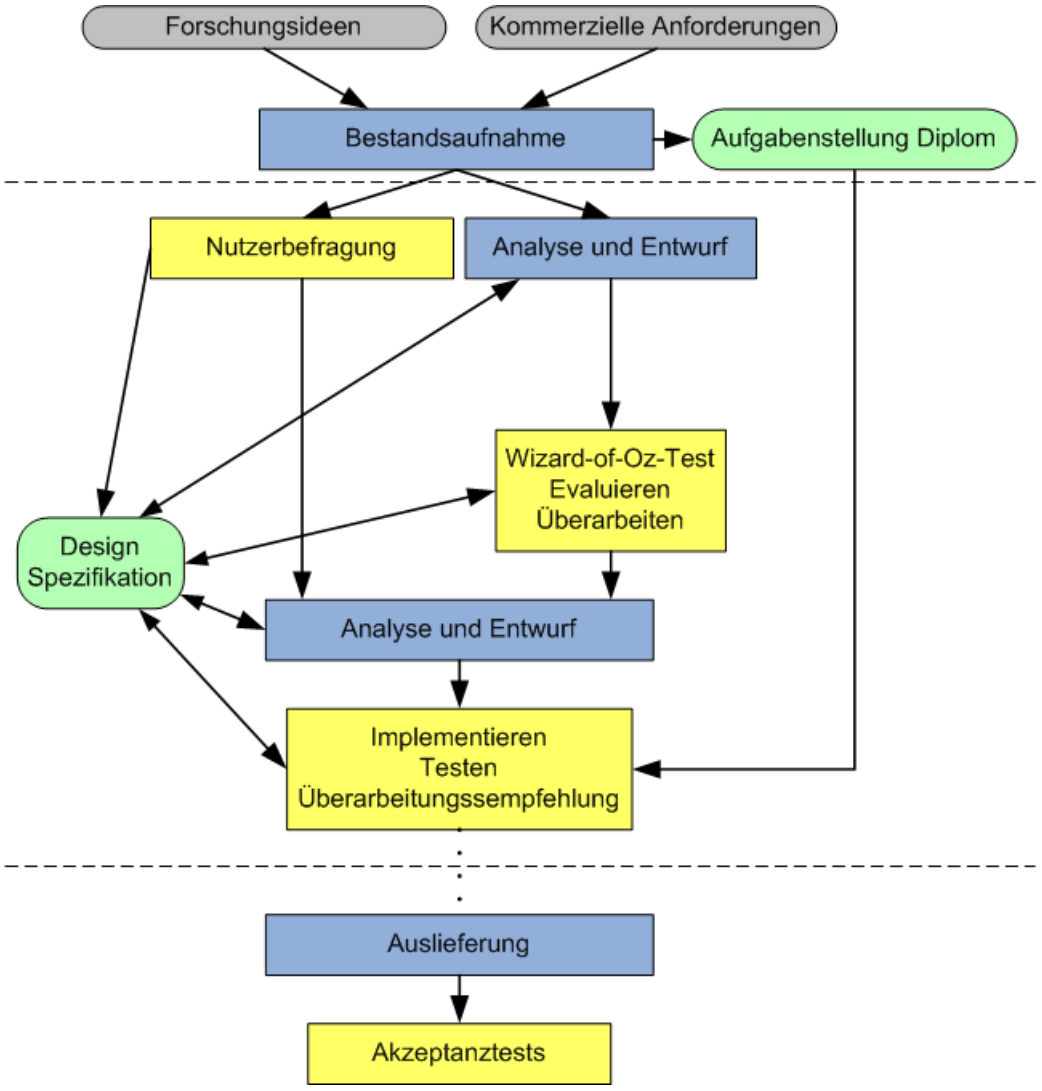

Abbildung 5.6: Vorgehensmodell dieser Arbeit für die Entwicklung und Evaluation eines Sprachdialogsystems - realistische Variante

In jeder Analysephase wurde dabei aus Usability-Vorgaben (wie sie in Abschnitt 5.2 und 5.3 diskutiert wurden), und vorliegenden Erkenntnissen über Musikauswahl aus der Literatur (wie in Abschnitt 3.4 dargestellt) und aus bereits durchgeführten Tests im Rahmen der Diplomarbeit jeweils Schlussfolgerungen für den weiteren Entwurf gezogen und weitere zu klärende Fragen festgelegt. Diese und jeweils auch die Ergebnisse aus den Tests wurden in einer Design Spezifikation zusammengetragen. Die jeweils aktuelle Design Spezifikation

implizierte dabei jeweils immer eine gewisse Vorstellung des Systems, aus der am Ende eine Systemhistorie (dargestellt in Abschnitt 7.1) ableitbar war.
Wie bereits in Abschnitt 4.4 erwähnt, wurden Betrachtungen über die Art und den Umfang der Integration non-verbaler Interaktionselemente erst entschieden, als durch Fragebogen und Wizard-of-Oz-Test eine Vorstellung entwickelt werden konnte, wie ein prinzipielles Systemdesign aussehen könnte.

6
Fragebogen und Wizard-of-Oz-Test

Im Rahmen der Arbeit wurden für die Bestandsaufnahme und Analyse eine Nutzerbefragung in Form eines Fragebogens und eines Wizard-of-Oz-Tests durchgeführt.
Beides diente, neben der Erfassung nützlicher Fakten über Nutzungshäufigkeiten und MP3-Erfahrung, dem Erkunden der „Denkwelten“ der Benutzer. Da Musik, wie bereits im Kapitel 3 diskutiert, ein zutiefst subjektiv geprägtes Thema ist, sollte unter anderem geklärt werden, was potentielle Nutzer mehrheitlich mit Musik verbinden und wie sie diese nutzen. Daraus sollte beispielsweise abgeleitet werden, ob eine Musikauswahl basierend auf Verzeichnissen, Alben oder gänzlich unscharf auf Kriterien wie Genre[1] oder Stimmung aufbauen sollte. Eine weitere interessante Frage ist, ob Musikauswahl schnell und unkompliziert, aber dafür ungenauer; oder etwas langsamer und hierarchielastiger, aber genauer stattfinden sollte.

Die Befragung mittels Fragebogen wurde parallel zur Vorbereitung des WOZ-Tests durchgeführt, so dass ausgewertete Daten des Fragebogens nur teilweise bei der Erstellung des WOZ-Tests zur Verfügung standen. Wurden Erkenntnisse dennoch genutzt, wird darauf im Text eingegangen.
Die Systemideen, die vor und für den WOZ-Test formuliert wurden, sind in der Systemhistorie im nächsten Kapitel dargestellt.

Im Folgenden werden Voraussetzungen und Anforderungen für die jeweilige Untersuchungsmethode, die Umsetzung und die Ergebnisse geschildert. Am Ende dieses Kapitels wird eine Bilanz gezogen, welche die Voraussetzung für die prototypische Umsetzung bildet.

[1] Die Kategorisierung von Genre als unscharfem Kriterium resultierte aus den Vorüberlegungen zur Musikauswahl in Kapitel 3. Aufgrund der Ergebnisse des WOZ-Tests (siehe Abschnitt 6.2.3) wurde später eine andere Betrachtungsweise angenommen. An dieser Stelle wird jedoch Genre weiter als unscharfes Kriterium begriffen, da aus dieser Betrachtungsweise die Entscheidungen im Rahmen dieses Kapitels verständlich werden.

6.1 Fragebogen

Die erste Befragung mittels eines Fragebogen sollte vor allem der Eingrenzung des Themas dienen, Faktenwissen sollte aufgebaut werden und schon erste Fragestellungen den Nutzern zur Entscheidung vorgelegt werden.

6.1.1 Voraussetzungen & Anforderungen

Grundlage des Fragebogens war ein Brainstorming zu möglichen interessanten Themengebieten. Kamen dabei Fragen auf, welche nicht erlebbar in den Wizard-of-Oz-Test integrierbar waren, wurden diese in den Fragebogen übernommen.
Der Fragebogen sollte im Rahmen von Vorlesungen von den Studenten ausgefüllt werden, was dem Prinzip der Gruppenbefragung von Oppenheim [Opp92] entspricht. Dies wurde vor allem angedacht, um eine große Anzahl von Teilnehmer zu erreichen, aber auch um den Aufwand der Gewinnung von Teilnehmern gering zu halten. Bei der Stichprobe konnte von einer weitgehenden Vertrautheit mit dem Thema MP3 ausgegangen werden. Dies war hilfreich, da die Befragung auch ideengenerierend wirken sollte.
Um auch ungewöhnliche Antworten und Ideen zu berücksichtigen, war es nötig den Befragten viel Freiheit in der Beantwortung zu ermöglichen.

6.1.2 Umsetzung

Der für die Befragung genutzte Fragebogen (siehe Anhang B.1.1) gliederte sich in drei Teile:

- Allgemeine Daten
- MP3
- Sprachbedienung im Auto

Fragen zur Sprachbedienung folgten bewußt erst nach dem Teil zu MP3. So konnte sichergestellt werden, dass die Nutzer aus ihrer normalen Nutzung von MP3 und nicht aus dem Blickwinkel der Sprachbedienung urteilen.

Die Fragen des **allgemeinen Teils** dienten zur Ermittlung von Informationen über die Befragten selbst (siehe Tabelle 6.1). So hätte eine Einteilung in Subgruppen ermöglicht werden können, wenn Auffälligkeiten hinsichtlich Alter oder technischen Erfahrungen aufgetreten wären.

Im **MP3-Teil** sollte unter anderem herausgefunden werden, welche Art und welchen Umfang die MP3-Nutzung bei den Befragten hat, woraus sich eventuell gewisse Prägungen und Vorlieben hätten ableiten lassen. Dazu sollte auch weitere Fragen zu üblichen Musiknutzung im Auto und Erfahrungen mit Playlisten dienen. Um eine Vorstellung von Struktur und Größe üblicher MP3-Sammlungen zu bekommen, wurde schließlich nach Sortierung und Größe der MP3-Sammlung gefragt.

Fragebogen	Teiln.	Alter	Geschlecht	Hintergrund	tech. Interesse
	227	90% <26	79% ♂	87% IT	81% hoch/sehr hoch

Tabelle 6.1: Zusammensetzung Stichprobe Fragebogen

Der Teil zu **Sprachbedienung** sollte der Bewertung möglicher Funktionen eines sprachbedienten MP3-Players dienen. Zunächst wurde jedoch gefragt, welche Funktion die Nutzer am liebsten per Sprache bedienen würden. Damit sollte herausgefunden werden, welche Funktionen die Nutzer intuitiv einer Sprachbedienung zuordnen. Um auch hedonische Überlegungen[2] zu integrieren, wurde auch eine Frage nach einer „besonders coolen Funktion“ gestellt.
Auf der nächsten Seite und damit für den Nutzer beim Ausfüllen der vorangegangenen Fragen nicht einsehbar, wurden dann die mögliche Funktionen vorgestellt, deren Wichtigkeit eingeschätzt werden sollte. Eine Frage nach der Bezahlung eines möglichen Aufpreises sollte abschließend darüber Aufschluss geben, ob überhaupt prinzipielles Interesse an einem solchen System besteht.+

Um einen möglichst breiten und umfassenden Eindruck zu gewinnen, wurde versucht, die Beeinflussung der Teilnehmer möglichst gering zu halten und wo möglich, offene Fragen zu stellen.
Für die konkrete Umsetzung wurden zwei Vorbedingungen festgelegt, so mussten die Befragten im Besitz eines Führerscheins und deutsche Muttersprachler sein. Ersteres stellte sicher, dass sie mit der Situation im Auto vertraut sind und letzteres, dass die Befragten den Wortlaut vorgeschlagener Funktionen verstehen können.[3]

An der Befragung beteiligten sich 227 Teilnehmer, hauptsächlich Studenten der TU-Dresden aus dem Grundstudium. Die Teilnehmer waren überwiegend männlich, jung, technisch interessiert und hatten einen IT-Hintergrund, wie aus Tabelle 6.1 hervorgeht. Wie bereits erwartet nutzten die Befragten sehr intensiv MP3s.

6.1.3 Ergebnisse

Die Auswertung des Fragebogens (ausführliche Ergebnisse finden sich im Anhang B.1.3) ergab zum Teil erwartete Erkenntnisse, aber auch Überraschungen. Wie erwartet gehört Musik hören für viele zum Autofahren dazu. Überraschend dagegen war die starke Verbreitung der MP3-Nutzung im Auto, ein Drittel der Befragten nutze dies bereits. Daraus lässt sich schlussfolgern, dass die Musikauswahl im Auto schon heute wichtig ist.
Am Computer benutzen die Mehrheit der Befragten die Software Winamp, um auf ihre MP3s zuzugreifen. Das spricht dafür, dass die dort dominierende Tracklist-Ansicht die Nutzer möglicherweise prägt. Diese Einschätzung diente als Bestätigung für die bereits zu diesem Zeitpunkt angedachte Form der Playlisten-Funktionen im WOZ-Test (siehe Abschnitt 6.2). Mehrheitlich wurden Playlisten benutzt, wenn auch meist wenige.

Die Vermutung, dass heute im Durchschnitt mit sehr großen MP3-Sammlungen umge-

[2] Also die Freude, eine Funktion zu benutzten (joy-of-use).
[3] Diese Voraussetzungen galten auch für alle anderen Tests im Rahmen dieser Arbeit. Bei den Sprachtests waren die muttersprachlichen Kenntnisse dann noch wichtiger, um sicherzustellen, dass der Dialog nicht aufgrund fehlender Sprachkenntnisse scheiterte.

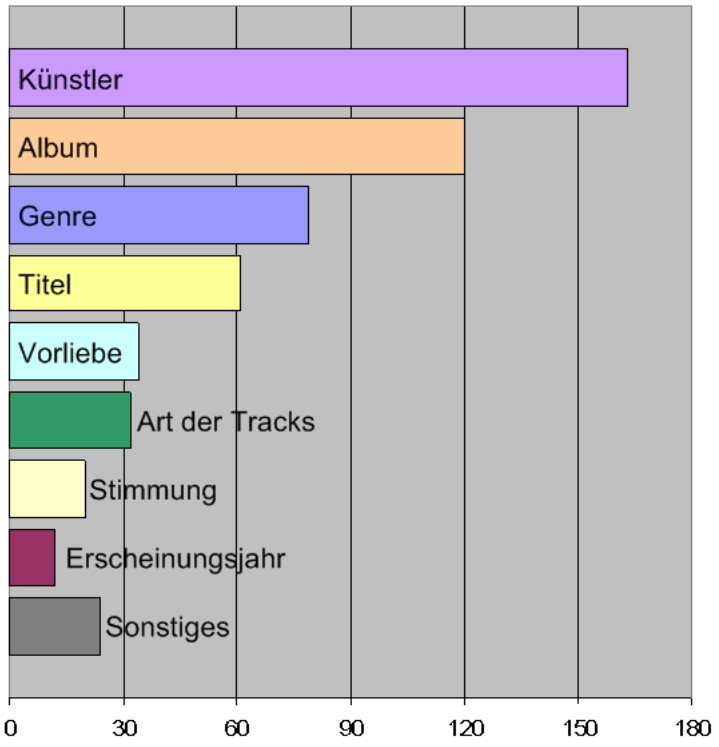

Abbildung 6.1: Kriterien Sortierung MP3-Sammlung (Fragebogen - Multiple Choice)

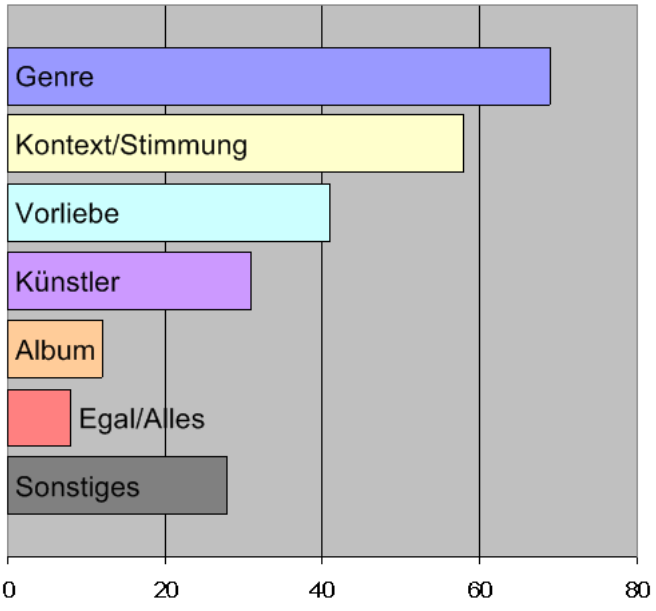

Abbildung 6.2: Kriterien Zusammenstellung Playlisten (Fragebogen - offene Frage)

gangen werden muss, bestätigte sich bei der Frage nach der Größe der eigenen Sammlung. Nach der Auswertung der Ergebnisse, welche später durch die Befragung vor dem WOZ-Test bestätigt wurden, kann beispielsweise davon ausgegangen werden, dass eine durchschnittliche Sammlung mehr als 1000 Titel und mehr als 10 Gigabyte[4] Musik beinhaltet.

Eine Sammlung wird meist klassisch nach Künstler und Album sortiert, wie Abbildung 6.1 zeigt. Das entspricht im Wesentlichen einer sehr klassischen, expliziten Sortierung. Das Merkmal Titel ist ebenfalls noch gebräuchlich.
Wesentlich seltener wird nach unscharfen Kriterien wie Genre, Vorliebe, Art der Tracks oder Stimmung sortiert. Trotz des relativ hohen Wertes für Genre kann davon ausgegangen werden, dass in den meisten Sammlungen eine sehr explizite und eher klassische Ordnung der MP3s vorherrscht. Interessant ist die Tatsache, dass in Zeiten von legalen und illegalen MP3-Downloads das Konzept des Albums bei der hier untersuchten jungen Stichprobe doch noch eine so hohe Bedeutung hat.

Der Fragebogen machte deutlich, dass nicht immer explizit Musik betrachtet wird. Bei der Frage nach den Kriterien für die Zusammenstellung von Playlisten ergab sich bei den Teilnehmern ein anderes Bild (Abbildung 6.2). Hier dominierten unscharfe Kriterien wie Genre, Kontext und Stimmung oder Vorliebe das Bild. Somit ergab sich ein zweiter Ansatz, um Musik zu finden bzw. zu hören.
Fraglich bleibt, ob diese Ergebnisse wirklich direkt auf die Musikauswahl anwendbar sind. Sie gaben aber zumindest Hinweise, dass die in Kapitel 3 unterschiedenen Nutzungsparadigmen für Musikauswahl durchaus existieren können und somit für beide eine Möglichkeit gefunden werden muss, auf Musik zuzugreifen.

Zur Eingrenzung der Funktionsumfangs eines sprachbedienten MP3-Players dienten die offenen und geschlossenen Fragen zu Funktionen. Überraschenderweise ergab sich auf die Frage nach der „liebsten Funktion“, die per Sprache gesteuert werden soll, eine sehr eindeutige Antwort. So wurden mit weitem Abstand drei Funktionen – mit der exakt gleichen Anzahl von Stimmen – genannt, die sich die Nutzer wünschen würden:

- Titel/Albumwahl
- Steuerkommandos (Play, Stop, Pause, Skip)
- Lautstärke-/EQ-Regelung

Während die ersten zwei Funktionalitäten zu erwarten waren, überraschte die Nennung von Lautstärke-/EQ-Regelung. Da die Lautstärke-Regelung per Sprache viele neue grundsätzliche Fragen aufwirft[5], wurde sich im Rahmen dieser Arbeit auf die beiden ersten Funktionalitäten als Schwerpunkte konzentriert.
Die Frage nach einer „möglichst coolen Funktion“ brachte leider keine neuen Erkenntnisse über allgemein gewünschte Funktionen, aber viele zum Teil individuelle, kreative Ideen

[4] Die meisten Befragten füllten entweder Textfeld Anzahl oder GB aus. Natürlich passen in 10 GB meist mehr als 1000 Titel. Die beide Arten von Werten wurden nicht in der Auswertung vermengt, um diese Tatsache auszudrücken.

[5] Wie z.B. könnte eine Lautstärke-Regelung per Sprache aussehen? Ist damit vielleicht nur eine „Mute“-Funktion gemeint? Ist nicht die Bedienung eines Drehschalters viel natürlicher?

(siehe Anhang B.1.1).

Die Auswertung der Bewertung zur Nützlichkeit verschiedener vorgegebener Funktionen bestätigte sich die Präferenz der bereits genannten Steuerkommandos und der direkten Musikwahl über Tags, neben Titel und Album auch Interpret und Genre (siehe Anhang B.1.3). Zusätzlich konnten auch die Steuerung von Spielmodi (Zufallswiedergabe, Wiederholung) und das Laden von Playlisten als wichtige Funktion ausgemacht werden.
Darüber hinaus wurden die drei Funktionen: „Alles wiederholen“, „Etwas zufälliges spielen“, „Informationen über den aktuellen Titel anzeigen“, in der weiteren Konzeption berücksichtigt. Zwar erreichten sie nicht ganz so gute Bewertungen, ließen sich aber einfach in ein Konzept integrieren.

6.2 WOZ

Der Wizard-of-Oz-Test war dazu vorgesehen, Fragestellungen zum Thema „sprachgesteuerter MP3-Player“ zu testen, welche auf das Erleben von Sachverhalten und Bewerten derer abzielten. Dazu mussten geeignete Repräsentationsformen gefunden und mit Aufgaben und Fragen in eine auswertbare Form gebracht werden.

6.2.1 Voraussetzungen & Anforderungen

Um das Dialogkonzept für die Musikauswahl (der Aufbau der restlichen Funktionen wie Steuerkommandos war bereits bekannt) zu entwickeln, musste zunächst die Datenbasis für diese Auswahl analysiert werden. Wie bereits in Abschnitt 3.4 festgehalten, wurden dafür ID3-Tags [Nil00] berücksichtigt. Bei dieser Analyse wurden Tags festgehalten, die prinzipiell für eine Musikauswahl interessant sein könnten. Danach wurde nach der Wahrscheinlichkeit der Benutzung dieser Tags eine Auswahl getroffen. Dabei wurden folgende Tags als nützlich für die Musikauswahl identifiziert[6]:

- Title
- Album
- Artist
- Genre
- Year

Auf Basis von eigenen Überlegungen zur Verknüpfung von Eigenschaften mit den entsprechenden Tags („Genre ist eine unscharfe Musikbeschreibung“) wurden Überlegungen zur Systemreaktion nach Benutzung der entsprechenden Tags („Nach ’Genre Pop’ sollte abgespielt werden“) angestellt. Ein Teil der Ergebnisse dieser Überlegungen finden sich in Tabelle 6.2. Weitere Überlegungen zu expliziten Modi für Play und Browse sowie zur Kombination der Tags finden sich im Anhang B.1.2. In der Abbildung steht „Konzept“

[6] Der darüber hinaus interessante Composer-Tag wurde wegen der hauptsächlichen Fokussierung auf populäre Musik zunächst ausgenommen. Wie aber bereits in Abschnitt 3.2 diskutiert, kann für die Musikauswahl ein Komponist als ein weiterer Artist betrachtet werden.

	Nutzeräußerung		**Systemreaktion**		
	Konzept	Argument	Aktion	Liste	Sonderfälle
Titel	X		Zeige	Titel (n)	
	X	X	Spiele	Titel (1)	
	X	X	Spiele	Titel (n)	gl. Titelname
Interpret	X		Zeige	Interpreten	
	X	X	Zeige	Alben	
Album	X		Zeige	Alben	
	X	X	Spiele	Titel d. Albums	
	X	X	Zeige	Interpreten	gl. Albumname
Jahr	X		Zeige	Jahre	
	X	X	Zeige	Interpreten	
Genre	X		Zeige	Genre	
	X	X	Spiele	Zufäll. Titel Genre	

Tabelle 6.2: Entwurf zum System-Verhalten nach Benutzung verschiedener Tags (WOZ)

jeweils für ein allein benutztes Konzept (wie „Titel“, „Album“ etc.) und „Argument“ für die Kombination eines Konzeptes mit einem Argument („Titel 99 Luftballons“). Die Spalte „Aktion“ gibt an, ob angezeigt oder abgespielt wird und die Spalte „Liste“ beschreibt. Da bei Titel vom Wort her unklar ist, ob es sich um einen oder mehrere handelt, ist dies dort in Klammern mit angegeben.
Im Wizard-of-Oz-Test sollte nun herausgefunden werden, inwieweit diese subjektive Basis (im Weiteren „Play/Browse-Hierarchie“ genannt) für ein Dialogkonzept mit den Erwartungen der Nutzer übereinstimmt. Insbesondere ging es um die grundsätzliche Frage, ob eher einer hierarchie-geleiteten Auswahl oder einer direkten, aber möglicherweise ungenauen Auswahl der Vorzug gegeben wird. In diesem Umfeld sollten auch die Auflösung von Mehrdeutigkeiten untersucht werden und ob diese automatisch vom System oder nutzerbestimmt behandelt werden sollten.

Dies führte unter anderem zu der Überlegung, den WOZ-Test als möglichst offenen Test durchzuführen. Das heißt, nicht in jedem Fall sollte mit einem festem Dialogkonzept an die zu lösenden Fragen herangegangen werden. Über den Test sollte herausgefunden werden, welches Konzept der Musikauswahl von einem solchen System intuitiv von Testpersonen erwartet wird. Dies sollte auch Klarheit darüber bringen, ob, wie in Kapitel 3 vermutet, wirklich verschiedene Nutzergruppen zu berücksichtigen sind, wie diese mit den Systemen umgehen und ob auf eine Dateiansicht verzichtet werden kann. Weiterhin hatten sich bereits konkrete Designfragen ergeben, die ebenfalls geklärt werden sollten. Konkret sollte der Test also:

- Designfragen beantworten,
- Das Grobkonzept des Auswahldialogs testen,
- Herausfinden, was Nutzer „einfach so sagen“.

Für die Umsetzung stand das schon in Abschnitt 5.4.3 diskutierte WOZ-Tool von Stefan

Petrik [Pet04] zur Verfügung, dass offene Tests durch seine freie, hierarchische Struktur ermöglicht. Durch die schnelle und auch informell mögliche Spezifikation des Dialogs, ebenso wie durch die einfache Integrierbarkeit von Musikstücken in den Dialog[7] konnte der Entwicklungsaufwand erheblich reduziert werden.

6.2.2 Umsetzung

Vor der Umsetzung wurde untersucht, ob es bereits Erfahrungen mit dem Aufbau von Wizard-of-Oz-Tests in diesem oder ähnlichem Umfeld gab. Insbesondere die Arbeit von Kruijff-Korbayová et al. [KKBG+05] erwies sich dabei als durchaus vergleichbar. Dort wurde ebenfalls sprachgesteuerte Musikauswahl im Auto mithilfe eines WOZ-Tests untersucht. Jedoch hatte diese Untersuchung gänzlich andere Ziele, es wurden vor allem die Wizards getestet und wie sie die Erledigung ihrer Aufgabe planten. Nach der freien Anfrage eines Nutzers mussten hierbei die Wizards die eingebundene Datenbank von freedb [fre06] befragen und konnten dann für die Antwort aus verschiedenen, bereits vorgegebenen Screens auswählen. Was auswählbar war, entschied eine bereits vorhandene Struktur. Eine solche Struktur sollte in den WOZ-Tests im Rahmen dieser Arbeit aber erst gefunden werden.[8]. Dafür wäre aber ein solches freies Vorgehen der Nutzer nicht möglich bzw. unverhältnismäßig aufwendig gewesen. Deswegen fiel die Entscheidung, begrenzte Aufgaben zu definieren, die nur eine überschaubare Anzahl möglicher Nutzerhandlungen zulassen, aber trotzdem dem Nutzer über das „Wie“ des Vorgehens freie Hand lassen. So konnte effektiv die Belastung für den Wizard gesenkt werden und die für die Illusion wichtige Reaktionszeit zwischen Äußerung und Aktion auf ein annehmbares Maß gebracht werden.

Bei der Erstellung des Testkonzeptes und der Aufgaben wurde allerdings schnell klar, dass manche Themenbereiche im Rahmen eines solchen Tests nicht angemessen überprüft werden können. So wäre bei mancher Aufgabe (insbesondere bei der Play/Browse-Hierarchie für die Musikauswahl) durch die Formulierung der Frage meist schon der Weg bzw. das konkrete Ziel ziemlich genau vorgegeben gewesen. Darüber hinaus wäre die Versuchsperson nach dem Versuch kaum in der Lage gewesen, den Dialogverlauf des Tests zu rekonstruieren, um eine Entscheidung für eine Art des Vorgehen bei der Musikauswahl zu treffen.
So wurde für diese Bereiche lieber auf eine Powerpoint-Präsentation der verschiedenen möglichen Systemzustände zurückgegriffen. Und nach der Präsentation aller Möglichkeiten der Nutzer jeweils zu einer Auswahl aufgefordert (im Weiteren „Powerpoint-Befragung“ genannt). Mit Hilfe dieser Methode wurde vor allem eine Vorstellung einer intuitiven Play/Browse-Hierarchie entwickelt. Abbildung 6.3 zeigt den grafischen Aufbau, so wurden als Zuordnungsmöglichkeit die Fragen und Antwortmöglichkeiten immer eingeblendet, bevor der jeweilige Ablauf gezeigt wurde (die Powerpoint-Präsentation findet sich auf der beiliegenden CD-ROM - siehe Abschnitt A).

Durch die Auslagerung dieser wichtigen Strukturentscheidungen in die Powerpoint-Befragung konnte im Test ein eher konventionelles Systemdesign angewendet und auch eine mögliche Verwendung von Playlisten getestet werden (genauere Informationen dazu finden sich im Abschnitt 7.1). Weitere Aufgaben im Test widmeten sich vor allem der typischen Wortwahl verschiedener Funktionen (wie beispielsweise Shuffle), um für die Grammatik des

[7] Wenn auch mit einer sehr geringen Abtastrate von 11 khz. Über die Qualitätseinbußen durch diese Konvertierung beschwerte sich allerdings – erschreckenderweise – keine einzige Versuchsperson.
[8] Kruijff-Korbayová et al. formulieren das Ziel einer intuitiven Struktur nur im Ausblick.

Abbildung 6.3: Frage und Simulation in der Powerpoint-Befragung (WOZ)

Prototypen eine ausreichende Zahl von Synonymen zu sammeln. Um die jegliche Beeinflussung des Nutzers zu vermeiden, wurden weder Hilfe noch aussagekräftige Fehlermeldungen in das System integriert. Ebenso wurden die Informationen für die Aufgaben, wenn möglich, auf Karteikarten präsentiert, um so verbale Beeinflussung zu minimieren.
Weiterhin wurde der Test von einigen Befragungen flankiert. So wurden in einer Vorbefragung, die im Wesentlichen dem in Abschnitt 6.1 beschriebenen Fragebogen entsprach, die Ergebnisse des Fragebogens noch einmal validiert. In einer Nachbefragung (mit einem mündlichen und schriftlichen Teil) der generelle Eindruck des Systems erfasst, um eine Vorstellung der Zufriedenheit mit dem System, über stark gewünschte Funktionen oder weitere Entwicklungspotentiale zu erhalten. Dabei wurde der SUS-Bogen (siehe Abschnitt 5.4.1) für die Ermittlung eines einfachen Usability-Maßes benutzt. Anschließend wurde die bereits angesprochene Powerpoint-Befragung durchgeführt.

Für die Implementierung des WOZ-Tests wurden Use Cases textuell spezifiziert (in Form von Dialogen), was dann als Basis für die Überführung in Prompts im WOZ-Tool diente. Diese Spezifikation, wie auch alle anderen Materialien, Anweisungen und Befragungsbögen, die während dieses Tests eingesetzt wurden, finden sich im Anhang B.1.2.

Zur Ablenkung wurde, wie auch bei Kruijff-Korbayová et al. [KKBG+05], der Lane-Change-Test (siehe Abschnitt 5.2) eingesetzt. Jedoch nur, um die Versuchspersonen in eine realitätsgetreue Umgebung zu versetzen. Die Fahrdaten wurden zwar aufgezeichnet, aber nicht ausgewertet, weil keine Vergleichswerte (Bedienung ohne Sprache) zur Verfügung standen und eine Erhebung dieser Daten den zeitlichen Rahmen gesprengt hätte.

Der Testaufbau schließlich wurde in den Räumlichkeiten von Harman/Becker eingerichtet und ist in Abbildung 6.4 dargestellt.
Dabei wurde zusätzlich zu dem üblichen, bereits im Abschnitt 5.4.3 vorgestellten Testaufbau, auf Nutzerseite eine Fahrsimulation mit Lenkrad und Monitor als Ablenkung hinzugefügt, ebenso wie ein 7-Zoll-Monitor (übliche Größe der Anzeigen im Auto) für grafische Rückmeldungen des Systems. Ein Versuchsleiter führte die Versuche durch und protokollierte schriftlich und per Videoaufzeichnung mit, um im Nachhinein wortgenau den Dialog nachvollziehen zu können.
Dieser Testaufbau wurde mittels zweier Pilottests getestet, kleinere auftretende Mängel konnten so erkannt und beseitigt werden.

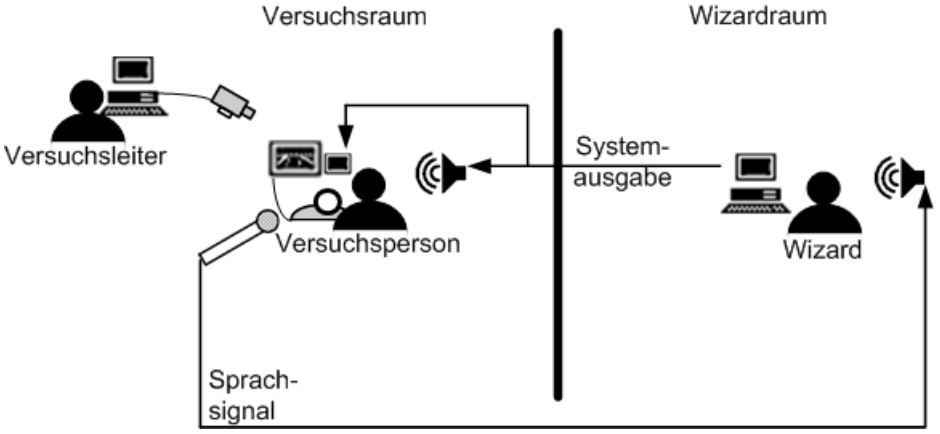

Abbildung 6.4: Testaufbau (WOZ)

Für den eigentlichen eineinhalbstündigen Test konnten 20 Testpersonen gewonnen werden, hauptsächlich Studenten der Ulmer Hochschulen aus dem Hauptstudium. Die Stichprobe ähnelten in der Zusammensetzung sehr jener der Befragung, mit den kleinen Abweichungen, dass die Teilnehmer im Schnitt etwas älter und zu einem geringeren Teil aus dem IT-Bereich kamen, wie aus Tabelle 6.3 hervorgeht. Die Mehrzahl davon hatte bisher noch keine Erfahrungen mit Sprachdialogsystemen.

WOZ-Test	Teiln.	Alter	Geschlecht	Hintergrund	tech. Interesse
	20	85% 21-30	85% ♂	55% IT	95% hoch/sehr hoch

Tabelle 6.3: Zusammensetzung Stichprobe WOZ-Test

6.2.3 Ergebnisse

Die Trennung des Tests in Powerpoint-Befragung und WOZ-Test bewährte sich, beide Teile brachten die von ihnen erwarteten Erkenntnisse. Die Powerpoint-Befragung brachte Klarheit über die Struktur der Play/Browse-Hierarchie, während der Test wertvolle Informationen über Wortwahl, Fahrablenkung und der Integration von Playlisten bot.

Im Test wurde klar, dass ein explizites Play-Kommando vor dem Abspielen, wie es aufgrund der Integration der Playlisten-Funktion nötig wurde, von den Nutzer eindeutig abgelehnt wurde. Vielmehr wurde ein direktes Abspielen von Titeln eingefordert.
Wie sich dies in eine Play/Browse-Hierarchie einfügt, konnte durch die Powerpoint-Befragung geklärt werden. Dabei wurden Überlegungen, einen jeweils durch Schlüsselwort aktivierbaren Play- und Browse-Modus zu schaffen (siehe Abschnitt 3.3.4) endgültig verworfen, da dieser Unterschied den Versuchspersonen im Test nicht auffiel und auch nicht vermittelbar war.
Somit ergab sich eine Struktur für die Play/Browse Hierarchie, welche unabhängig von solchen Schüsselwörtern ist und im Anhang B.1.3 dargestellt ist. Die Veränderung gegenüber den Vorstellungen vor dem WOZ-Test sind darin hervorgehoben. Diese Struktur konnte aus einer Reihe von Erkenntnissen abgeleitet werden.
Die Nutzer bevorzugten bei der Wahl von Titel oder Album jeweils das direkte Abspielen.

Weiterhin konnte aus Antworten auf eine weitere Frage geschlossen werden, dass auch nach der Titelauswahl die gesamte Albenliste in die aktuelle Wiedergabeliste geladen werden soll.
Bei doppeldeutigen Album- oder Titelnamen wünschten sich die Nutzer mehrheitlich eine Disambiguierung[9] über die Künstler-Auswahl.
Keine klare Tendenz zeigte sich jedoch bei den Fragen zu Künstler und Genre. Vielmehr war eine starke Polarisierung zwischen zwei Gruppen zu beobachten. Die Einen wollten direktes Abspielen eines Titels des gewählten Künstlers oder Genres, die Anderen ein Anzeigen von Listen der Alben oder Künstler. Dies entsprach der Erwartung aus dem Vorwissen zu den zwei Arten von Musikauswahl, wie bereits aus der Auswertung des Fragebogens vermutet.[10]
Da sich bei der Auswahl eines Künstlers lediglich rund ein Drittel der Testpersonen für direktes Abspielen entschieden hatte, wurde hier das Anzeigen einer Liste von Alben als Standardverhalten gewählt.
Bei Genre ergab sich zwischen Befürwortern und Gegnern direktem Abspielens ein Patt, weswegen im Anschluss eine kleine hausinterne Umfrage unter 15 Mitarbeitern von Harman/Becker durchgeführt wurde, in der sich nur eine Minderheit der Befragten für ein solches direktes Abspielen begeistern konnten. Dagegen wollten mehr als die Hälfte lieber aus einer Liste der Künstler des Genres auswählen, weswegen dieses Verhalten im Weiteren bevorzugt wurde.
Die Auswahl über die Jahreszahl löste bei einigen Teilnehmer Verwundern aus, einige machten sogar eindeutig klar, dass sie diese Möglichkeit nie benutzen würden. Da üblicherweise Jahresangaben bei ID3-Tags sehr spärlich oder ungenau vorhanden sind, wurde daraufhin von einer weiteren Berücksichtigung abgesehen.

Playlisten wurden direkt über den Namen aufgerufen, so dass die Behandlung von Playlisten analog zu Tags sinnvoll erschien. Darüber hinaus wurden jegliche Funktionen zur Playlisten-Erstellung verworfen, da kein wirkliches Interesse daran gezeigt wurde und außerdem die bereits oben erwähnten Probleme mit dem expliziten Play-Kommando bestanden.

Bei der Wortwahl im Versuch zeigte sich die erwartete Tendenz der Benutzung von sowohl englischen als auch deutschen Begriffen (beispielsweise sowohl „Play“ als auch „Abspielen“). Es wurde empfohlen, Synonyme beider Sprachen für Kommandos zu berücksichtigen.
Aus der Befragung zur Fahrablenkung konnte geschlossen werden, dass vor allem die Blicke zum Display zur Ablenkung beigetragen hatten. Mehr als zwei Drittel der Versuchspersonen fühlten sich davon vom Fahren abgelenkt.[11] Deswegen wurden Möglichkeiten gesucht, das System möglichst blind bedienbar zu gestalten, um die Notwendigkeit solcher Blicke zu reduzieren.

Bei der abschließenden Bewertung über eine Schulnote (1-6) und den SUS-Bogen (0-100)

[9] Die Disambiguierung (auch Desambiguierung) ist der Vorgang und das Ergebnis der Auflösung von Doppeldeutigkeiten sprachlicher Ausdrücke durch den sprachlichen oder außersprachlichen Kontext [BN05].

[10] Die Ergebnisse des Fragebogens konnten im Wesentlichen durch die durchgeführte Vorbefragung bestätigt werden.

[11] Ebenso fühlten sich die Versuchspersonen durch die Karteikarten mit Musikinformationen abgelenkt, da sie zum Ablesen den Blick von der Straße wenden mussten. Dies führte dazu, in der Evaluation des Prototypen „Dorothy“ einen anderen Weg der Aufgabenpräsentation zu wählen, siehe Abschnitt 8.2.

ergaben sich mit 2,05 und 82,5 gute Werte, was sicher auch auf die „perfekte Spracherkennung“ des Wizards zurückzuführen ist, wurde diese doch von einer Mehrheit als die Stärke des Systems bezeichnet.

Die vollständigen Ergebnisse des WOZ-Tests finden sich in Anhang B.1.2.

6.3 Schlussfolgerungen für die weitere Entwicklung

Aus den bisher dargestellten Ergebnissen ließen sich viele Einzelerkenntnisse ziehen, die auch bei der Erstellung des Prototypen und der anschließenden Evaluation beachtet wurden. Doch lohnt es sich, noch einmal die wesentlichen Erkenntnisse für die weitere Entwicklung zusammenhängend darzustellen.

Ausgehend von der, im Rahmen der Befragung bestätigten, immer stärkeren MP3-Nutzung im Auto und der Größe durchschnittlicher privater Musiksammlungen, kann von einer Notwendigkeit einer einfachen und intuitiven Bedienung für diese Datenmengen im Auto ausgegangen werden. Doch erscheint dies schwierig angesichts verschiedener Nutzungsszenarien, wie sie im Rahmen der Untersuchungen auftraten. So ist die in Kapitel 3 diskutierte Unterscheidung in Stöberer und Bibliothekare wohl zutreffend. Ein Konsens für die Bedienung solcher Systeme zeichnete sich jedoch nicht ab.[12] Im WOZ-Test wurde zudem deutlich, dass eine nutzergesteuerte Auswahl dieser Modi (Play- und Browse-Mode) nicht bevorzugt wurde.
Vielmehr musste eine Strategie gefunden werden, beide Nutzungsparadigmen in einem System weitestgehend zu vereinen. Dieses System musste möglichst schnell und unkompliziert funktionieren, aber auch genaue Auswahl ermöglichen, ohne dabei zwischen den Modi umschalten zu müssen. Da dies nur ein Kompromiss sein konnte, wurde dieses System nach der Fertigstellung wieder einem Nutzertest unterzogen werden.

Eine zweite wesentliche Schlussfolgerung war die Entscheidung, das System möglichst blind bedienbar zu machen. Dies resultierte aus der starken Ablenkung, die die Blicke auf das Display im ersten Test bei den Testpersonen ausgelöst hatten. Die Reduzierung dieser Display-Gebundenheit hatte angesichts des starken Einflusses von Listen innerhalb des Systems seine Grenzen. Hier boten die in Kapitel 4 diskutierten non-verbalen Interaktionselemente einen gangbaren Ausweg, wobei natürlich auch die Auswirkungen auf das gesamte Systemverhalten sorgfältig analysiert werden mussten.

Schließlich musste das System auch einmal unter realistischen Bedingungen ausprobiert werden, da viele Testpersonen von der vermeintlich perfekten Spracherkennung so geblendet waren, dass sie vermutlich eventuelle Probleme im restlichen System großzügig übersahen. Das sprach für einen Test mit einem funktionstüchtigen Prototypen, in dem auch getestet werden konnte, inwieweit der geplante Wortschatz unter realistischen Bedingungen funktionieren würde.

[12] In den Nachbefragungen des WOZ-Tests äußerten sich sogar meist radikalere und unvereinbare Forderungen („Wenn das nicht immer gleich abspielt, ist es sinnlos“, „Ich möchte immer volle Kontrolle über meine Auswahl haben“ usw.).

7 Prototypische Umsetzung

In diesem Kapitel soll zunächst beschrieben werden, wie sich die Vorstellung des Systems vom Anfang der Arbeit bis zur Definition der Anforderungen für den Prototypen entwickelt hatte. Dabei wird jeweils kurz die Grundidee beschrieben und ein kurzer Einblick in mögliche Dialogabläufe gegeben.
Danach werden die Anforderungen für den Prototypen, dessen technische Rahmenbedingungen und schließlich die Umsetzung erläutert.

7.1 Systemhistorie

Bei der Entwicklung der Systemidee im Laufe dieser Arbeit wurden, ausgehend von prinzipiellen Vorüberlegungen zu Vorteilen von Sprache und einigen Ideen zu einem prinzipiellen Aufbau der Musikauswahl, die einzelnen Zwischenentwürfe nach und nach immer mehr an technischen Gegebenheiten und den Ergebnissen der Vorbefragung und des WOZ-Tests orientiert und angepasst.
Um die Situation vor der Erstellung des endgültigen Prototypen zu verstehen, ist es deswegen notwendig neben dem Systementwurf, der beim WOZ-Test zum Einsatz kam, auch die anderen Systementwürfe aufzuzeigen. Dabei werden auch die zeitlich vor den in Kapitel 6 diskutierten Untersuchungen liegenden Systemvorstellungen an dieser Stelle vorgestellt, damit ein ganzheitlicher Blick auf die Entwicklung der Systemidee möglich wird.

Grundlage für die ersten Ideen bildeten Vorüberlegungen zu Sprachdialogsystemen im Allgemeinen und deren Einsatz im Auto im Speziellen, welche in Kapitel 2 diskutiert wurden. Insbesondere stand dabei der Gedanke der Schnelligkeit im Mittelpunkt, die Auswahl per Sprache sollte also schneller als mit der klassischen haptisch-grafischen Schnittstelle möglich sein. Durch dieses Vorgehen sollte also ein Finden ohne Anstrengung möglich sein. In erster Konsequenz bedeutete dies, dass möglichst jede Hierarchie und jedes „Durchhangeln“ durch Menüs verhindert und so eine Art freie Suche in großen Datenmengen ermöglicht werden sollte. Da bot es sich an, Ideen beim Marktführer in Sachen Suchen zu sammeln und erstmal fernab technischer Randbedingungen Ideen zu entwickeln.

Die Idee: Google Ansatz

Google [Goo06] ist seit einigen Jahren zum Inbegriff für Suche geworden. Beginnend mit der Websuche bietet das Unternehmen heute viele verschiedene Dienste an, die sich jedoch größtenteils ebenfalls mit dem einfache Finden von Informationen in Spezialgebieten beschäftigen. Dabei ist der Ansatz immer gleich, mit einem minimalen Aufwand für den Nutzer soll schnell und effektiv in riesigen Datenbeständen gesucht werden.
Ähnlich diesem Prinzip sollte nun ein möglichst kontextfreier und schneller Weg gefunden werden, auf Musik zuzugreifen, weswegen der Name „Google Ansatz“ gewählt wurde. Ziel war es, in wenigen Schritten, ohne große Korrekturen in den meisten Fällen die richtige Musik auszuwählen, und das möglichst den eigenen Vorlieben entsprechend.

Dazu mussten zuerst Ideen gesammelt werden, über welche Daten ein solchen Zugriff ermöglichen werden könnte. In Abschnitt 3.2 wurden bereits die dabei nutzbaren Metadaten diskutiert, neben den ID3-Tags boten sich hier neben unscharfen Kriterien wie Musikgeschwindigkeit, Stimmung oder Lyrics auch die Benutzung von Nutzerwertungen oder -charts an.

Durch die, wie auch immer geartete, Benutzung und Kombination dieser Daten wäre eine Filterung der Gesamtmenge der MP3s möglich. Mögliche Ergebnisse einer solchen Filterung wären „kein Treffer“, „genau ein Treffer“ oder „mehrere Treffer“. In den letzten beiden Fällen würde sich jeweils eine Liste (mit Einträgen ≥ 1) ergeben, die so genannte Tracklist. Auf einer solchen Liste könnte dann mit den üblichen Play-Controls (Play/Pause, FF, Rew, Next, Prev) navigiert werden.[1]
Dagegen ist die Möglichkeit der ergebnislosen Anfrage eher zu vermeiden, vielmehr sollte bei Unklarheiten lieber ein „unscharfes Matching“ durchgeführt werden. Dies würde bedeuten, bei ungenauen oder widersprüchlichen Anfragen zunächst eine interne Korrektur vorzunehmen und direkt nach der Auswahl Musik als direktes Feedback abzuspielen, um die (aufwendige) Korrektur der Anfrage möglichst oft zu umgehen (siehe Dialog 7.1). Prinzipiell war dabei die Regel angedacht: „Lieber ein schlechtes Ergebnis als gar keins“.

Dialog 7.1: „unscharfes Matching“ (Google Ansatz)

usr: Ich möchte das 'American Album' von Van Morrison hören.
sys: Spiele 'An American Album' von Jim Morrison.

Wenn dann eine Korrektur doch nötig werden sollte, sollte eine History mit bisherigen Anfragen bereitstehen, um eine kontextsensitive Bearbeitung der Anfrage möglich zu machen. So könnte nach der Auswahl einer Stimmung „harmonisch“ der Nutzer den Interpret „Elvis Presley“ auswählen, und das System würde nur harmonische Stücke von Elvis Presley als Treffer anbieten.
Ein solches Vorgehen wäre aber nur sinnvoll, wenn die Nutzer nicht das Gefühl hätten, sich über eine Struktur oder eine Zurück-Funktionalität im System zu bewegen. Insofern sollten sie motiviert werden, jeweils neue Anfragen an das System zu stellen, wenn ihnen eine Systemreaktion einmal nicht gefällt. Dies sollte durch Allverfügbarkeit der möglichen Kommandos geschehen.

[1] Als weitere Funktionen sind auch Funktionen zum Speichern der Liste als Playlist oder Informationen zum Titel denkbar.

Die Einbindung der Vorlieben des Nutzers sollten aber nicht nur durch schon in Tags gespeicherte Daten erfolgen, vielmehr war angedacht, auch während des Hörens eine Bewertung (analog zu den in Abschnitt 3.3.1 beschriebenen Bewertungsmethoden) zu ermöglichen, die Einfluss auf die Auswahl bei Doppeldeutigkeiten haben sollte.

Diese Systemvorstellung, die in Abbildung 7.1 visualisiert ist, war natürlich weit von jeder Machbarkeit entfernt, da sie zum Beispiel keine genauere Betrachtung der eigentlichen Auswahl vornahm („Black Box"), sondern nur wünschenswerte Ergebnisse betrachtete. Die Vorstellung diente jedoch in allen weiteren Schritten als Idealvorstellung, an der realistische Ansätze gemessen werden sollten.

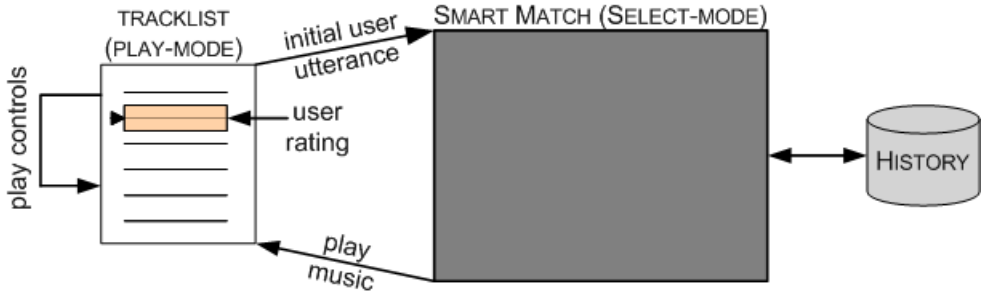

Abbildung 7.1: Grundaufbau Musikauswahl mit Black-Box „Smart Match" (Google Ansatz)

Erster Realitätscheck: Query-MP3 Ansatz

Ohne schon direkt ins technische Detail zu gehen, wurden dann die Bestandteile des Google-Ansatzes realistisch bewertet und Schlussfolgerungen für das Systemdesign formuliert.

Zunächst wurden die Schlussfolgerungen über die Metadatenverfügbarkeit aus Abschnitt 3.2 zum Anlass genommen, sich auf ID3-Tags zu beschränken. Diese massive Verkleinerung der Datenbasis führte zu neuen Betrachtungen über die mögliche Kombinierbarkeit von Tags, die geringe Anzahl erlaubte erstmals eine explizite Aufstellung über mögliche Kombinationsmöglichkeiten (wie bereits in Abschnitt 6.2.1 erläutert). Bei der Betrachtung dieser wurde klar, dass eine Verfeinerung einer Anfrage an das System bei diesen meist strikt hierarchischen Tags kaum mehr sinnvoll einsetzbar war.[2]
Auch die unscharfe, automatische Auswahl stellte sich bei näherer Betrachtung als kaum direkt umsetzbar dar, die Arbeit von Forlines et al. [FSNR+05] zeigte, was für ein Aufwand hätte getrieben werden müssen, um diese Ideen verwirklichen zu können (wie bereits in Abschnitt 3.4 diskutiert).

Doch wie konnte ohne diese ganzen eleganten Methoden trotzdem das Ziel eines effektiven, schnellen und einfachen Musikzugriffs erhalten werden?
Um jeden zusätzlichen Dialogzustand zu vermeiden, wurde zunächst die Benutzung der

[2] Zwar wurde in der WOZ-Powerpoint-Befragung noch eine Frage dazu gestellt, aber wurde dann von den Nutzern auch massiv der Nutzen einer solchen Funktion in Frage gestellt.

ID3-Tags in einer Äußerung weiterhin als Suchanfrage angesehen, nach der direkt als Feedback die Musik abspielen sollte. Ohne Verfeinerungsfunktionalität stellte sich aber die Frage, wie Nutzer einerseits mit „Interpret Elvis Presley“ direkt Musik von Elvis abspielen, aber anderseits ein bestimmtes Album von ihm auswählen könnte.
Die Lösung bot sich mit dem Ansatz von Wang et al. [WHHS05] an, welche die Verwendung von zwei Modi für den Zugriff auf die Musik vorschlugen:

- Play-Mode (direkte Auswahl, direktes Abspielen)
- Browse-Mode (Hierarchie-Zugriff, wie iPod)

Der Browse-Mode bezeichnet dabei die Auswahl über Listen, wie das im iPod bereits heute schon realisiert ist. Zu jeder Anfrage wird dabei eine Liste erstellt, in der der Nutzer einen Eintrag wählt und dann durch einzelne Hierarchiestufen bis zum Abspielen geführt wird. Im Play-Modus wird dagegen eine Anfrage direkt ausgewertet und direkt danach mit dem Abspielen des MP3-Files direkt begonnen. Interessant wurde dies nun, wenn die Unterscheidung der Modi nicht wie standardmäßig über zusätzliche Kommandowörter („Spiele“, „Zeige“) getroffen wurde, sondern Äußerungen ohne Schüsselwort verarbeitet werden mussten. Dazu musste ein Standardverhalten (die Play/Browse-Hierarchie) definiert werden, um auch ohne Kommandowort eine Aktion auszuführen. Dabei immer nur ein Verhalten zu bevorzugen, erschien nicht sinnvoll, so dass eine entsprechende Struktur nach eine subjektiven Einschätzung definiert wurde. Diese Systemvorstellungen sind in Abbildung 7.2 skizziert.

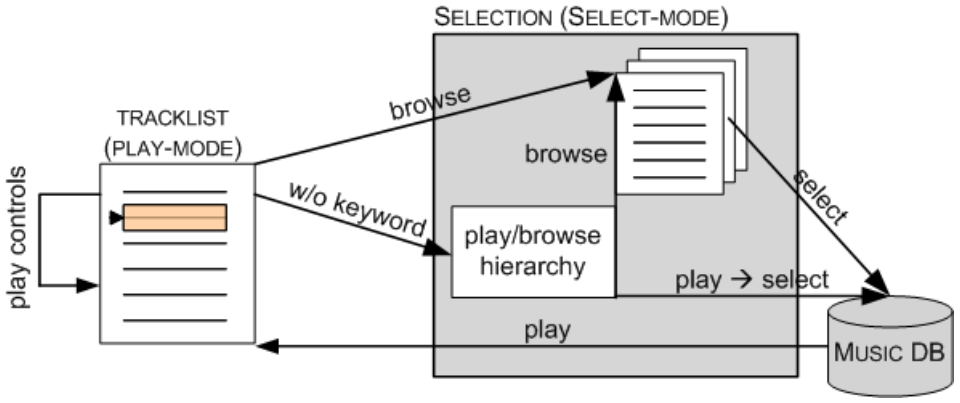

Abbildung 7.2: Grundaufbau Musikauswahl mit Play/Browse-Mode sowie Standardverhalten (Query-MP3 Ansatz)

Diese (Tag-)Struktur, wie aber auch alle anderen Annahmen über die Wünsche der Nutzer (Wollen sie denn überhaupt so ein schnelles, aber auch eher ungenaues Vorgehen?), waren jedoch bis zu diesem Zeitpunkt nicht hinterfragt wurden. Durch den Fragebogen und den WOZ-Test sollte dem nun abgeholfen werden, was zur Erstellung des WOZ-System-Konzeptes führte.

Nutzereinbindung: WOZ-System

Bei der Konzeption des Test-Systems für den WOZ-Versuch musste zunächst beachtet werden, dass die Fragen über die Struktur der Anwendung in die Powerpoint-Befragung

ausgelagert wurden (siehe Abschnitt 6.2.2). Da der Test vor der Powerpoint-Befragung stattfinden sollte, durfte durch die Struktur des Test-Systems keine Vorbeeinflussung entstehen.[3]
Dies führte zu einer einheitlichen Festlegung des Verhaltens in der Play/Browse-Hierarchie, so wurde immer nach der Auswahl eine Ergebnisliste präsentiert, die dann nur explizit auf Äußerung des Nutzers abgespielt werden sollte. Dieses Verhalten ermöglichte außerdem das Testen einer umfangreichen Playlisten-Funktionalität, mit welcher Lieder von der aktuellen Wiedergabeliste hinzugefügt und entfernt werden könnte sowie die Liste unter einem eigenen Namen abgespeichert werden konnte.
Dieses Konzept wurde erstellt, um an einem konkreten Beispiel herauszufinden, ob Playlisten-Funktionen sinnvoll im Auto umzusetzen sind. Dabei wurde bei der Entwicklung des Konzepts schnell klar, dass einige grundlegende Fragen nach dem allgemeinen Verständnis von Aufbau und Umgang mit Playlisten zu beantworten waren.
Denn zunächst war unklar, was ein Hinzufügen funktional bedeuten würde, zu welcher Liste dieses Hinzufügen stattfinden würde. Auch das Entfernen war schwierig umzusetzen, da dafür ein Eintrag zunächst ausgewählt werden musste. Ein gewissen Anhalt gab die, im Fragebogen festgestellte, breite Verwendung des Software-MP3-Players „Winamp“, dessen Prinzipien des Hinzufügen und Entfernen letztendlich als Vorbild der weiteren Entwicklung benutzt wurde.
Aus diesen Überlegungen heraus ergab sich die in Abbildung 7.3 visualisierte Struktur.

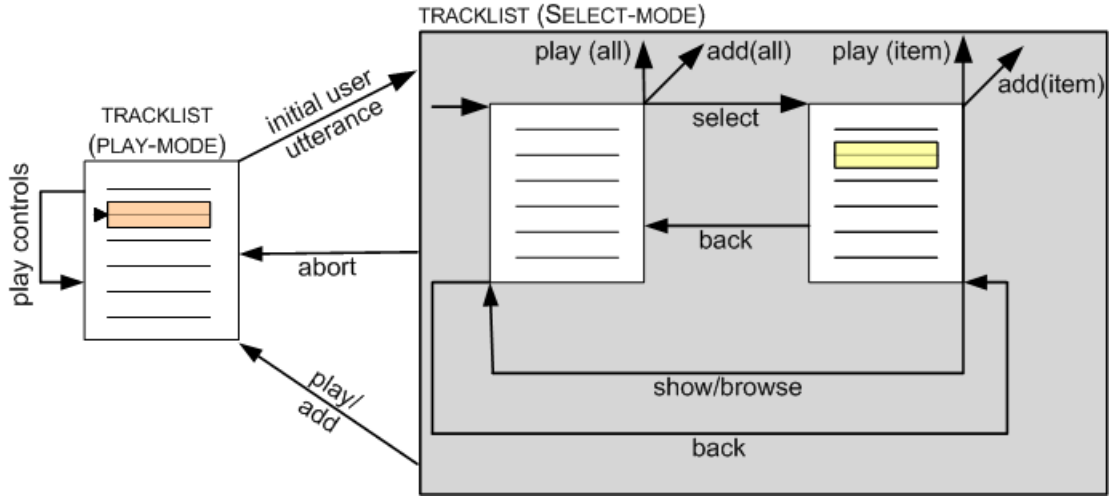

Abbildung 7.3: Veranschaulichung einer möglichen Dialog-Struktur (WOZ)

Ausgehend von einer initialen Nutzeräußerung würde dem Nutzer dabei zunächst eine Liste ohne Auswahl präsentiert. Diese Liste könnte nun abgespielt oder der bisherigen Liste angefügt werden, was jeweils die Rückkehr in den Abspielmodus bedeuten würde. Allerdings könnte alternativ auch ein Eintrag ausgewählt werden, um diesen einzeln anzufügen oder abzuspielen. Mit einem Anzeigen-Kommando („Show/Browse“) könnte zusätzlich die tiefer gelegene Ebene der Tag-Hierarchie angezeigt werden, auf der wiederum die gleichen Kommandos möglich wären. Jederzeit wäre ein Abbruch-Kommando möglich, welches zur letzten ausgewählten Musik zurückführen würde.

[3]Bei diesem System ging es also weniger um das Erreichen aller Usability-Ziele, vielmehr sollte ein System geschaffen werden, mit dem möglichst viel über die Erwartungen und das Verhalten der Nutzer herausgefunden werden konnte.

Dieses Konzept hatte den Vorteil, ziemlich genau bisher bekannte grafisch-haptische Playlisten-Funktionen abzubilden. Allerdings würden die Dialoge dadurch sehr lang, weswegen für den WOZ-Test auf die Stufe der Eintragsauswahl verzichtet wurde, wie Abbildung 7.4 zeigt.

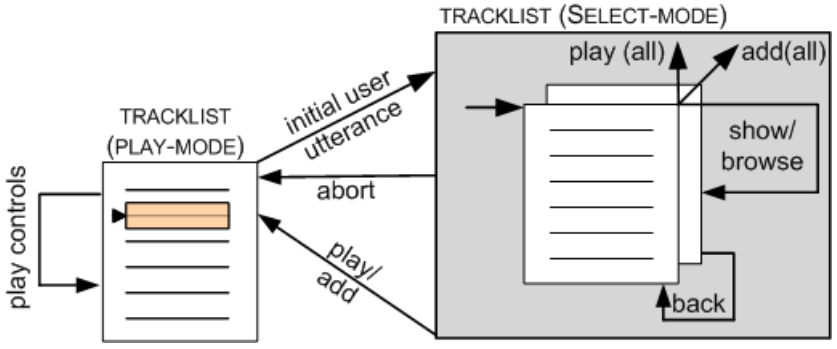

Abbildung 7.4: Veranschaulichung der verwendeten Dialog-Struktur (WOZ)

Daraus wird deutlich, dass die nächste Hierarchiestufe damit wesentlich schneller erreicht werden konnte.

Ein Beispiel soll diese Struktur veranschaulichen helfen. Abbildung 7.5 zeigt die Verwendung dieser Playlisten-Funktionalität für das Anfügen eines Titels an eine bereits vorhandene Titelliste.

User	System	Display
	(spielt „Chameleon“)	**1) Chameleon** 2) Watermelon Man 3) Sly 4) Vein Melter
PTT	Beep	dito
„Whenever, Wherever“ von Shakira	Interpret „Shakira“ – Titel „Whenever, Wherever“	Whenever, Wherever
	Beep	dito
Hinzufügen	Der Titel wurde zur Wiedergabeliste hinzugefügt. (spielt „Chameleon“ weiter)	**Chameleon** Watermelon Man Sly Vein Melter Whenever, Wherever

Abbildung 7.5: Beispiel Playlisten (WOZ)

Dabei unterbricht der Nutzer mit dem PTT-Knopf das Abspielen und wählt einen Titel aus. Dieser wird erkannt und im Display dargestellt, aber nicht abgespielt. Durch ein Hinzufügen-Kommando kann er nun das Anfügen des Titels an die aktuelle Playlist (die Trackliste) erreichen. An gleicher Stelle hätte der Nutzer auch ein Abspielen-Kommando verwenden können, durch den die Auswahl als neue aktuelle Playliste (Trackliste) übernommen geworden wäre.

Eine genaue Übersicht über das Systemverhalten im WOZ-Test gibt Anhang B.2.2 in Form von Dialog-Use-Cases. Über diese konnte das Systemverhalten vollständig beschrieben werden, wenn auch nicht so detailliert wie über State-Charts oder Flussdiagramme. Doch war ein solcher Detaillierungsgrad gar nicht nötig, da das WOZ-Tool von einem Menschen bedient werden sollte, für den verbale Repräsentationen einfacher in entsprechende Prompts und Abläufe umzusetzen waren.

Nach Nutzerwunsch: Ein vorläufiges Idealsystem

Nach dem WOZ-Test wurden aus Schlussfolgerungen Anforderungen für den Systementwurf abgeleitet. Doch bevor dieser Entwurf technische Rahmenbedingungen in Betracht zog, wurden betrachtet, wie ein Idealsystem aussehen könnte, welches sich direkt aus den Ergebnissen der Tests ableitet.

Im Wesentlichen sollte dieses System blind bedienbar sein, das Display nur noch Zusatzinformationen liefern. Dazu bot es sich an, den Dialog, wo es möglich war, zu kürzen und möglichst immer abzuspielen, um Musik als direktes Feedback zu benutzen. Dabei musste allerdings die im WOZ-Test ermittelte einheitliche Play/Browse-Struktur eingehalten werden, also bei Auswahl von Genre und Interpret jeweils in Auswahllisten verzweigt werden. Um auch in diesem hierarchischen Teil schneller vorgehen zu können, erschien die Nutzung von Barge-In ratsam.[4]
Die sich aus diesen Überlegungen ableitende Grundstruktur ist in Abbildung 7.6 festgehalten, daraus ist ersichtlich, dass der Nutzer nur noch Anfragen stellt oder eine Auswahl trifft, aber nicht mehr den Weg durch die Hierarchie bestimmt. Dieser ist vielmehr durch die Play-/Browse-Hierarchie bestimmt.

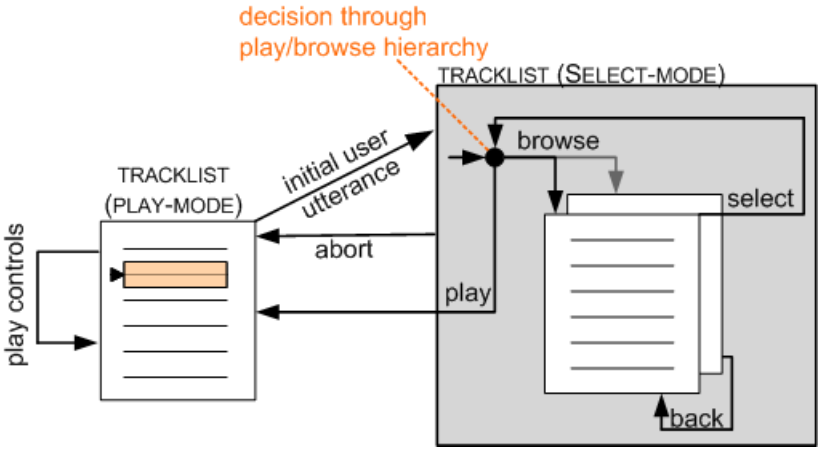

Abbildung 7.6: Beispieldialog im Idealsystem (Nach WOZ)

Äußerungen der Nutzer sollten dabei relativ formlos akzeptiert werden, also weder feste Bestandteile wie Bezeichner eines Modi („Spiele“, „Zeige“), noch verpflichtende Schüsselwörtern beinhalten.

[4] Die Bedeutung von Barge-In insbesondere beim Vorlesen von Listen wird insbesondere von McGlaun et al. [MAR+01] betont, 74% wollten hier Vorlesen unterbrechen können.

Dieses Idealsystem umfasste noch viele weitere Überlegungen, die jedoch vollständig in den nachfolgend beschriebenen Prototypen übernommen werden konnten. Für diesen konnte die an diskutierte Struktur größtenteils verwendet werden, bei der Freiheit der Nutzeräußerungen mussten allerdings Kompromisse geschlossen werden. Genauere Informationen zur konkreten Umsetzung finden sich in Abschnitt 7.2.3.

7.2 Prototyp „Dorothy“

„Dorothy“ war der Name des Mädchens, dass, aus Kansas weggetragen in das zauberhafte Land von Oz, den mächtigen Zauberer von Oz enttarnte [Bau00]. Für den Prototypen, dessen Struktur und Aufbau sich aus den Erkenntnissen des Wizard-of-Oz-Tests herleitete, war dieser Name Verpflichtung, ohne die Zauberei des WOZ-Tests eine trotzdem ebenbürtige oder sogar bessere Musikauswahl zu ermöglichen.

7.2.1 Anforderungen

Neben den aus den ersten Tests entwickelten Schlussfolgerungen, die in die Vorstellung eines im vorangegangenen Abschnitt diskutierten Idealsystem mündeten, bestanden noch andere Anforderungen an den zu entwickelnden Prototypen. Dieser sollte nicht nur einem Nutzertest standhalten, sondern auch geeignet sein, gezielt Erkenntnisse zu bestimmten Fragestellungen zu ermitteln. Da der Prototyp danach nicht weiter verwendet werden sollte, war somit eine Fixierung auf Untersuchungsziele möglich und nötig. So wurden neben strukturellen Anforderungen auch grundsätzliche Fragestellungen diskutiert, deren Test wiederum neue Anforderungen an das System stellte.

Doch zunächst musste der Prototyp natürlich auch die üblichen Steuerkommandos wie „Abspielen“, „Pause“, „Stopp“, sowie die Playmodi „Wiederholen“ und „Zufallswiedergabe“ beherrschen. Durch die Verwendung von Listen im Browse-Modus wurde auch die Unterstützung von zusätzlicher Listenkommandos wie „Nächster Titel“, „vorheriger Titel“ oder „2. Zeile“ nötig.

Aus der Verwendung von Listen folgten aber auch andere Überlegungen. Denn Listen widersprachen eindeutig dem Ziel der Unbeeinflussung beim Fahren, da durch Blicke zum Display Aufmerksamkeit gebunden würde. Als eine gute Idee wurde daher die auditive Präsentation solcher Listen (über Vorlesen oder „Anspielen“ der Listeneinträge) betrachtet, deren Einsatz getestet und bewertet werden sollte.
Darüber hinaus sollte auch ganz grundsätzlich geklärt werden, an welchen Stellen der Einsatz non-verbaler Interaktionselemente sinnvoll und hilfreich ist. Neben der Frage der grundsätzlichen Sinnhaftigkeit des Einsatzes war auch die der Form der Kombination (Töne oder Sprache, beides, Reihenfolge in Kombination) zu beantworten.[5] Letztendlich blieb die Anforderung, dem Einsatzzweck angemessene Töne zu integrieren, und deren Testbarkeit sicherzustellen.

In diesem Konzept eines Dialogs voller auditiver Elemente (Sprache, Töne, Musik) kam der Orientierung eine zentrale Rolle zu. Die bereits im WOZ-Test benutzte Informationsfunktion sollte eine Möglichkeit bieten, hier Orientierung zu schaffen. Insbesondere die Art

[5] Im Anhang B.2.1 findet sich eine solche Analyse.

und der Umfang der multimodalen Präsentation stand dabei im Mittelpunkt, was bereits auch in der Arbeit von McGlaun et al. [MAR+01] diskutiert worden war. Ihre Empfehlung nach einem umfassenden Nutzertest war auf auditiver Seite ein Vorlesen der Information und eine dazu kontextsensitiv die Seiten wechselnde visuelle Zusammenfassung der Information. Dieser Ansatz bildete den Start für entsprechende Überlegungen im Rahmen der Erstellung des Prototypen.
Aber auch Hilfe sollte Orientierung bieten, wenngleich nur eine rudimentäre Hilfe implementiert werden sollte. Trotzdem sollte es durch die Hilfe möglich sein, alle Kommandos abzurufen.

Schließlich wurde, basierend auf den diskutierten Überlegungen, ein Anforderungskatalog von Muss- und Kann-Kritierien für die Implementierung aufgestellt, welcher in Tabelle 7.1 dargestellt ist.

Kriterium	MUSS	KANN
Musikauswahl über Tags		
Genre	X	
Interpret	X	
Album	X	
Titel	X	
Playlisten		X
Egal/Alles spielen		X
Nur Konzept - Listen	X	
Abfangen Mehrdeutigkeiten		
Album/Titel	X	
Spracherkennung		X
Listenhandling		
Vorlesen	X	
Reinhören/Scannen	X	
Listenkommandos	X	
Modi		
Zufall	X	
Wiederholung	X	
Ablesen/Vorlesen/Reinhören	X	
Wiedergabesteuerung	X	
globale Kommandos		
Information	X	
Hilfe	X	
Zurück		X
Töne	X	

Tabelle 7.1: Muss-/Kann-Kriterien („Dorothy“)

Die „Zurück“-Funktion wurde dabei nur als Kann-Kriterium klassifiziert, der Aufwand der Implementation stand in keinem vernünftigen Verhältnis zu seinem Nutzen, auch da neue Anfragen global möglich waren und somit eine Korrektur auch anders erreicht werden konnte.

7.2.2 Technische Rahmenbedingungen

Neben den diskutierten Anforderungen waren nunmehr, anders als beim WOZ-Test, auch technische Rahmenbedingungen entscheidend für die Umsetzung des Systems. Für die Implementation wurde dabei auf die vorhandene Systemarchitektur für Sprachdialogsysteme von Harman/Becker Automotive Systems zurückgegriffen.
Diese richtete sich an den Einschränkungen des automobilen Einsatzes von Sprachdialogsystemen (siehe 2.4) aus und ermöglichte mittels einer Dialogbeschreibungssprache GDML (Generic Dialog Modelling Language, [HH04]) die Spezifikation des Dialogs. Um diesen Code dabei auf eingebetteten Systemen weiter lauffähig zu halten, werden Dialog und Grammatik kompiliert, wie dies in Abbildung 7.7 dargestellt ist.

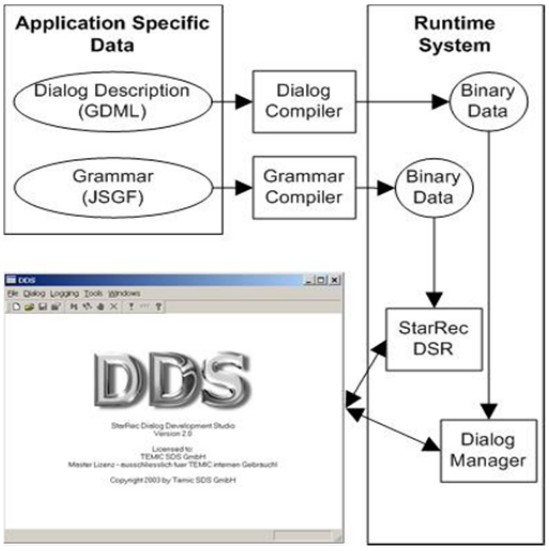

Abbildung 7.7: Toolkette für GDML-Dialoge [HH04]

Das DDS (Dialog Development Studio) ermöglicht, diesen erzeugten Code interaktiv auf normaler PC-Hardware zu entwickeln und zu testen, gleichzeitig ist der damit entwickelte Code in den eingebetteten Systemen lauffähig.
Im Vergleich zu standardisierten Beschreibungssprachen wie VoiceXML [MBC+04][6] ermöglicht GDML weiterhin zusätzliche Freiheiten in der Dialoggestaltung durch einen expliziten Dialogablauf und erweiterte Steuermöglichkeiten des Spracherkenners. Dadurch ist eine sehr genaue und systemnahe Entwicklung des Dialogs möglich, was für die Benutzung in eingebetteten Systeme unerlässlich ist.

Weiterhin ermöglicht GDML den Aufruf externer Funktionen über ein erweitertes Programminterface, dadurch ist es möglich, fast beliebige Funktionen aufzurufen und deren Ergebnisse im Dialog zu verarbeiten. Ein solche Funktion stellt dabei beispielsweise der Display-Service dar, der die Ansteuerung eines Displays erlaubt und somit eine multimodale Systemausgabe ermöglicht.
Eine Basis dieser Arbeit bot eine weitere Funktion, der MP3-Service. Dieser ermöglicht sowohl die Steuerung von Wiedergabefunktionen für MP3s als auch die Auswahl von MP3s über das Sprechen ihrer Tags. Während ersteres relativ einfach über die Benutzung einer

[6] Einen genauen Vergleich dieser beiden Beschreibungssprachen findet in einer Arbeit von Hamerich [Ham03].

MP3-Libary möglich ist, wird letzteres über die Benutzung eines „Grapheme to Phoneme“ (G2P) genanntes Verfahren erreicht, wobei einfach gesprochen aus einer vorhandenen Datenbasis Teile der Grammatik des Systems erzeugt werden und somit diese Einträge sprechbar sind. Wie dieses Verfahren speziell für die Musikmetadaten funktioniert, ist in Abbildung 7.8 schematisch dargestellt.

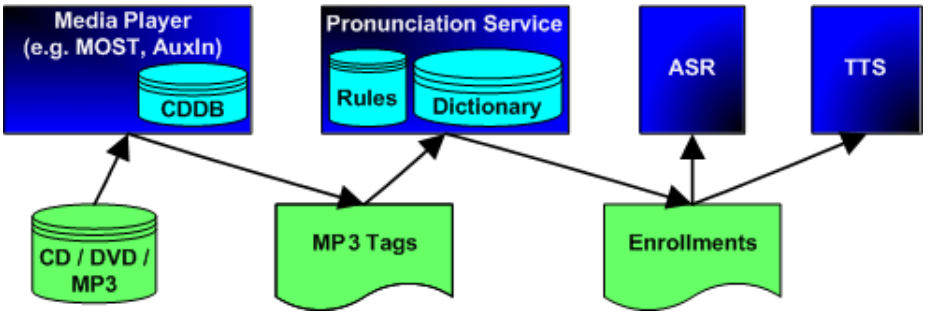

Abbildung 7.8: Grapheme to Phoneme (G2P) für ID3 Tags [WHHS05]

Zunächst werden dabei die Metadaten aus den MP3-Dateien ausgelesen, fehlende Informationen aus eventuell vorhandenen Datenbanken wie CDDB ergänzt. Diese MP3-Tags werden dann über gewisse Regeln und ein Wörterbuch in so genannte „Enrollment“-Regeln umgewandelt, die nun für diese Worte Spracherkennung und -synthese ermöglichen, sie also „sprechbar“ machen.

Die wichtigsten Komponenten des Prototypen „Dorothy“ sind in Abbildung 7.9 noch einmal zusammenfassend zusammengetragen.

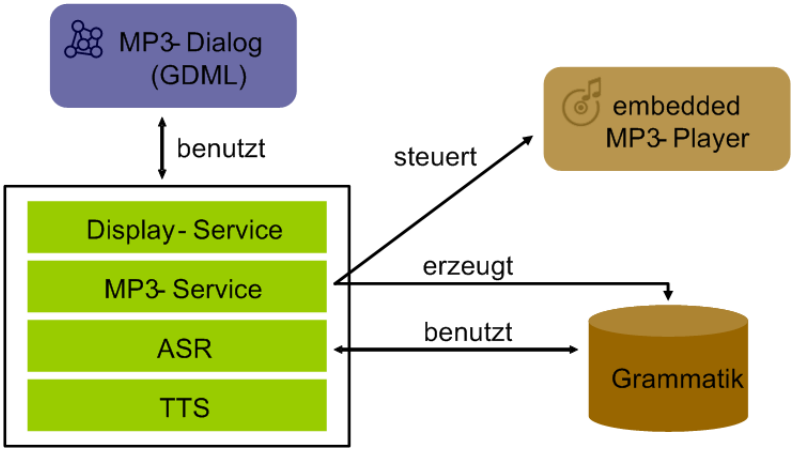

Abbildung 7.9: Technische Komponenten („Dorothy“)

Die Erstellung der konkreten Systemausgaben wurde durch das Prompt-Konzept von GDML vereinfacht, dass die einfache Einbindung von sowohl textueller als auch aufgenommener Sprache ermöglichte. Diese einfache Einbindung von Sounddateien erleichterte auch die Einbindung der Töne im weiteren Verlauf der Arbeit.

7.2.3 Umsetzung

Die Überlegungen zu dem am Ende von Abschnitt 7.1 beschriebenen Idealsystem bildeten die Basis für die hier diskutierte Umsetzung. So konnte die in Abbildung 7.6 veranschaulichte Struktur fast unverändert übernommen werden. Dennoch ist aus dieser nur das Prinzip, nicht der konkrete Ablauf der Musikauswahl ersichtlich. Diese ergab sich aus der im WOZ-Test ermittelten Play-/Browse-Hierarchie und ist in Abbildung 7.10 veranschaulicht.[7]

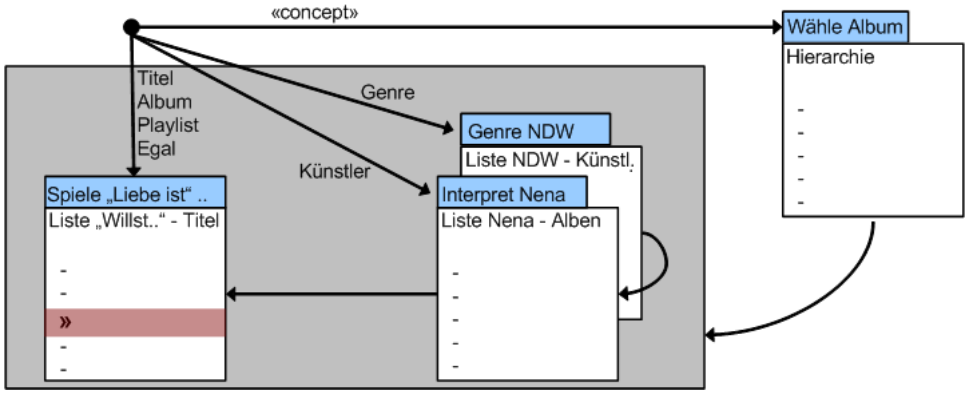

Abbildung 7.10: konkreter Ablauf der Musikauswahl („Dorothy“)

Aus dieser Darstellung wird ersichtlich, dass nach der Auswahl eines Titels oder Albums, einer Playliste oder durch die Auswahl „Egal“ jeweils direkt ein Titel abgespielt wird (in den blauen Kästen stehen die auditiven Systemausgaben, in den weißen ist der prinzipielle Bildschirminhalt abgebildet). Wird dagegen ein Genre oder ein Interpret gewählt, wird zunächst eine Liste angezeigt, aus der der Nutzer auswählen kann. Wählt der Nutzer hingegen nur das Konzept selbst aus („Zeige Alben“, „Alle Interpreten“ u.ä.), wird er in eine vorgeschaltete Liste gefühlt, in der er aus der gesamten Datenbasis, geordnet nach seinem Filter, auswählen kann.
Wie aus dieser Abbildung klar wird, ist der Vorgang nach maximal vier Äußerungen abgeschlossen (z.B. „Genre“,“NDW“, „Nena“, „Willst du mit mir gehen“), meist jedoch nach ein bis zwei Äußerungen. Dies entspricht im Wesentlichen der Forderung nach schneller Auswahl und berücksichtigt gleichzeitig die Ergebnisse des WOZ-Tests.

Bei den Anfragen selbst mussten allerdings Abstriche gegenüber dem Idealsystem gemacht werden, so musste auf jede Art von Multilingualität verzichtet werden. Es wurde die Entscheidung getroffen, für den Test mit deutschen Versuchspersonen einen deutschen Prototypen zu erstellen, der nur auf in deutscher Sprache vorliegende Musikmetainformationen zugreifen kann.
Ebenfalls konnte auch nicht auf die Schüsselworte verzichtet werden, damit eine eindeutige Erkennung der einzelnen Informationen möglich wurde.

Unter den weiteren, in Abschnitt 7.2.2 diskutierten Kriterien wurden zwei weitere Kriterien im Laufe der Bearbeitung ausgeklammert. Zum einen der Playlistenaufruf, da dies substantielle Erweiterungen am MP3-Service erfordert hätte, aber eigentlich zu Playlisten

[7] Die genauen Spezifikation findet sich im Anhang B.2.3.

nichts Wesentliches mehr herauszufinden war. Insofern wurde von einer Umsetzung abgesehen.
Ähnliche Probleme ergaben sich bei der Umsetzung der Disambiguierung nach Doppeldeutigkeiten in der Spracherkennung. Um dieses Prinzip testen zu können, hätte eine umfangreiche Emulation des Zustandes herbeigeführt werden müssen. Darauf wurde, auch im Blick auf die begrenzte Zeit des Tests, verzichtet.

Bei der integrierten Hilfe wurde die Entscheidung getroffen, diese im Wesentlichen in drei Teile zu teilen, einer allgemeinen Hilfe, einer Hilfe zu Funktionen der Wiedergabesteuerung und einer Hilfe zur Musikauswahl. Diese Teile bestanden aus langen Aufzählungen der Befehle und ihrer Bedeutungen, die Abbrechbarkeit mittels Barge-In war eine wesentliche Voraussetzung für die Implementierung des Hilfesystems in dieser Form.[8] Auf die gleichzeitige visuelle Präsentation dieser Hilfen im Display wurde verzichtet, um jegliche Ablenkung vom Fahrer fernzuhalten.[9] In Systemausgaben nach Erkennungsfehlern wurden diese Hilfen jeweils erwähnt, so dass den Nutzern Informationen zu möglicher Hilfe im Fehlerfall zur Verfügung standen. Die Verifikation der Richtigkeit erkannter Äußerungen des Nutzers wurden über einen Papageienmodus[10] und natürlich auch über die abgespielte Musik erreicht.

Neben diesen Entscheidungen, die vor allem die Weiterentwicklung der Musikauswahl im Blick hatten, wurden als neue Elemente des Dialogs non-verbale Interaktionselemente eingeführt. Als Grundlage diente dabei die Diskussion zum Einsatz non-verbaler Interaktionsobjekte für die sprachgesteuerte Musikauswahl im Auto, die in Abschnitt 4.4 diskutiert wurde.
Um Listen verständlicher zu präsentieren, wurde dem Vorlesen der Listeneinträge auch eine Funktion zum „Reinhören“ implementiert, Dialog 7.2 zeigt dafür ein Beispiel.

Dialog 7.2: Dialog unter Verwendung der „Reinhören“-Funktion

usr: Sängerin Madonna.
sys *zeigt Albenliste von Madonna*
sys: Interpret Madonna.
sys: Album 1.
sys *spielt Soundprobe von Album 1*
sys: Album 2.
sys *spielt Soundprobe von Album 2*
usr *<PTT>*
sys **BEEP**
usr: Spiele.
sys *spielt Album 2*

Damit wurde eine Möglichkeit geschaffen, eine Liste „erlebbar“ zu machen. Auch ohne genaue Kenntnisse über seine eigene MP3-Datenbank konnte so ein Eindruck gewonnen werden, welche Art Musik den Nutzer bei der Auswahl des jeweiligen Interpreten oder des

[8] Hierbei wurde PTT-Barge-In verwendet, bei dem der Nutzer durch das Drücken des PTT-Knopfes anzeigen muss, dass er die Systemäußerung unterbrechen und selbst sprechen möchte.
[9] Dies galt ebenso für das Kommando „Information“, welches nur auditive Informationen zum Titel übermittelte.
[10] Die explizite Wiederholung jedes Befehls durch das System.

jeweiligen Albums erwarten würde. Um aber selbst in diesem Modus eine verbale Orientierung zu bieten, wurde eine kontextsensitive Informationsfunktion eingeführt, die jeweils zum angespielten Titel eine entsprechende Information liefert.
Wird nun das Vorlesen oder Reinhören sehr langer Listen betrachtet (beispielsweise die aller Titel der MP3-Sammlung), bestand hierbei jedoch die Möglichkeit, beim naiven Vorgehen des kompletten Vorlesens oder Anspielens ein Übermaß an Information zu erzeugen, welches der Nutzer nicht mehr in seinem Kurzzeitgedächtnis speichern könnte [Mil56]. Dem Nutzer musste also eine gewisse Art Struktur vorgegeben werden. Durch die grafische Anzeige der Seite (jeweils sechs Einträge der Liste gleichzeitig) existierte eine solche Struktur bereits, welche dem Nutzer auch intuitiv aus Erfahrungen mit grafischen Benutzungsoberflächen verständlich war, durch Kommandos wie „nächste Seite“ oder „Blättern“ sollte dann auch die Verzweigung auf weitere Seiten möglich sein. Um dem Nutzer zu signalisieren, dass weitere Seiten vorhanden sind (und ihm somit auch eine Vorstellung von dem Umfang seiner Ergebnisliste zu geben), sollte ein non-verbaler Hinweis am Anfang der Liste integriert werden.
Weitere Verwendung sollten non-verbale Elemente bei der Identifizierung von Arten der Ergebnislisten bieten. Dazu soll angemerkt werden, dass Listen aus verschiedenen Gründen dem Nutzer präsentiert werden können. Zunächst kann er eine Auswahl getätigt haben, die ihn in den Browse-Modus versetzt hat (Genre, Interpret,...), aber ebenso kann seine Anfrage mehrdeutig sein (inhaltlich oder aufgrund fehlerhafter Spracherkennung) und die Liste die Möglichkeit zur Disambiguierung sein. Für diese drei Fälle sollte ein weiterer non-verbaler Hinweis dem Nutzer ermöglichen, sich schon vorher eine Vorstellung von der Art der Liste zu machen. Dies könnte Vorteile bringen, um Orientierungsproblemen vorzubeugen, da Nutzer in der Regel Mehrdeutigkeiten nicht vorausahnen und somit von der Präsentation solcher Listen überrascht werden könnten. Ein non-verbaler Hinweis könnte an dieser Stelle Klarheit über den Systemzustand vermitteln.

Weiter wurden Überlegungen angestellt, die Play-/Browse-Hierarchie in einer geeigneten Art auditiv darzustellen. Dabei war jedoch unklar, welche Informationen dieser Hierarchie hierbei im Mittelpunkt stehen sollten. Eine Möglichkeit wäre gewesen, jeweils die gewählten Tags (Genre, Künstler, Titel...) durch entsprechende non-verbale Elemente zu kodieren, eine andere die Stufen der Auswahl, die noch bis zum Abspielen folgen. Die Gefahr dabei war, dass beide Möglichkeiten miteinander verwechselt werden könnten und ebenfalls beim Nutzer zusammen mit den anderen auditiven Elementen eine Informationsüberflutung eingetreten wäre, weswegen von der Implementation Abstand genommen wurde.
Ebenfalls wurde die Verwendung von „Hyperlinks“ diskutiert, durch die beispielsweise beim Vorlesen die einzelnen Einträge als Hyperlinks (hervorgehobene durch Hintergrundgeräusche, siehe Abschnitt 4.4) betrachtet und durch einfaches Drücken von PTT ausgewählt werden können. Doch aus Gründen der Einheitlichkeit im System wurde auch diese Idee nicht umgesetzt.

Zusammenfassend ist in Abbildung 7.11 nochmals ein genauerer Ablauf der Musikauswahl abgebildet, der nun auch die strukturelle Einordnung der benutzten non-verbalen Interaktionselemente in das Konzept der Musikauswahl aufzeigt.
Dabei sind in dieser Abbildung alle Nutzeräußerungen türkis, alle Systemausgaben blau und die non-verbalen Interaktionselemente rot dargestellt. Die in der Abbildung erwähnten Einstellungen wurden im Prototypen über explizite Modi umgesetzt, die der Nutzer auch

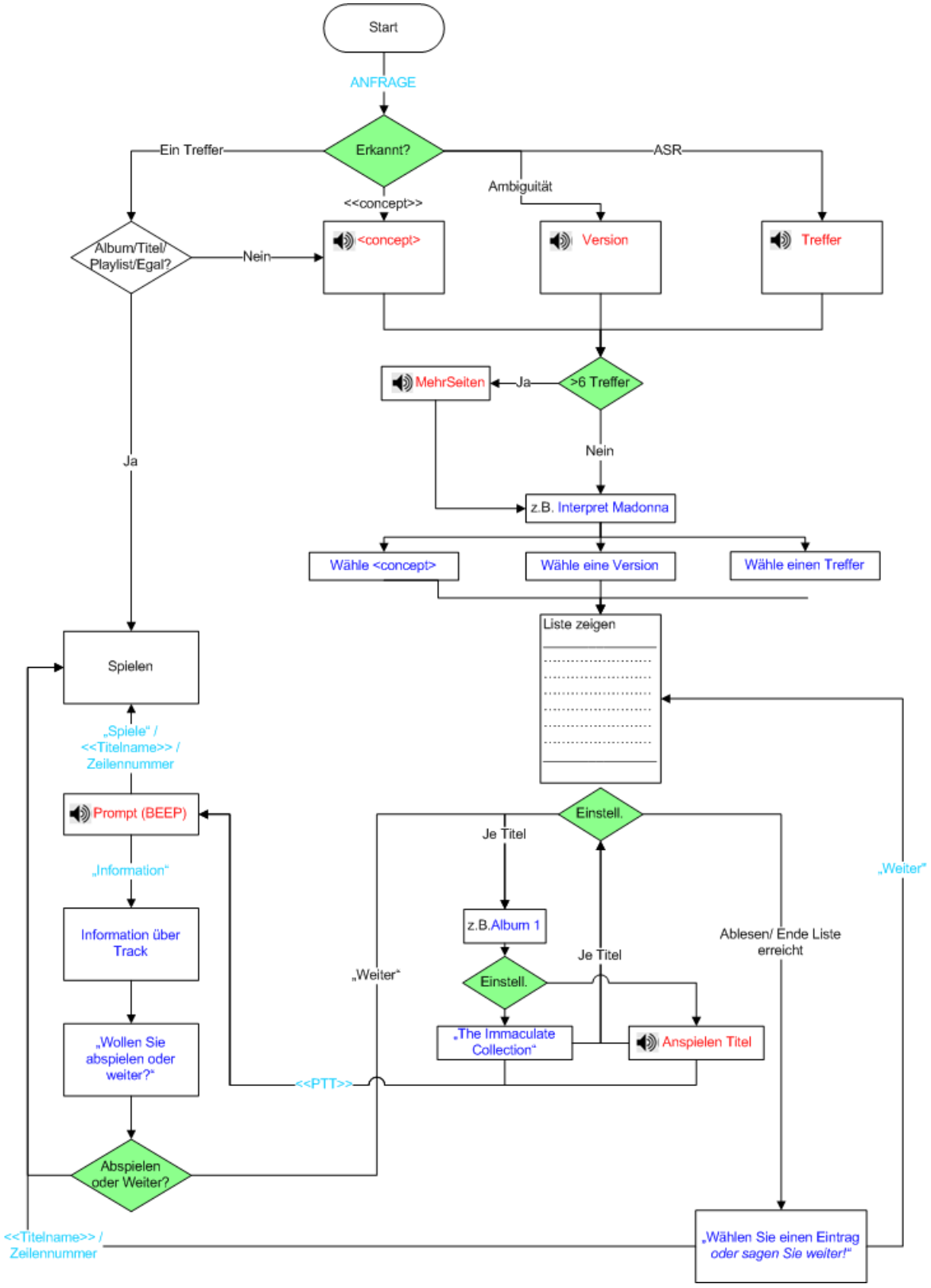

Abbildung 7.11: Struktur Musikauswahl über Einbeziehung non-verbaler Interaktionselemente („Dorothy“)

nicht umschalten konnte (siehe Abschnitt 8.1).

Nachdem durch diese Überlegungen klar geworden war, an welchen Stellen non-verbale Interaktionsobjekte eingesetzt werden sollten, musste überlegt werden, wie diese konkret ausgestaltet werden sollten.
Zunächst musste die Grundsatzentscheidung getroffen werden, welche Art non-verbaler Interaktionsobjekte für die Systemzustände in Frage kommen. Basierend auf der Diskussion in Abschnitt 4.4 konnten dabei frühzeitig Earcons ausgeschlossen werden, für eine Verwendung musikalischer Parameter gab es zu wenige abzubildende Systemzustände, und eine richtige Hierarchie sollte ebenfalls nicht abgebildet werden.
Weniger klar war die Sachlage bei der Verwendung von Auditory Icons. Es war nicht offensichtlich, ob eher ikonische oder symbolische Interaktionselemente benutzt werden sollten, für beide Möglichkeiten gab es eine Reihe von Argumenten. Die Benutzung ikonischer Interaktionselemente hätte dabei mehr der Lernfreiheit und Prägnanz der automobilen Nutzung entsprochen, aber dadurch auch die Ablenkung verstärkt, eine Verwechslung mit Warntönen bedeuten können und die Erstellung von einem einheitlich-wirkenden System von Tönen erschwert. Bei der Benutzung symbolischer Interaktionselemente waren diese Vor- und Nachteile jeweils vertauscht.

Deswegen wurde zunächst parallel sowohl nach sehr ikonischen Tönen und Geräuschen in freien Sounddatenbanken (wie zum Beispiel freesound [Jon06]) im Internet als aber auch nach einer Möglichkeit zur Erzeugung eigener, eher symbolischer, aber zusammenhängender Töne gesucht. Für Letzteres sollte die bereits in Abschnitt 1.2 erwähnte Kooperation mit der Hochschule für Musik „Carl Maria von Weber“ Dresden [mus06a] dienen. Die zeitlichen Restriktionen der Entwicklung ließen es jedoch nicht zu, auf die Ergebnisse dieser Arbeit zu warten, vielmehr musste, nachdem die Recherche in den Sounddatenbanken trotz intensiver Suche keine passenden Töne und Geräusche geliefert hatte, selbst an die Erstellung von Tönen herangegangen werden.

Es fiel dabei die Entscheidung für eine eher symbolischere Repräsentation, um die Ablenkung gering zu halten und gleichzeitig eine Einheitlichkeit und Verbindung zwischen den Tönen zu ermöglichen. Eingespielt wurden die Töne schließlich unter Mitwirkung eines Gitarristen, mit dem verschiedene Akkorde immer wieder variiert wurden, bis die gewünschte Prägnanz, aber auch Einheitlichkeit erreicht war (die verwendeten Töne finden sich auf der beigelegten CD-ROM, siehe Abschnitt A).[11]
In einer nachfolgenden Bearbeitungsphase wurde die Kombination der Töne und verbalen Systemausgaben festgelegt. Dabei wurde von der ursprünglich geplanten Reihenfolge der Form

> Töne → Wiederholung Nutzeräußerung → Handlungsanweisung
> *(Bsp: Töne - „Interpret Nena - Bitte wählen Sie ein Album“)*

abgewichen, um einheitlich im ganzen System die Wiederholung der Nutzeräußerung an den Anfang der Systemausgabe zu stellen. Die Töne ans Ende der Sequenz zu stellen wurde jedoch ebenfalls nicht als sinnvoll erachtet, da durch eine so späte Positionierung

[11] Eigentlich hätte sich an diese subjektive Bewertung ein Assoziationstest zukünftiger Nutzer anschließen müssen, in dem Nutzer nach verknüpften Eigenschaften zu Tönen gefragt und nach der Auswertung passende Töne bestimmt werden könnten. Aus Zeitgründen musste jedoch darauf verzichtet werden.

der Töne die eigentliche Aufgabe, den Nutzer schnell über Art und Fortführung der Liste zu informieren, nicht mehr gegeben wäre. So wurde entschieden, die Töne zwischen der Wiederholung der Nutzeräußerung und der Handlungsanweisung einzufügen:

> Wiederholung Nutzeräußerung → Töne → Handlungsanweisung
> *(Bsp: „Interpret Nena“ - Töne - „Bitte wählen Sie ein Album“)*

Am Ende entstand ein Prototyp, der die Vorgaben aus Fragebogen und WOZ effektiv in einem System integrierte, die Integration non-verbaler Interaktionselemente möglich machte und einige Komponenten zur Testerleicherung enthielt. In Abbildung 7.12 ist schließlich die grafische Oberfläche dieses Prototypen zu sehen, eine ausführbare Version des Prototypen findet sich auf der beigelegten CD-ROM, siehe Anhang A.

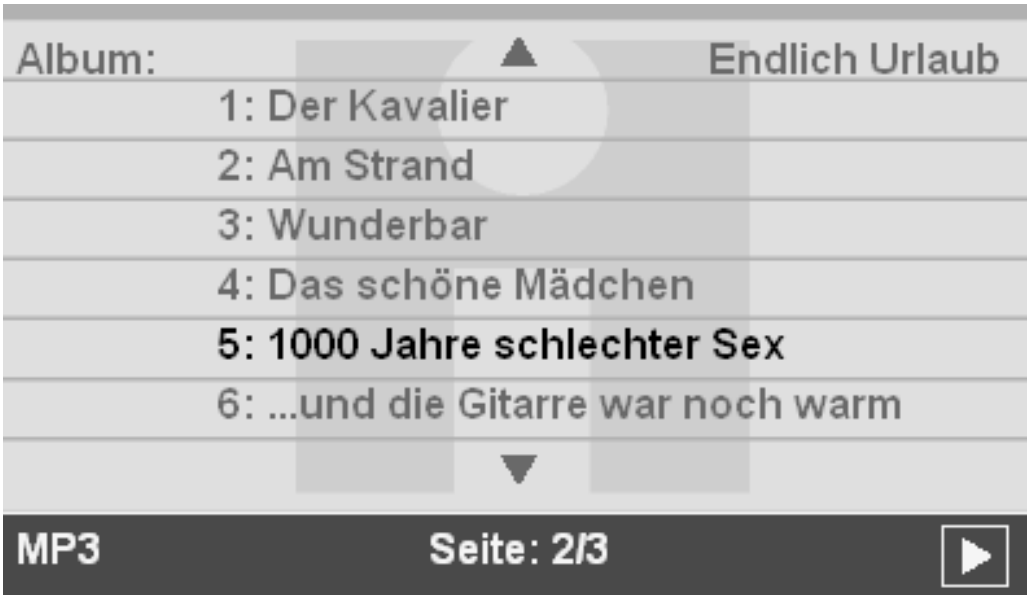

Abbildung 7.12: grafische Oberfläche (Prototyp „Dorothy“)

8 Abschlussevaluation

Nach der Erstellung wurde eine Evaluation des Prototypen „Dorothy“ durchgeführt. Bei diesem Test waren eine Reihe von Voraussetzungen einzuhalten und Anforderungen zu erfüllen, welche in Abschnitt 8.1 vorgestellt werden. Danach wird in Abschnitt 8.2 die Umsetzung diskutiert und schließlich in Abschnitt 8.3 die Ergebnisse präsentiert und Überarbeitungsempfehlungen gegeben. Schließlich wird diskutiert, inwiefern und unter welchen Umständen sich die Ergebnisse dieser Arbeit verallgemeinern lassen.

8.1 Voraussetzungen & Anforderungen

Die Evaluation des Prototypen diente zunächst dazu, nachzuprüfen, inwieweit die Schlussfolgerungen aus dem WOZ-Test richtig waren, also einer Validierung dieser Daten. Dazu sollte neben der konkreten Bewertung der Play/Browse-Hierarchie auch die allgemeine Systemzufriedenheit bewertet werden. Neben qualitativen Befragungen und Auswertungen von quantitativen Effizienzmaßen (Erfolgsrate, Zeitdauer) sollten auch die Beobachtung der Nutzer genutzt werden, Schwächen des Systems zu identifizieren.
Darüber hinaus sollte auch die Gebrauchstauglichkeit der Listenmodi (Ablesen, Vorlesen und Reinhören) und der verwendeten Auditory Icons bewertet werden.

Zur Durchsetzung der Testziele sollte jede Beeinflussung minimiert und auch konsequent objektive Testmaße zur Auswertung herangezogen werden.
Angedacht wurde die Fahrablenkung auch quantitativ über die Auswertungsoptionen des Lane Change Tests zu messen, um Selbsteinschätzung der Fahrer und Realität miteinander vergleichen zu können. Weiterhin sollten die Blickbewegungen der Nutzer während des Test in einer nachfolgenden Videoauswertung erfasst werden, um die Häufigkeit und Art der Blicke zum Display feststellen zu können und daraus den Erfolg der Maßnahmen zur Senkung der Ablenkung durch das Display ableiten zu können.
Ebenso erschien die Auswertung über Effizienzmaße wie Aufgabenerfüllrate oder Dauer der Bearbeitung überdenkenswert. Für deren Benutzung war es nötig, dass die zu testenden Modi (Ablesen, Vorlesen, Anspielen sowie Töne an/aus) gleichmäßig über alle Aufga-

bentypen verteilt wurden, aber auch Reihenfolgeeffekte durch randomisierte Anordnung minimiert wurden.

Die große Herausforderung des Tests bestand jedoch darin, einen gangbaren Weg für den Test der verwendeten Auditory Icons zu finden. Vollständig würde sich deren Nutzen erst nach längerem Einsatz bemerkbar machen, da manche Bedeutungsrepräsentationen erst gelernt werden müssten [VH05]. Allerdings musste wegen zeitlicher Einschränkungen der Arbeit von einem Langzeittest abgesehen werden.
Eine Möglichkeit zur Simulation eines solchen Langzeittests hätte sich mit einer Art Lernphase für die Töne vor dem eigentlichen Test geboten. Bei dem Aufbau einer solche Lernphase oder Tutoriums wäre jedoch zu bedenken, wie die Übertragbarkeit von Wissen sichergestellt werden sollte.
Als erste Möglichkeit hätte die Bedeutung abstrakt vermittelt werden können, auf der Basis von abstrakten Listen oder mit der einfachen Information, was welche Töne bedeuten. Dabei wäre allerdings die Übertragbarkeit der Wissens fraglich. Wird dagegen als zweite Möglichkeit die Lernphase realistischer gestaltet, gar durch eine Präsentation am Prototypen selbst, würden die Versuchspersonen massiv beeinflusst.
Ganz offensichtlich war die erste Möglichkeit nicht ausreichend und die zweite anderen Zielen des Tests zuwiderlaufend. Bevor jedoch über eine weitere Aufsplittung der Testgruppen in weitere Untergruppen (einmal mit Lernphase, einmal ohne) diskutiert wurde, wurde der Zweck einer solchen Lernphase nochmals kritisch hinterfragt.
Durch diese hätte danach bei den Versuchspersonen die Integration von bekannten Tönen mit dem System getestet werden können, also wie durch bekannte Töne eine Interaktion befördert werden könnte. Neben der Frage, welcher Wert für die Realität eine solche Aussage hätte, würde sich aber keine Aussage dazu treffen lassen, ob überhaupt diese Töne den Ereignissen zugeordnet würden.[1] Aber gerade auf dieser Prägnanz der automatischen Verbindung mit den untersuchten Ereignissen lag der Hauptfokus der gesamten Untersuchung. Insbesondere gilt dieses, da die typische Anwendung im Auto betrachtet wurde, in der die Nutzer üblicherweise keinerlei Lernphasen akzeptieren, sondern unnütze Funktionen (und als solche würde eine nicht intuitive Funktion verstanden werden) einfach abschalten.

Die zentrale Frage war also, welche Form und Ausgestaltung Töne im System Auto haben müssen, um intuitiv wahrgenommen zu werden, nicht ablenkend zu wirken und weder mit Signaltönen noch Musik verwechselt werden zu können. Diese Frage konnte nicht durch eine Lernphase vor dem Test beantwortet werden, sondern nur durch das intuitive Erkennen im Test und durch eine Befragung danach. Deswegen wurde auf eine Lernphase grundsätzlich verzichtet, jedoch der Nutzer auf die Beachtung der Töne hingewiesen und am Ende eine umfangreiche Nachbefragung zu den Tönen angedacht, genaueres wird im nun folgenden Abschnitt beschrieben.

8.2 Umsetzung

Nach der Klärung der grundlegenden Fragen musste nun ein Weg gefunden werden, für die verbliebenen Blöcke (Ablesen, Vorlesen, Reinhören sowie Töne ein/aus) eine Aufteilung

[1] Ganz abgesehen von der Tatsache, dass dies zu realitätsfremden Tests geführt hätte, wenn den Versuchspersonen Teile des Systems bekannt gewesen wären (Töne) und andere gänzlich unbekannt (Prompts, Struktur).

unter den Versuchspersonen abzuleiten. Dabei sollten in jedem Block ungefähr ähnliche Aufgaben (von Dauer und Schwierigkeitsgrad) bearbeitet werden, dass später auch über quantitative Maße feststellbar wäre, unter welchen Modi sich Probleme besonders häuften. Um Reihenfolgeeffekte zu verhindern, wurde die Abfolge der Blöcke so durchmischt, dass alle möglichen Kombination gleich oft auftraten. Die konkrete Anordnung der Aufgabenblöcke ist in Abbildung 8.1 dargestellt.

Ablaufplan

VPs					*Aufgabenteile* I.	II.	III.
1	7	13	19		A	B	C
2	8	14	20		A	C	B
3	9	15	21		B	A	C
4	10	16	22		B	C	A
5	11	17	23		C	B	A
6	12	18	24		C	A	B

Legende:
A ... Ablesen
B ... Vorlesen
C ... Reinhören

Abbildung 8.1: Anordnung Aufgabenblöcke nach Versuchspersonen („Dorothy“ Evaluation)

Unabhängig davon musste auch die Benutzung der Auditory Icons einmal pro Experiment ein- oder ausgeschaltet werden. Dies wurde jeweils in der Mitte des Experiments und während des laufenden Betriebs getan. Theoretisch hätte die Versuchsperson auch diese Umschaltung auslösen können, praktisch passierte dies jedoch nur ein Mal. Um auch hier Reihenfolgeeffekte zu vermindern, wurde das Experiment jeweils zur Hälfte mit eingeschalteten oder ausgeschalteten Tönen gestartet.

In den Aufgabenblöcken selbst wurden jeweils typische Aufgaben zu den implementierten Funktionen durchgeführt. Dabei wurde bei Aufgaben zur Musikauswahl jeweils direkt eine Nachfrage gestellt, ob das Verhalten den Erwartungen entsprach. Dadurch sollte direkt Verwirrung, wenn sie durch das Verhalten des Systems verursacht wurde, aufgefangen werden, bei späteren Nachfragen wäre wohl ein Großteil der Informationen verloren gegangen.
Weiterhin wurden bei der Präsentation der Aufgaben neue Wege beschritten. Beim WOZ-Test hatte die Präsentation über Karteikarten zwar zu einer sehr unbeeinflussten Präsentation der Aufgaben geführt, durch das notwendige Lesen der Karten aber die Fahrablenkung erhöht. Diesmal wurden die Aufgaben vorgelesen, aber dabei bei der Aufgabenentwicklung noch sorgfältiger darauf geachtet, das keine auf Lösungen hindeutende Formulierungen gewählt wurden.
Dies galt allerdings nicht für die freie Aufgabe, mit der der Test begann. In dieser Aufgabe sollte der Nutzer aus einer Liste von Titel und Alben (in der auch alle zusätzlichen verfügbaren Informationen wie Interpret oder Genre aufgeführt waren) frei einen Eintrag auswählen und das System dazu bringen, ihn abzuspielen.Hierbei wurden keinerlei Vorgaben über Wortwahl, Auswahl der Tags oder Ähnliches gemacht. Somit ermöglichte diese Aufgabe, bei der die Auswahl des Eintrags noch vor dem Beginn der Fahrsimulation stattfand, völlig unbeeinflusste Herangehensweisen an die Musikauswahl zu untersuchen und auch zu vergleichen, ob und wie dieses durch die im weiteren benutzten verbalen Aufgabenpräsentationen verändert wird.

Neben diesem Test wurde ebenfalls wieder eine Vorbefragung unter Verwendung des bereits im WOZ-Test genutzten Fragebogens durchgeführt und dem Nutzer die Fahrsimulation und die Grundlagen der Sprachbedienung (PTT, PTT-Barge-In, Vorhandensein einer Hilfe) vorgestellt. im Anschluss an den Test wurde wiederum eine Befragung mündlicher und schriftlicher Natur zu den Systemeindrücken (Allgemeine Einschätzungen, SUS), zu einzelnen Funktionen und den vorgestellten Modi durchgeführt. Bei den Tönen wurde dabei eine Zuordnung durchgeführt, bei der die Nutzer die im Prototyp benutzten Töne ihrer Bedeutung zuordnen mussten. Genauen Aufschluss über den Testablauf gibt der Versuchsleiter-Bogen im Anhang B.1.4, der die genauen Aufgaben und auch Anweisungen für den Versuchsleiter beschreibt.

Der Testaufbau orientierte sich im Wesentlichen an dem des WOZ-Tests, natürlich übernahm jetzt ein Rechner die Aufgabe des Wizards (siehe Abbildung 8.2).[2]
Zugunsten einer freieren und realistischeren Aufgabengestaltung wurde auf die eigentlich geplante formalisierte Benutzung des Lane Change Tests verzichtet (mit welcher höchstens die Hälfte der Aufgaben hätte durchgeführt werden können) und stattdessen dieser nur zur reinen Ablenkung eingesetzt. Zur Ermittlung der Fahrablenkung wurde wie im WOZ-Test eine qualitative Nachbefragung und zusätzlich die Videoauswertung zur Häufigkeit der Blicke zum Display eingesetzt.

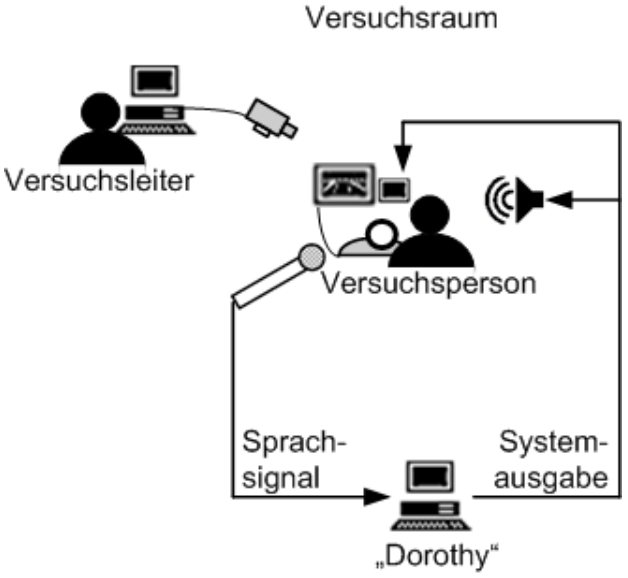

Abbildung 8.2: Versuchsaufbau („Dorothy“ Evaluation)

Als Testpersonen wurden nicht wieder die WOZ-Teilnehmer herangezogen, da diese massiv vorbeeinflusst gewesen wären. Weiterhin sollte eine andere Zusammensetzung der Stichprobe die Erkenntnisse aus den relativ jungen ersten Stichproben validieren helfen.
Schließlich konnten 24 Angestellte anderer Harman/Becker Standorte gewonnen werden, sich an der Untersuchung zu beteiligen.Wie aus Tabelle 8.1 hervorgeht waren die Teilnehmer in ihrer Mehrzahl Ergonomen und Gestalter, jedoch diese nicht im Bereich von

[2]Gegenüber dem WOZ-Test konnte dabei eine Verbesserung erzielt werden. Als PTT war nun ein Button am Lenkrad benutzbar und nicht wie beim WOZ die Leertaste einer in der Nähe liegenden Tastatur.

Sprachdialogsystemen tätig. Trotzdem war der Großteil schon einmal (meist oberflächlich) mit Sprachdialogsystemen in Berührung gekommen. Auffällig war die im Vergleich zu den anderen Stichproben seltene MP3-Nutzung, fast die Hälfte der Versuchsteilnehmer nutzte MP3s nur selten, dagegen fand sich keiner, der MP3s mehrfach täglich nutzte.

Eval	Teiln.	Alter	Geschlecht	Hintergrund	tech. Interesse
	24	75% 26-40	42% ♂	58% Gestalt/Psych	67% hoch/sehr hoch

Tabelle 8.1: Zusammensetzung Stichprobe Evaluation Dorothy

8.3 Ergebnisse & Überarbeitungsempfehlungen

Die wichtigste Erkenntnis aus dem Test war die Bestätigung der Play/Browse-Hierarchie, welche aus Fragebogen und Powerpoint-Befragung im Rahmen des WOZ-Tests abgeleitet wurden war. Auf die zu diesem Zweck während des Tests gestellte Nachfrage („War das Verhalten so wie erwartet?“) kritisierten die Versuchspersonen nur in 4% der Fälle die Orientierung oder die unklare Struktur. Auch gab es keine explizite Äußerung irgendeiner Versuchsperson, dass diese eine andere Struktur erwartet hätte.
Bei der Video-Auswertung der Displaybenutzung zeigte sich, dass zwar mehrheitlich kurze bis mittellange Blicke nach der Äußerung nach einer visuellen Bestätigung der getätigten Aktion suchten, aber selten länger dort verharrten. Dies bestätigte sich auch in der Nachbefragung, in der weniger als ein Drittel der Befragten angaben, sich vom Display abgelenkt gefühlt zu haben (beim WOZ-Test waren dies noch mehr als zwei Drittel der Befragten gewesen). Daraus konnte geschlussfolgert werden, dass die angestrebte stärkere Entkopplung der Interaktion vom Display erreicht werden konnte. Sicher ist auch die Verwendung des expliziten Papageienmodus dafür mitverantwortlich verantwortlich.
Weiterhin wurde PTT-Barge-In von jeder Versuchsperson verwendet und diese Möglichkeit einstimmig als positiv bewertet. Eine weitere Verwendung, auch unter Berücksichtigung, dass einige andere Funktionen nur dadurch richtig sinnvoll nutzbar werden, wird deswegen dringend empfohlen.

Neben diesen Erkenntnissen zeigte sich ebenfalls, wie zuvor bereits im WOZ-Test festgestellt, dass die (technisch notwendige) Benutzung der Schlüsselwörter („Interpret“, „Album“) nicht intuitiv ist, in ihrer ersten Äußerung versuchten viele Nutzer es zunächst ohne Schlüsselwort.
Dagegen nahm überraschenderweise die Komplexität in den Äußerungen gegenüber dem WOZ-Test ab. Waren im WOZ noch eher natürlichsprachliche Äußerungen unter Benutzung mehrerer Tag-Informationen gleichzeitig benutzt wurden („Ich möchte 'In the Ghetto' von 'Elvis Presley' hören“), wurde nun viel häufiger Kommandosprache und vor allem meist nur eine Tag-Information gleichzeitig genutzt. Über die Ursache lässt sich an dieser Stelle nur spekulieren, einerseits könnte die veränderte Zusammensetzung der Stichprobe dafür verantwortlich sein (mehr Ergonomen/Gestalter, die besser wissen, dass man eher „kommandobasiert“ mit Maschinen interagieren sollte), andererseits auch veränderte Art der Aufgabenpräsentation über verbale Äußerungen. Gegen diesen letzten Punkt spricht allerdings, dass der Effekt auch bei der freien Aufgabe auftrat. Hier sind weitere Untersuchungen nötig, um die Ursache dieses Effektes zu klären.

Die auf dieser unsicheren Basis erhobenen Daten wiesen dann auch nur bei einer Aufgabe größere Probleme nach, dem Abspielen der gesamten Liste in einer Auswahl. Diese „Spiele alles“-Funktion war sowohl von der Wortwahl als auch der eigentlichen Herangehensweise einer Reihe von Versuchspersonen unklar.
Für eine kritische Betrachtung der Qualität der im Experiment erhobenen Daten und deren korrekte Interpretation, müssen an dieser Stelle zunächst Abweichungen vom geplanten Experiment betrachtet werden. Die Aussagekraft der erhobenen Daten wird durch eine unterschiedliche Interpretation der Anweisungen durch die verschiedenen Versuchsleiter eingeschränkt. Dies erschwerte die Ableitung von Erkenntnissen aus der Anzahl der erfolgreich bearbeiteten Aufgaben. Im Speziellen handelt es sich hierbei um das Abbruchkriterium, wonach der Versuchsperson drei Versuche bis zum Abbruch der Aufgabe eingeräumt wurden. So kann dieses Abbruchkriterium als exakt drei nicht erfolgreiche Äußerungen verstanden werden, aber auch als drei komplette Wiederholungen der Aufgabe von Beginn an. Hinzu kam, dass die Versuchspersonen dazu tendierten, einmal aufgetretene Fehler durch sofortige Wiederholung ihrer vorherigen Anfrage zu korrigieren. Auch hier war es von Seiten der Versuchsleiter nicht immer möglich die vorgegebenen Anweisungen genau umzusetzen.
Nur bei einer Aufgabe konnten aus diesen Daten Erkenntnisse über bestehende Usability-Probleme gewonnen werden. Diese „Spiele alles“-Funktion war sowohl von der Wortwahl als auch der eigentlichen Herangehensweise einer Reihe von Versuchspersonen unklar.
Ein stärkerer Hinweis auf die Verfügbarkeit und mögliche Wortwahl dieser Funktion schien sinnvoll, eine Anmerkung einer Testperson, ähnlich der Darstellung beim iPod, ein „Spiele alles“ als ersten Eintrag in die Liste der auszuwählenden Titel zu integrieren, schien dafür ein gangbarer Weg. Dieser „Spiele alles“ sollte jedoch nur auf der ersten Seite der Liste auftauchen (an dieser Stelle wird der Hinweis gebraucht), um mehr Platz für die Einträge der Liste auf den restlichen Seiten zu erhalten.

Weitere Erkenntnisse zu den Funktionen wurden nun vor allem aus Kommentaren der Nutzer im Test und der dem Test folgenden Nachbefragung gewonnen.
In dieser Nachbefragung gaben beispielsweise etwas mehr als die Hälfte der Versuchsteilnehmer an, die „Zurück“-Funktion zu vermissen, im Test jedoch benutzten es weniger. Personen, die von der „Zurück“-Funktion Gebrauch machten, passten sich jedoch schnell an und verzichteten schnell auf die weitere Benutzung dieses Kommandos. Hier scheint das Konzept der Korrektur durch Neuanfrage trotz eigentlich gegenläufiger Erwartungshaltung der Nutzer größtenteils angenommen worden zu sein, eine Implementation eines „Zurück“-Kommandos für zukünftige Systeme scheint trotzdem sinnvoll, wenngleich nicht notwendig, zu sein.
Dagegen gab es einen großen Bedarf an einem Kommando für ein Hauptmenü, zu dem bei Fehlerfall zurückgekehrt werden könnte. Zwei Drittel der Testpersonen versuchten so, sich aus ausweglosen Situationen zu befreien, und, anders als beim „Zurück“-Kommando, versuchten sie mehrheitlich immer wieder dieses Kommando zu benutzen. Eine Integration eines solchen Kommandos erscheint daher absoult notwendig.
Weiterhin wurde bei der Bearbeitung der Aufgaben zum Einschalten des Wiederholmodus mehrfach die Frage gestellt, ob jetzt der Titel oder das Album wiederholt werden sollte. Diese intuitive Erwartung zweier Wiederholungsmodi sollte Anlass für die zukünftige Umsetzung beider Modi bieten. Weiterhin fiel auf, dass bei den Aufgaben zu Wiederholung und Zufallswiedergabe am häufigsten die Hilfe bemüht wurde, diese Funktionen sollten also in der Hilfe an prominenter Stelle präsentiert werden.

Die Hilfe wurde generell gut aufgenommen, doch öfters die fehlende Kontextsensitivität bemängelt. Diese erscheint jedoch angesichts der Masse der global im System verfügbaren Kommandos nicht sinnvoll umsetzbar.

Auch bei den untersuchten Listenmodi ergaben sich bei der Auswertung aus den quantitativen Maßen keine verwertbaren Informationen. Über die Nachbefragung konnte jedoch festgestellt werden, dass der „Vorlesen"-Modus den Versuchspersonen am besten gefallen hatte (siehe Tabelle 8.2). Dieser Modus wurde im Durchschnitt am besten bewertet, und wurde sowohl von den meiste Versuchspersonen bevorzugt wie von den wenigsten abgelehnt. Ein Problem des „Vorlesens" im Nutzertest bestand jedoch darin, dass Nutzer fälschlicherweise annahmen, dass nicht die gesamte, sondern nur die vorgelesenen Einträge der Liste sprechbar wären. Die Sprechbarkeit aller Listeneinträge müsste bei der Verwendung von „Vorlesen" als Standardmodus stärker motiviert werden. Kann dies nicht gewährleistet werden, sollte „Ablesen" als Standard benutzt werden. In jedem Fall sollte zwischen beiden Modi umgeschaltet werden können.
Das „Reinhören" gefiel den wenigsten Testpersonen. Doch zeigte sich hier eine große Streuung, wie einerseits die hohe Standardabweichung zeigt. Andererseits bezeichneten die Nutzer diesen Modus im Test abwechselnd als völlig überflüssig oder nervig, andere dagegen als gute Idee oder absolut notwendig. Aufgrund dieser Kontroverse wird die Verwendung von „Reinhören" als Kommando statt als Modus empfohlen, so dass der Nutzer bei Bedarf auf diese Funktion Rückgriff hat und sich auch ein auditives Abbild seiner Auswahl machen kann. Gerade zur Unterstützung für den Nutzertypus des „Stöberers" wäre dieses Kommando eine sinnvolle Unterstützung.

Listenmodi	Bewert.	StdAbw	Lieblingsmodus von	Am meisten abgelehnt von
Ablesen	2,79	1,14	7 Versuchspersonen	10 Versuchspersonen
Vorlesen	2,33	1,24	7 Versuchspersonen	5 Versuchspersonen
Reinhören	3,42	1,59	4 Versuchspersonen	15 Versuchspersonen

Tabelle 8.2: durchschnittliche Bewertung Listenmodi (Evaluation „Dorothy")

Der Test der Auditory Icons endete dagegen mit einem überraschend negativen Ergebnis. Sowohl aus den quantitativen als auch den qualitativen Daten der Nachbefragung konnte kein Mehrwert der Töne für die Bedienung des Systems festgestellt werden. Im Test konnte keiner der 24 Testpersonen einen eindeutigen Eindruck davon gewinnen, was die Töne ausdrücken sollten. Zwar wurden die Töne im Nachhinein in 42% der Fälle richtig ihren Bedeutungen zugeordnet, weit mehr, als die Wahrscheinlichkeit für zufällige Zuordnung (1/6, rund 17%) dafür betrug. Ebenfalls wurde trotzdem von einem Drittel der Befragten ein Mehrwert der Töne im konkreten Fall bestätigt.[3] Generell sollte eine Verwendung von Tönen also weiter angedacht werden.
Trotzdem bleibt die Frage, warum die speziellen Töne im Testverlauf so schlecht erkannt wurden. Sicher hätte ein Langzeittest hierbei andere Ergebnisse gebracht. Nach mehr Zeit der Beschäftigung mit dem System hätten sicher mehr Personen den Tönen Bedeutungen zuordnen können. Auch war durch den Test einer solchen Neuheit wie der sprachgesteuer-

[3] Allgemein waren sogar 88% der Befragten der Meinung, dass eine Verwendung von Tönen in Sprachdialogsystemen einen Mehrwert bieten würde.

ten Musikauswahl sicher die Aufmerksamkeit vor allem auf die Kernfunktionen gerichtet. Doch war, wie bereits in Abschnitt 8.1 diskutiert, vor allem das Ziel, **prägnante** und **schnell selbst erklärende** Töne zu finden, die auch unter solchen Umständen hätten erkannt werden sollen. Die Verbesserung beider Eigenschaften soll nachfolgend betrachtet werden.

Um die Prägnanz von Tönen zu verbessern, sollen sie zukünftig ikonischer mit ihrer Bedeutung in Zusammenhang stehen, auch auf die Gefahr hin, auf die Dauer nervig oder mit vorhandenen Signaltönen im Auto verwechselt zu werden. Wie stark überhaupt eine Verwechslungsgefahr mit solchen Tönen besteht, wäre zuvor jedoch lohnendes Ziel einer Untersuchung.
Weiterhin war die fehlende Prägnanz eventuell auch auf die zeitliche Einordnung der Töne, wie am Ende von Abschnitt 7.2.3 dargestellt, zurückzuführen. „Eingeklemmt“ zwischen Bestätigung der Nutzeräußerung und Handlungsanweisung ist es möglich, dass für die Nutzer die Töne sich mit den verbalen Äußerungen überlagert wahrgenommen und damit die Bedeutung der Töne maskiert wurden. Ein Umbau dieser Reihenfolge würde jedoch mit dem guten Abschneiden des Papageienmodus kollidieren, wodurch eine Lösung dieses Teilproblems nicht absehbar ist.
In jedem Fall sollten bei der Einbindung von Tönen diese von einem Musiker erstellt werden (wie dies geplant war, aber zeitlich nicht umgesetzt werden konnte) und in anschließenden Assoziationstests den Nutzern zur Begutachtung vorgelegt werden, bevor sie im Rahmen eines Systems getestet werden.

Zur Verbesserung der Lerngeschwindigkeit und der Selbsterklärungsfähigkeit dieser Töne, brachte ein Nutzer im Test eine eigentlich nahe liegende Lösung ins Gespräch, die Kopplung von visuellen und auditiven Elementen. Konsequent umgesetzt würde dies bedeuten, dass mit dem Erklingen eines Tons ein dem Ton entsprechendes Icon aufleuchtet, um so dem Nutzer die Verbindung beider Elemente zu signalisieren. Beispielsweise könnte so der „mehr Seiten“-Ton an das Dreieck-Icon mit der gleichen Bedeutung gekoppelt werden (wie in Abbildung 8.3 angedeutet). Natürlich würde eine solche Displayaktivität Einfluss auf die Fahrablenkung haben, deren Ausmaß untersucht werden müsste.
Zusätzlich könnten die Töne auch in die Online-Hilfe in der Form eingebunden werden, dass sie dort direkt vorgespielt werden und dazu die Bedeutung erläutert wird.

Die Fahrablenkung konnte, wie bereits erwähnt, nur qualitativ über die Nachbefragungen betrachtet werden. In dieser gaben die Nutzer an, hauptsächlich durch das Finden der Kommandos abgelenkt worden zu sein, was sicher mit der Erstbenutzung zusammenhängt. Erst dann wurde die Menünavigation und das Warten auf die Systemreaktion als ablenkende Faktoren benannt. Letzteres war die Folge eines Bugs im verwendeten MP3-Service, der die Systemreaktion verzögern konnte.
Einige Versuchspersonen äußerten darüber hinaus die Vermutung, ein solch neues System üblicherweise zuerst im Stand kennenzulernen und auch häufig zu bedienen, eine Untersuchung eines solchen zweiten wichtigen Anwendungsfalls wäre sicher nicht nur zur Validierung der Ergebnisse sinnvoll. Ebenso bestanden einige Versuchspersonen auf der Feststellung, dass sie besser mit dem System klargekommen wären, wenn die Untersuchung mit ihrer eigenen Musiksammlung durchgeführt worden wäre. Eine solche Untersuchung könnte eine weitere Möglichkeit zur Validierung der bisher gewonnen Erkenntnisse bieten.

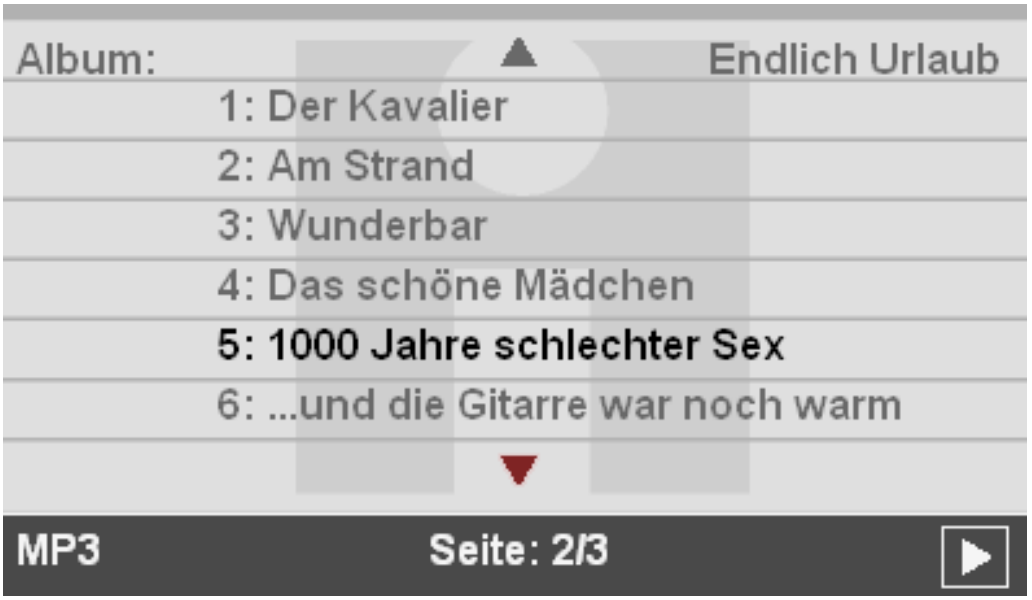

Abbildung 8.3: grafische Hervorhebung bei gleichzeitigem Abspielen des „mehr Seiten"-Tons (Idee nach Evaluation „Dorothy")

Schließlich wurde das System in seiner Gesamtheit gegenüber dem WOZ-Test fast genauso gut bewertet, erreichte mit einer Schulnote von 2,5 und einem SUS-Score von 77,5 noch überwiegend gute Einschätzungen. Weiterhin äußerten 63% der Versuchspersonen, die Musikauswahl wäre mit dem System besonders einfach gefallen.

Die vollständigen Ergebnisse finden sich in Anhang B.1.4.

8.4 Verallgemeinerbarkeit

Angesichts solch positiver Nutzerreaktionen stellt sich die Frage, ob sich die Prinzipien des im Rahmen dieser Arbeit entwickelten Dialogdesigns auf andere Sachverhalte übertragen lassen. Dabei wird zunächst die Übertragbarkeit auf andere Musik- und Audioformen, danach auf andere Interaktionsumgebungen, aber auch auf andere Anwendungen im Auto diskutiert. Schließlich wird betrachtet, ob sich auch Erkenntnisse für Umgang mit großen Datenmengen in Sprachdialogsystemen ableiten lassen.

Die triviale Übertragbarkeit der Erkenntnisse auf andere Dateiformate als MP3 folgt dabei schon aus der allgemeinen Diskussion von Musik in Kapitel 3. Die in der Arbeit verwendeten ID3-Tags sind auch für andere Dateiformate verfügbar. Damit sollte eine Erweiterung des Systems auf andere Formate keinerlei Änderungen am Dialog erforderlich machen.
Enthalten solche Medien aber etwas anderes als populäre Musik wie beispielsweise Klassik, Hörbücher oder Podcasts, ist die Übertragbarkeit nicht sicher. Klassik baut in seiner Beschreibung mit Komponist, Werksverzeichnissen und ausführenden Interpreten auf einer ganz anderen Beschreibungsstruktur als der hier betrachteten auf. Zwar lässt sich einiges ineinander überführen (wie bereits in Abschnitt 6.1.1 erwähnt), dennoch wäre ein Test angeraten, um spezielle Anforderungen zu ermitteln. Dies gilt analog auch für Hörbücher und Podcasts.

Auch in einem nicht-automobilen Umfeld kann die in dieser Arbeit entwickelte Dialogstruktur eingesetzt werden, da viele anfängliche Überlegungen zur Musikauswahl auf eher allgemeinen Einstellungen und Erwartungen zu Musik beruhten. Obwohl das Ziel der weitestgehenden Display-Unabhängigkeit direkt aus Beobachtungen zur Nutzung im Auto abgeleitet wurde, ist dies sicher auch in Situationen sinnvoll, in denen die visuelle Modalität nur zeitweise (Smart Home) oder gar nicht zur Verfügung (Telefon) steht. Die Anwendung der in der Arbeit erstellten Play/Browse-Hierarchie könnte helfen, dieses Ziel zu erreichen. Dazu müsste jedoch konsequent auf eine Art der auditiven Präsentation der Listen gesetzt werden, ein Ablesemodus wäre dem Einsatzszenario nicht angemessen. Auch könnte sich insbesondere bei Anwendungen im Telefoniebereich die Reinhören-Funktion als sinnvoll erweisen, wenn beispielsweise eine Anwendung zum Musikverkauf erstellt werden soll. Vorteil wäre hier jeweils, dass trotz Bedienung des Systems der Nutzer nicht an einer Stelle vor einem Display gebunden ist, sondern sich frei bewegen kann.
Soll aber das volle multimodale Potential ausgeschöpft werden, visuelle und auditive Modalität also zu einem effektiven System gekoppelt werden (beispielsweise für die Bedienung des Mediencenters im Wohnzimmer von der Couch aus), könnte die Struktur so nicht direkt übernommen werden. Hier wäre zunächst eine umfangreiche Modalitätsanalyse vonnöten, um zu bestimmen, welche Informationen wie präsentiert werden sollten.

Bei der Diskussion der Übertragbarkeit einzelner Prinzipien auf andere sprachbediente Automobilanwendungen muss zunächst PTT-Barge-In betrachtet werden. Im Test konnte eindeutig der Mehrwert des Einsatzes dieser Methode aufgezeigt werden. Die Zustimmung war dabei weniger von der konkreten Musikauswahl abhängig, sondern wurde vor allem mit der Beschleunigung der Kommunikation begründet. Deswegen sollte sie auch in anderen sprachbedienten Automobilanwendungen eingesetzt werden. Weiterhin sollte darüber nachgedacht werden, die Ablenkung zu senken, indem möglichst viele Sprachfunktionen ohne Blickkontakt zum Display bedienbar gemacht werden.
Nicht geklärt werden konnte, welche Art von Tönen zur Erweiterung des sprachlichen Dialogs im Auto eingesetzt werden können. Im Spannungsfeld zwischen zu hoher Ablenkung und Wahrnehmbarkeit konnte lediglich festgestellt werden, dass symbolische Töne wahrscheinlich nicht geeignet sind.

Dagegen lassen sich für den sprachlichen Umgang mit großen Datenmengen nur einige Aussagen zu Listen treffen, die bei dem Zugriff auf große Datenmengen zwangsläufig entstehen. Wurde am Anfang der Arbeit noch vermutet, dass eine möglichst geschickte oder fehlerkorrigierende Darstellung dieser Listen den Umgang mit großen Datenmengen vereinfachen würde und somit nur eine „günstige“ Listendarstellung gefunden werden müsste, wurde später klar, dass eine Vermeidung von Listen viel hilfreicher sein konnte. Für die Musikauswahl konnte dies recht effektiv erreicht werden, doch ob diese „Verbergung“ der Komplexität auch in anderen Themenbereichen sinnvoll sein kann, ist nicht abzuschätzen.

9 Zusammenfassung und Ausblick

In diesem Kapitel sollen die Erkenntnisse der Arbeit noch einmal zusammengefasst betrachtet werden. Anschließend wird eine Bewertung der Arbeit vorgenommen und ein Ausblick auf mögliche weiterführende Arbeiten gegeben.

9.1 Abschließende Betrachtung und Zusammenfassung

In dieser Arbeit wurde die sprachgesteuerte Navigation in komplexen Strukturen am Beispiel eines MP3-Players im Auto betrachtet.
Dazu erfolgte zu Beginn ein umfangreiches Literaturstudium zu den Themen Sprachdialogsysteme, Musik, non-verbale Interaktionselemente und Usability Engineering. Basierend auf diesen Überlegungen wurde eine Befragung und ein Wizard-of-Oz-Test durchgeführt, mit denen Erkenntnisse zu potentiellen Nutzern, ihren MP3-Nutzungsgewohnheiten und Vorstellungen einer idealen Musikauswahl gewonnen wurden. Durch den Wizard-of-Oz-Test konnte zusätzlich ein Eindruck der intuitiv von den Nutzern erwarteten Struktur und Vorgehensweise einer sprachgesteuerten Musikauswahl erhalten werden.
Mit diesen Erkenntnissen war es möglich, die bereits vorher entwickelten Ideen zur Systemstruktur entsprechend anzupassen und dabei auch Ideen zur Benutzung non-verbaler Interaktionselemente in das Konzept zu integrieren. Anschließend wurde ein Prototyp erstellt und die Umsetzung der Nutzererwartungen in einer abschließenden Evaluation überprüft. Dabei wurde auch untersucht, inwieweit die non-verbalen Interaktionselemente den Dialog verbessern konnten.
Schließlich wurden diese Evaluation ausgewertet und neben Überarbeitungsempfehlungen auch die Verallgemeinerbarkeit der Erkenntnisse diskutiert.

9.2 Bewertung des Ergebnisses

Im Ergebnis dieser Arbeit entstand eine realistisch umsetzbare, intuitive und an neuester Forschung orientierte Dialogsteuerung für sprachgesteuerte Musikauswahl im Auto, welche

direkt in der Praxis anwendbar ist. Dabei konnte gezeigt werden, dass auch ohne einen Eingriff in die Spracherkennung oder die Benutzung erweiterter Metadaten ein System erstellbar ist, welches zu großen Teilen den Erwartungen der Nutzer entspricht.
Dies wurde durch Erkenntnisse aus einem umfassenden Literaturstudium und der strikten Orientierung am Nutzerwillen möglich. Die starke Einbindung von Nutzern im Form einer Befragung, eines Wizard-of-Oz-Tests und einer Abschlussevaluation ermöglichte die frühzeitige Konzentration auf von Nutzern gewünschte Funktionen, welche daraufhin schwerpunktmäßig verbessert und an den Nutzerwillen angepasst werden konnten. Durch die verschiedenen Zusammensetzungen der Stichproben war es möglich, eine Validierung gewonnener Ergebnisse vorzunehmen und so die Verallgemeinerbarkeit abzusichern.
Einen Schwerpunkt bildete die Play/Browse-Hierarchie, welche Ansprüche verschiedener Nutzertypen integrierte und deren Usability-Nutzen in der abschließenden Evaluation bestätigt werden konnte. Ebenso reduzierte ein konsequent auditiver Entwurf wirkungsvoll die Ablenkung durch das Display. Durch die globale Verfügbarkeit der meisten Befehle konnte eine starre Struktur vermieden und eine intuitiv zugängliche Struktur geschaffen werden.
In einer Betrachtung zur Verallgemeinerbarkeit der Erkenntnisse konnte dargestellt werden, dass die Ergebnisse der Arbeit nicht nur für die sprachgesteuerte Auswahl von MP3s im Auto, sondern auch allgemeiner nutzbar waren.

Bei der Verwendung non-verbaler Interaktionselemente bewirkte das Fehlen einer solchen Struktur allerdings, dass eine wesentliche Anwendungsmöglichkeit – Struktur „hörbar zu machen“ – größtenteils entfiel. Wesentliche Arbeiten in diesem Gebiet beschäftigten sich vor allem mit dieser Anwendungsmöglichkeit. Trotz der freien Struktur wurde dennoch versucht, strukturelle Informationen in non-verbale Repräsentation zu überführen, aber diese erfassten nur Teilbereiche (die Listen). Nach Auswertung der Literatur wurden eher symbolische Auditory Icons als geeignet befunden, diese Informationen zu transportieren. Damit konnte aber nicht die gewünschte Prägnanz erreicht werden – im Nutzertest bemerkten viele Testpersonen sie gar nicht. Mehrere Gründe dafür wurden diskutiert, so spielten sowohl die Auswahl der konkreten Töne, der Ort ihrer Einbettung als auch die zu wenig im System verfügbaren lernförderlichen Hinweise eine Rolle. Ein weiterer Versuch der Integration unter Beachtung dieser Punkte wurde empfohlen, da sich die Nutzer sehr positiv zu einer prinzipiellen Verwendung von Tönen im Dialog äußerten.
Als weitere Verwendungsmöglichkeit für non-verbale Interaktionselemente wurde die Idee untersucht, diese zur Repräsentation der Daten selbst zu nutzen, praktisch also durch Musik den Zugang zur Musik zu ermöglichen. Dies sollte entdeckendes Suchen, wie es mehrfach in dieser Arbeit thematisiert wurde, ermöglichen. Mit der Umsetzung der „Reinhören“-Funktion wurde eine Möglichkeit dafür im Rahmen der Arbeit getestet. Die Nutzer bewerteten die Funktion als Modus zwar mehrheitlich ablehnend, jedoch zeigte sich eine starke Polarisierung der Meinungen. Eine weitere Verwendung als Funktion konnte deswegen empfohlen werden.
Zusammenfassend konnte zwar beim Einsatz non-verbaler Interaktionsobjekte das angestrebte Ziel der Verbesserung der Usability größtenteils nicht erreicht werden, jedoch wurden in der Arbeit verschiedene Möglichkeiten diskutiert, festgestellte Probleme in zukünftigen Entwicklungen zu umgehen. Weiterhin entstanden Fragestellungen für weiterführende Forschungsarbeiten in diesem Kontext, die im Ausblick diskutiert werden.

9.3 Ausblick

Nachdem in dieser Arbeit die Machbarkeit eines intuitiv bedienbaren sprachgesteuerten MP3-Players gezeigt werden konnte, kann kein Zweifel daran bestehen, dass solche Systeme in den nächsten Jahren nach und nach auf den Markt kommen werden. Mit der immer weiter fortschreitenden Mobilisierung des Musikhörens (siehe Abschnitt 3.3.2) werden jedoch schnell die Ansprüche an solche System steigen. Deswegen scheint es ratsam, die grundsätzlichen Fragestellungen, die im Verlauf dieser Arbeit auftraten, weiterzuverfolgen. Auch ergab sich durch die Beschäftigung mit non-verbalen Interaktionselementen eine Reihe von Ideen für anknüpfende Arbeiten, mit deren Darstellung zunächst begonnen werden soll.

Motiviert durch die schlechte Wahrnehmung der verwendeten symbolischen Auditory Icons im abschließenden Test stellte sich die grundsätzliche Frage, welche Art von Tönen im Auto überhaupt sinnvoll eingesetzt werden können. Während in Literatur zu Warntönen im Auto (beispielsweise [GLPS95]) eine sehr geringe Anzahl von Tönen empfohlen wird, um die Unterscheidbarkeit sicherzustellen, versuchen andere Arbeiten ([SSM+05], [VH05]) pauschale Empfehlungen für die Verwendung von Tönen im Auto zu geben. Nötig wäre jedoch die Identifizierung typischer Anwendungsgebiete für die Einbindung von Tönen und die konkrete Empfehlung von Tönen und Instrumenten (nicht nur von deren Arten), die in diesen Fällen verwendet werden sollten. Fröhlich und Hammer [FH05] versuchen dies, beschränken ihre Analyse jedoch auf eine spezielle Anwendung. Weiterhin sollte diskutiert werden, inwiefern diese Töne mit den Warntönen im Auto interferieren und ob dadurch die Fahrablenkung gesteigert wird.
Solange solche einheitlichen Empfehlungen nicht vorliegen, muss die einfache Erlernbarkeit der Töne verbessert werden. Die Verbindung von dem Erklingen der Töne und dem Hervorheben des entsprechenden grafischen Icons scheint an dieser Stelle erfolgversprechend (siehe Abschnitt 8.3). Jedoch müsste untersucht werden, ob und in welchem Ausmaß die zusätzliche Aktivität im Display Einfluss auf die Fahrablenkung hat. Eine sorgfältige Diskussion der Arten von Hervorhebung für diesen Zweck scheint dabei angebracht.

Zur Musikauswahl musste zunächst festgestellt werden, dass eine Vielzahl von Ideen in dieser Arbeit nicht umgesetzt werden konnten, weil diese umfangreiche Neuentwicklungen von Komponenten bedeutet hätte, für deren Entwicklung im Rahmen der Arbeit die Zeit fehlte. Doch werden demnächst sicher unscharfe Auswahlmöglichkeiten wie Query-By-Humming, automatisiert erstellte Playlisten oder Stimmungsauswahl an Bedeutung gewinnen, ein Versuch der Integration in das bestehende Konzept sollte angedacht werden.
Allerdings muss Musik im Auto nicht auf Musikauswahl durch den Nutzer beschränkt werden, Vorschläge wie die der Arbeit von Rist [Ris04] oder auf der Ideenplattform Halfbakery [Hal06] legen nahe, in absehbarer Zeit auch über situative Erzeugung bzw. automatische Auswahl nachzudenken. So könnte mittels Sensoren die Stimmung bzw. der Zustand des Nutzers oder einzelne Fahrparameter des Autos abgefangen werden, um damit die Auswahl oder Erzeugung von geeigneter Musik zu steuern. Damit würden Autofahren und Musikhören endgültig zu einer untrennbaren Einheit zusammenfinden.

Literaturverzeichnis

[AN02] Akesson, Karl-Petter ; Nilsson, Andreas: Designing Leisure Applications for the Mundane Car-Commute. In: Personal and Ubiquitous Computing 6 (2002), S. 176 – 187

[Ans00] Ansorg, Jürgen. MP3-Grundlagen, Aufbau und Funktion. URL: http://www.fh-jena.de/contrib/fb/et/personal/ansorg/mp3/mp3_2_res.htm. Oktober 2000

[AP03] Aucouturier, Jean-Julien ; Pachet, François: Representing Musical Genre: A State of the Art. In: Journal of New Music Research Vol. 32 (2003), S. 83–93

[ARV05] Alty, James L. ; Rigas, Dimitrios ; Vickers, Paul: Music and speech in auditory interfaces: When ione mode more appropiate than others. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[Bak05] Baker, Janet M.: Milestones in Speech Technology – Past and Future! In: Speech Technology Magazine 10 (2005), September/Oktober, Nr. 5

[Bau00] Baum, Lyman F.: The Wonderful Wizard of Oz. George M. Hill Company, 1900

[BD92] Kap. Introduction: The trend toward multimedia interfaces. In: Blattner, M. ; Dannenberg, R.B.: Multimedia Interface Design. ACM Press, Addison-Wesley, 1992, S. xvii–xxv

[BD96] Button, Graham ; Dourish, Paul: Technomethodology: Paradoxes and Possibilities. In: Proceedings of CHI '96, 1996

[BDD97] Bernsen, Niels O. ; Dybkjear, Hans ; Dybkjear, Laila: Designing Interactive Speech Systems: From First Ideas to User Testing. Springer-Verlag New York, 1997

[BH00] Bussemakers, Myra P. ; de Haan, Ab: When it Sounds like a Duck and it Looks like a Dog... Auditory icons vs. Earcons in Multimedia Environments. In: Proceedings of the International Conference on Auditory Display, 2000

[BK02] Baumann, Stephan ; Klüter, Andreas: Super Convenience for Non-Musicians: Querying MP3 and the Semantic Web. In: Proceedings of the ISMIR 2002, 2002

[Böl97] Bölke, Ludger: Ein akustischer Interaktionsraum für blinde Rechnerbenutzer, Carl von Ossietzky Universität Oldenburg, Diss., 1997

[BN05] Burger, Götz ; Naumann, Sven. WikiLingua. URL: http://www.uni-trier.de/uni/fb2/ldv/ldv_wiki/index.php/Hauptseite. Letzte Änderung: Dezember 2005

[Bre03] Kap. Chapter 12: Nonspeech auditory output In: Brewster, Stephen: Human Factors And Ergonomics. Lawrence Erlbaum Associates, Inc., 2003, S. 220–239

[Bro96] Kap. SUS: A "quick and dirty" usability scale. In: Brooke, John: Usability Evaluation in Industry. Taylor and Francis, 1996, S. 189–194

[Bro06] Brown, Robert: Talking Windows: Exploring New Speech Recognition And Synthesis APIs In Windows Vista. In: MSDN Magazine 21 (2006)

[BSG89] Blattner, M. ; Sumikawa, D. ; Greenberg, R.: Earcons and icons: Their structure and common design principles. In: Human Computer Interaction 4 (1989), Nr. 1, S. 11–44

[Bux89] Buxton, W.: Introduction to this special issue on nonspeech audio. In: Human Computer Interaction 4(1) (1989), S. 1–9

[Cam00] Cameron, Hugh: Speech at the Interface. In: Proceedings of the COST249 Workshop on Speech in Telephone Networks, 2000

[Com01] van Compernolle, Dirk: Recognizing speech of goats, wolves, sheep and ... non-natives. In: Speech Communication 35 (2001), S. 71–79

[Con98] Conversy, Stéphane: Ad-hoc synthesis of auditory icons. In: International Conference on Auditory Display '98, 1998

[DKG02] Kap. Gestaltung einer auditiven Benutzungsoberfläche für Blinde In: Donker, Hilko ; Klante, Palle ; Gorny, Peter: Mensch & Computer 2002: Vom interaktiven Werkzeug zu kooperativen Arbeits- und Lernwelten. B. G. Teubner, 2002, S. 383–392

[EK00] Eysenck, Michael W. ; Keane, Mark T.: Cognitive Psychology: A Student's Handbook (4th Edition). Taylor & Francis Group, 2000

[FH05] Fröhlich, Peter ; Hammer, Florian: A user-centred approach to sound design in voice-enabled mobile applications. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[fre06] freedb.org. freedb FAQ. URL: http://www.freedb.org/modules.php?name=Sections\&sop=viewarticle\&artid=26. Letzter Zugriff: Januar 2006

[FSNR+05] Forlines, Clifton ; Schmidt-Nielsen, Bent ; Raj, Bhiksha ; Wittenburg, Kent ; Wolf, Peter: A Comparison Between Spoken Queries and Menu-based Interfaces for In-Car Digital Music Selection. In: IFIP TC13 International Conference on Human-Computer Interaction, 2005

[Gav94] Kap. Using and Creating Auditory Icons In: Gaver, William W.: Auditory Display: Sonification, Audification, and Auditory Interfaces. Addison-Wesley Publishing Company, 1994, S. 417– 446

[GL85] Gould, John D. ; Lewis, Clayton: Designing for Usability: Key Principles and What Designers Think. In: Communications of the ACM 28 (1985), Nr. 3, S. 360–411

[GLCS95] Ghias, Asif ; Logan, Jonathan ; Chamberlin, David ; Smith, Brian C.: Query by humming: musical information retrieval in an audio database. In: MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia, 1995. – ISBN 0–89791–751–0, S. 231–236

[GLPS95] Green, Paul ; Levison, William ; Paelke, Gretchen ; Serafin, Colleen: Preliminary human factors design guidelines for driver information systems / The University of Michigan. 1995. – Forschungsbericht

[GMN04] Gruhn, Rainer ; Markov, Konstantin ; Nakamura, Satoshi: A Statistical Lexicon for Non-Native Speech Recognition. In: Proceedings ICSLP, 2004

[Goo06] Google. Google Webseite. URL: http://www.google.com/. Letzter Zugriff: Januar 2006

[Gär02] Gärdenfors, Dan: Designing Sound-Based Computer Games. In: Proceedings of cybersonica symposium, 2002

[Gra06] Gracenote. What is CDDB®? URL: http://www.gracenote.com/music/corporate/FAQs.html/faqset=what/page=1. Letzter Zugriff: Januar 2006

[Gri06] Griffiths, Richard. Usability Evaluation by Query Techniques. URL: http://www.it.bton.ac.uk/staff/rng/teaching/notes/UsabilityEvalQuery.html. Letzter Zugriff: Januar 2006

[Hal06] Halfbakery.com. halfbakery Webseite. URL: http://www.halfbakery.com/. Letzter Zugriff: Januar 2006

[Ham00] Hamerich, Stefan W.: Strategien für Dialogsegmente in natürlichsprachlichen Anwendungen, Universität Hamburg, Diplomarbeit, 2000

[Ham03] Hamerich, Stefan W.: Gegenüberstellung von VoiceXML und GDML / Temic Speech Dialog Systems. 2003. – Internes Dokument

[Han04] Hanrieder, Gerhardt: Sprachbedienung im KFZ - Eine Erfolgsgeschichte. In: Tagungsband der 34. Jahrestagung der Gesellschaft für Informatik e.V. – 'Informatik 2004', 2004, S. 220–224

[Hei01] Heisterkamp, Paul: Linguatronic - Product-Level Speech System for Mercedes-Benz Cars. In: Proceedings of HLT, 2001

[HF04] Herrmann, Christoph ; Fiebach, Christian: Gehirn & Sprache. Fischer Taschenbuch Verlag, 2004

[HH04] Hamerich, Stefan W. ; Hanrieder, Gerhard: Modelling Generic Dialog Applications for Embedded Systems. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), 2004

[HHLC05] Hamer-Hodges, Dan ; Li, Simon Y. ; Cairns, Paul: Cueing hyperlinks in auditory interfaces. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[HU04] Haegler, Simon ; Ulmer, Andreas: Sprachsteuerung eines Audiogerätes / Eidgenössische Technische Hochschule Zürich. 2004. – Forschungsbericht

[IP02] Infineon-Presseinformationen. Infineon stellt Basis-Technologien für „intelligente“ Kleidung vor. URL: http://www.interactive-wear.de/cms/front_content.php?idcat=49\&idart=52. Letzte Änderung: April 2002

[ISO96] ISO: ISO 9241-10: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 10: Dialogue principles / International Organization for Standardization. 1996. – Standard

[ISO98] ISO: ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability / International Organization for Standardization. 1998. – Standard

[ISO04] ISO: ISO 15006: Road vehicles — Ergonomic aspects of transport information and control systems — Specifications and compliance procedures for in-vehicle auditory presentation / International Standards Organization. 2004. – Standard. First edition

[JM00] Jurafsky, Daniel ; Martin, James H. ; Horton, Marcia (Hrsg.): Speech and Language Processing: An Introduction to Natural Language Processing. Prentice Hall, 2000

[Jon06] de Jong, Bram. The Freesound Project. URL: http://freesound.iua.upf.edu/. Letzte Zugriff: Januar 2006

[Jou01] Jourdain, Robert ; Wigger, Frank (Hrsg.) ; Peters-Hofmann, Marlis (Hrsg.) ; Alton, Bianca (Hrsg.): Das wohltemperierte Gehirn. Wie Musik im Kopf entsteht und wirkt. Spektrum Akademischer Verlag, Heidelberg, 2001

[JSF05] Johannes, Jendrik ; Schäfer, Ullrich ; Felbecker, Tobias: Audio gesteuerter DVDPlayer fürs interaktive Kino / TU-Dresden. 2005. – Forschungsbericht

[Kir00] Kirakowski, Jurek. Questionnaires in Usability Engineering - A List of Frequently Asked Questions (3rd Ed.). URL: http://www.ucc.ie/hfrg/resources/qfaq1.html. Letzte Änderung: Juni 2000

[KKBG+05] Kruijff-Korbayová, Ivana ; Blaylock, Nate ; Gerstenberger, Ciprian ; Rieser, Verena ; Becker, Tilman ; Kaisser, Michael ; Poller, Peter ; Schehl, Jan: An Experiment Setup for Collecting Data for Adaptive Output Planning in a Multimodal Dialogue System. In: Proceedings of the 10th European Workshop on Natural Language Generation, 2005

[Kla03a] Kap. Praxisbericht zur Gestaltung auditiver Benutzungsoberflächen In: Klante, Palle: Proceedings of the 1st annual GC-UPA Track, Stuttgart. German Chapter der UPA e.V., 2003, S. 57–62

[Kla03b] Klante, Palle: Werkzeuggestütze Entwicklung auditiver Benutzungsoberflächen. In: Informatiktage 2002, Fachwissenschaftlicher Kongress, 2003

[KSC+00] Klemmer, Scott R. ; Sinha, Anoop K. ; Chen, Jack ; Landay, James A. ; Aboobaker, Nadeem ; Wang, Annie: Suede: a Wizard of Oz prototyping tool for speech user interfaces. In: UIST '00: Proceedings of the 13th annual ACM symposium on User interface software and technology, 2000, S. 1–10

[LAB+05] Larson, James A. ; Applebaum, Ted ; Byrne, Bill ; Cohen, Michael ; Giangola, James ; Gilbert, Juan E. ; Green, Rebecca N. ; Hebner, Thomas ; Houwing, Tom ; Hura, Susan ; Issar, Sunil ; Kaiser, Lizanne ; Kaushansky, Karen ; Kilgore, Robby ; Lai, Jennifer ; Leppik, David ; Mailey, Stephen ; Margulies, Ed ; McArtor, Kristen ; McTear, Michael ; Sachs, Richard: Ten Guidelines for Designing a Successful Voice User Interface. In: Speech Technology Magazine 9 (2005), Januar/Februar, Nr. 7

[Lar03] Larsen, Lars B.: On the Usability of Spoken Dialogue Systems, Aalborg University, Diss., 2003

[Las06] Last.fm. last.fm. URL: http://www.last.fm/. Letzter Zugriff: Januar 2006

[MAR+01] McGlaun, Gregor ; Althoff, Frank ; Rühl, Hans-Wilhelm ; Alger, Michael ; Lang, Manfred: A Generic Operation Concept for an Ergonomic Speech MMI under Fixed Constraints in the Automotive Environment. In: HCI 2001, 9.th Int. Conference on Human Computer Interaction, 2001

[Mar04] Martínez, José M.: ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio / International Standards Organization. 2004. – Forschungsbericht

[Mat03] Mattes, Stefan: The Lane Change Task as a Tool for Driver Distraction Evaluation. In: Proceedings of the Conference of the International Society for Occupational Ergonomics and Safety (ISOES), 2003

[May99] Mayhew, Deborah J.: The usability engineering lifecycle - a practitioner's handbook for user interface design. Morgan Kaufmann, 1999

[May04] Mayhew, Deborah J. How - The Usability Engineering Lifecycle. URL: http://www.deborahjmayhew.com/index.php?loc=11\&nloc=1. Letzte Änderung: 2004

[MBC+04] McGlashan, Scott ; Burnett, Daniel C. ; Carter, Jerry ; Danielsen, Peter ; Ferrans, Jim ; Hunt, Andrew ; Lucas, Bruce ; Porter, Brad ; Rehor, Ken ; Tryphonas, Steph: W3C Voice Extensible Markup Language (VoiceXML) Version 2.0 - URL http://www.w3.org/TR/voicexml20/ / W3C. 2004. – Standard

[MBW+98] Mynatt, E. D. ; Back, M. ; Want, R. ; Baer, M. ; ELLIS, J. B.: Designing audio aura. In: Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI '98, 1998

[McK05] McKeown, Denis: Candidates for within-vehicle auditory displays. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[McT02] McTear, Michael F.: Spoken dialogue technology: enabling the conversational user interface. In: ACM Comput. Surv. 34 (2002)

[McT04] McTear, Michael F. ; Pittermann, Johannes (Hrsg.): Spoken dialogue technology: toward the conversational user interface. Springer-Verlag London Limited, 2004

[Med06] MediaInterfaceDresden. SpeaKING Control Webseite. URL: http://www.mediainterface.de/index.php?module=control1\&menu_module=control. Letzter Zugriff: Januar 2006

[Mic81] Michels, Ulrich: dtv-Atlas zur Musik, Bd. 1. Systematischer Teil. Historischer Teil: Von den Anfängen bis zur Renaissance. Deutscher Taschenbuch-Verlag, München, 1981

[Mic06] Microsoft. Voice Command for Pocket PC and Pocket PC Phone Edition. URL: http://www.microsoft.com/germany/windowsmobile/vm_content.mspx. Letzter Zugriff: Januar 2006

[Mil56] Miller, George A.: The Magical Number 7, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. In: Psychological Review 63 (1956), S. 81–97

[mus06a] Webseite Hochschule für Musik „Carl Maria von Weber" Dresden. URL: http://www.hfmdd.de/. Letzter Zugriff: Januar 2006

[mus06b] musicline.de. Query By Humming Melodiesuche. URL: http://www.musicline.de/de/melodiesuche/. Letzter Zugriff: Januar 2006

[Nat90] Nattiez, Jean-Jacques: Music and Discourse: Toward a Semiology of Music. Princeton University Press, 1990

[Nie89] Kap. Usability Engineering at a Discount In: Nielsen, Jakob: Designing and Using Human-Computer Interfaces and Knowledge-Based Systems. Elsevier Science, 1989, S. 389–401

[Nie93] Nielsen, Jakob: Usability Engineering. Academic Press Limited, 1993

[Nil00] Nilsson, Martin: ID3 tag version 2.4.0 - Main Structure. 2000. – Informal Standard

[OP02] Olsen, Dan R. ; Peachey, Jon R.: Query-by-critique: spoken language access to large lists. In: UIST '02: Proceedings of the 15th annual ACM symposium on User interface software and technology, 2002, S. 131–140

[Opp92] Oppenheim, A. N.: Questionnaire Design, Interviewing and Attitude Measurement (New Edition). Pinter Publishers Ltd., 1992

[Ovi95] Oviatt, Sharon: Predicting Spoken Disfluencies during Human-Computer Interaction. In: Computer Speech Language 9 (1995), Nr. 1, S. 19–36

[PBZA04] Pachet, Francois ; Burthe, Amaury L. ; Zils, Aymeric ; Aucouturier, Jean-Julien: Popular music access: the Sony music browser. In: J. Am. Soc. Inf. Sci. Technol. 55 (2004), Nr. 12, S. 1037–1044

[PDB+03] Pieraccini, Roberto ; Dayanidhi, Krishna ; Bloom, Jonathan ; Dahan, Jean-Gui ; Phillips, Michael ; R.Goodman, Bryan ; Prasad, K. V.: A Multimodal Conversational Interface for a Concept Vehicle. In: Proceedings of Eurospeech 2003, 2003, S. 2233–2236

[Pet04] Petrik, Stefan: Wizard of Oz Experiments on Speech Dialogue Systems.Design and Realisation with a New Integrated Simulation Environment, Technische Universität Graz, Diplomarbeit, 2004

[PH05] Pieraccini, Roberto ; Huerta, Juan: Where do we go from here? Research and commercial spoken dialog systems. In: Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue, 2005

[PM06] Pandora Media, Inc. Discover Music - Pandora. URL: http://www.pandora.com/. Letzter Zugriff: Januar 2006

[PR06] Philips-Research. Easy Access. URL: http://www.research.philips.com/technologies/syst_softw/easyaccess/. Letzter Zugriff: Januar 2006

[Pre99] Preim, Bernhard ; Strothotte, Thomas (Hrsg.): Entwicklung interaktiver Systeme. Springer-Verlag, 1999

[Pre04] O_2Germany Presseabteilung. Music bei O_2. URL: http://de.o2.com/ext/common/download?file_id=792\&state=online\&style=standard\&link_id=24110. März 2004

[Pre06] connect.de Presseinformationen. connectFachkongress „music meets mobile": Handynutzer wünschen sich integrierte MP3-Player. URL: http://www.connect.de/presseinformationen/_connect_fachkongress_music_meets_mobile_handynutzer_wuenschen_sich_integrierte_mp3_player.59761.htm. Letzter Zugriff: Januar 2006

[Res06a] Research, DSS. Questionnaire Design. URL: http://www.dssresearch.com/toolkit/resource/papers/QD03.asp. Letzter Zugriff: Januar 2006

[Res06b] Research, DSS. Tips for Writing a Good Questionnaire. URL: http://www.dssresearch.com/toolkit/resource/papers/QD01.asp. Letzter Zugriff: Januar 2006

[Ris04] Rist, Thomas: Affekt und physiologische Verfassung als Parameter für hochgradig personalisierte Fahrerassistenzdienste. In: Proceedings Workshop Automobile Cockpits und HMI, 2004

[RLL+04] Reeves, Leah M. ; Lai, Jennifer ; Larson, James A. ; Oviatt, Sharon ; Balaji, T. S. ; Buisine, Stephanie ; Collings, Penny ; Cohen, Phil ; Kraal, Ben ; Martin, Jean-Claude ; McTear, Michael ; Raman, TV ; Stanney, Kay M. ; Su, Hui ; Wang, Qian Y.: Guidelines for multimodal user interface design. In: Communications of the ACM 47 (2004), Nr. 1, S. 57–59

[Rol05] Rolandi, Walter: Some Things Are Better Left Unsaid. In: Speech Technology Magazine 10 (2005), September/Oktober, Nr. 5

[ROR01] Rosenfeld, Ronald ; Olsen, Dan ; Rudnicky, Alex: Universal speech interfaces. In: interactions 8 (2001)

[SC93] Salber, Daniel ; Coutaz, Joelle: Applying the Wizard of Oz Technique to the Study of Multimodal Systems. In: EWHCI '93: Selected papers from the Third International Conference on Human-Computer Interaction, 1993, S. 219–230

[Sch04] Schulz, Stefan: Hyperaudio Browser, TU Dresden, Belegarbeit, 2004

[SRSR01] Stutts, Jane C. ; Reinfurt, Donald W. ; Staplin, Loren ; Rodgman, Eric A.: The Role of Driver Distraction in Traffic Crashes. AAA Foundation for Traffic Safety, 2001

[SS00] Sawhney, Nitin ; Schmandt, Chris: Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. In: ACM Transactions on Computer-Human Interaction 7 (2000), Nr. 3, S. 353–383

[SSM+05] Suied, Clara ; Susini, Patrick ; Misdariis, Nicolas ; Langlois, Sabine ; Smith, Bennett K. ; McAdams, Stephen: Toward a sound design methodology: Application to electronic automotive sounds. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[Suh03] Suhm, Bernhard: Towards Best Practices for Speech User Interface Design. In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), 2003, S. 2217–2220

[Tim97] Timm, Friedrich ; Helmin, Christina (Hrsg.): Das moderne Fremdwörterlexikon. Naumann&Göbel, 1997

[Usa03] UsabilityNet. Questionnaire resources. URL: http://www.hostserver150.com/usabilit/tools/r_questionnaire.htm. Letzte Änderung: 2003

[Ven96] Venkatesh, Alladi: Computers and Other Interactive Technologies for the Home. In: Communications of the ACM 39 (1996), S. 47–54

[VGD+05] Voida, Amy ; Grinter, Rebecca E. ; Ducheneaut, Nicolas ; Edwards, W. K. ; Newman, Mark W.: Listening in: practices surrounding iTunes music sharing. In: CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, 2005, S. 191–200

[VH05] Vilimek, Robert ; Hempel, Thomas: Effects of Speech and Non-Speech-Sounds on Short Memory and Possible Implications for In-Vehicle Use. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display, 2005

[VK05] Vodafone-KundenBetreuung. Vodafone Info-Fax Nummer 600: Vodafone MusicFinder - 22 11 22. URL: http://www.vodafone.de/infofaxe/600.pdf. Juni 2005

[Wan03] Wang, Kuansan: A study of semantics synchronous understanding for speech interface design. In: Proceedings of ACM Symposium UIST-2003, 2003

[WHHS05] Wang, Yu-Fang H. ; Hamerich, Stefan W. ; Hennecke, Marcus E. ; Schubert, Volker M.: Speech-controlled Media File Selection on Embedded Systems. In: Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue, 2005

[Wik05a] Wikipedia, Die freie E. Artikel Autoradio. URL: http://de.wikipedia.org/w/index.php?title=Autoradio\&oldid=10566189. Letzter Zugriff: Dezember 2005

[Wik05b] Wikipedia, Die freie E. Artikel Musik. URL: http://de.wikipedia.org/w/index.php?title=Musik\&oldid=11629906. Letzter Zugriff: Dezember 2005

[Wil03] Wilson, Chauncey: Methods and Guidelines to Avoid Common Questionnaire Bloopers. In: Usability Interface 9 (2003), Nr. 3

[Win06] Winamp.com. Product Walkthrough. URL: http://www.winamp.com/player/walkthrough.php. Letzter Zugriff: Januar 2006

[Wir06] Wirth, Thomas. KommDesign.de - Texte - Usability (1) - Die EN ISO 9241 - 10. URL: http://www.kommdesign.de/texte/din.htm. Letzter Zugriff: Januar 2006

[Wit03] Witte, Marc: Weiterentwicklung einer Bibliothek von Interaktionsobjekten für auditive Benutzungsoberflächen und ihre Evaluation am Anwendungsbeispiel MP3 Player, Carl von Ossietzky Universität Oldenburg, Diplomarbeit, Juni 2003

[WRS04] White, Kenneth ; Ruback, Harvey ; Sicconi, Roberto: Is There a Future for Speech in Vehicles? In: Speech Technology Magazine 9 (2004), November/Dezember, Nr. 6

[WW04] Williams, Jason D. ; Witt, Silke M.: A Comparison of Dialog Strategies for Call Routing. In: International Journal of Speech Technology 7 (2004), Nr. 1, S. 9 – 24

[Yah06] YAHOO! Flickr Webseite. URL: http://www.flickr.com. Letzter Zugriff: Januar 2006

[Zag95] ZAGLER, Wolfgang L.: Overview About Typical Scenarios of Speech Technology Applications for Elderly and Disabled Persons. In: Proceedings of the COST 219 - Seminar Speech Technology Applications for Disabled and Elderly People, 1995

[ZPDG02] ZILS, A. ; PACHET, F. ; DELERUE, O. ; GOUYON, F.: Automatic Extraction of Drum Tracks from Polyphonic Music Signals. In: Proceedings of WEDEL MUSIC, 2002

A
CD-ROM-Inhalt

Auf der beiliegenden CD befindet neben den für die Nutzertests verwendeten Materialien (inklusive dem im WOZ-Test benutzten Powerpoint-Folien) auch der in der Abschlussevaluation verwendete Prototyp mit dem Namen „Dorothy“. Weiterhin finden sich auf der CD alle im Rahmen der Arbeit benutzten Töne.

Dabei finden sich folgende Ordner auf der CD:

`\_Diplom`	Diese Arbeit.
`\Dorothy`	Prototyp „Dorothy“.
`\Materialien`	Materialien für die Nutzertests.
`\Toene`	Im Rahmen der Arbeit benutzte Töne.

Im Diplom-Verzeichnis findet sich lediglich diese Arbeit als PDF-Dokument, jeweils als öffentliche und nicht-öffentliche Version. Erläuterungen zu den weiteren Verzeichnissen finden sich in den nachfolgenden Abschnitten.

A.1 Dorothy

In diesem Verzeichnis befindet sich der Prototyp „Dorothy“, an dem die abschließende Evaluation (Kapitel 8) durchgeführt wurde. Dieser Prototyp darf bis zum 21.12.2006 an der Fakultät Informatik der TU Dresden benutzt werden. Er wurde unter Windows 2000 und XP auf Lauffähigkeit getestet. Bei der Benutzung von Windows XP bitte die weiteren Hinweise beachten.
Zum Starten wird der Inhalt des Ordners einfach in ein beliebiges Verzeichnis auf der Festplatte kopiert. Danach muss das Verzeichnis `C:\Musik\` mit den MP3s gefüllt werden, die per Sprache auswählbar sein sollen. Gestartet wird der Prototyp dann in dem Unterverzeichnis `Dorothy\Dorothy` mit einem dieser vier Kommandos:

- `start_normal`
- `start_ablesen`
- `start_vorlesen`
- `start_reinhören`

Diese drücken den beim Start voreingestellten Listen-Modus aus (siehe Kapitel 7).
Bei der Benutzung von Windows XP ist zu beachten, dass dort das in der Batchdatei eingesetzte kill-Kommando nicht funktioniert. Sollen die Fehlermeldungen umgangen werden, bietet es sich an, die die Datei `startdds_xp.bat` in `startdds.bat` umzubenennen, das geht zum Beispiel in der cmd-Kommandozeile von Windows mit `move /y startdds_xp.bat startdds.bat`. Danach wird das kill-Kommando nicht mehr aufgerufen, die Programme müssen dann aber per Hand beendet werden.[1]

A.2 Materialien

In diesem Verzeichnis befinden sich die in den jeweiligen Unterverzeichnissen die Vorarbeiten, Testdokumente und Ergebnisse zum Fragen, WOZ-Test und der Evaluation von „Dorothy“. Unter anderem findet sich unter `WOZ\PP` die im WOZ-Test benutzte Powerpoint-Präsentation. Ebenfalls befinden sich in den jeweiligen Unterverzeichnissen von WOZ und Evaluation die Excel-Dateien, die sämtliche Auswertungsdaten zu den Versuchen enthalten.

A.3 Töne

In diesem Verzeichnis finden sich die im Versuch verwendeten Töne sowie die von einem Studenten der Hochschule für Musik „Carl Maria von Weber“ Dresden (HFMDD) [mus06a] erstellte Töne, die leider nicht mehr rechtzeitig eintrafen, um berücksichtigt zu werden.

`\Dorothy` Im Prototyp „Dorothy“ verwendete Töne.

`\HFMDD` Für Prototyp „Dorothy“ erstellte Töne von Studenten der HFMDD

[1] Es ist auch möglich das für Windows XP verfügbare `pskill` als Ersatz für kill einzusetzen. Diese Einbindung wird jedoch bei gelegentlicher Nutzung kaum nötig sein.

B
Dokumente

In diesem Anhang finden sich Dokumente, die im Verlauf der Arbeit angefertigt wurden. In dieser öffentlichen Version der Arbeit fehlt ein Großteil der Dokumente dieses Anhangs, da sie der Geheimhaltung der Firma Harman/Becker Automotive Systems unterliegen. Wenn Dokumente fehlen, wird dies in den Unterkapiteln kenntlich gemacht.

B.1 Untersuchungsmaterialien

B.1.1 Fragebogen

Fragebogen

Fragebogen MP3-Player

Temic SDS GmbH

Speech Dialog Systems

Söflingerstraße 100

D-89077 Ulm, Germany

In Kooperation mit der

Technische Universität Dresden

Fakultät Informatik

Institut für Software- und Multimediatechnik

Dozentur Kooperative multimediale Anwendungen

D-01062 Dresden, Germany

Temic-SDS GmbH, Stefan Schulz, Söflinger Str. 100, 89077 Ulm
stefan.schulz@temic-sds.com

I. Allgemeine Daten

Alter o 18-21 o 22-25 o 25-30 o 30-40 o >40

Geschlecht o männlich o weiblich

Studiengang/Beruf .....

Mein Interesse für technische Dinge ist

sehr groß eher groß mittel eher gering sehr gering
o----------o----------o----------o----------o

II. MP3

Wie häufig hören sie MP3s?

mehrfach täglich (fast) täglich mehrmals wöchentlich mehrmals monatlich seltener
o----------o----------o----------o----------o

Wie hören Sie häufiger MP3s? o am Computer o mit mobilen MP3-Playern

Falls am Computer:
Welches Programm verwenden Sie hauptsächlich zum Abspielen ihrer MP3s?
o WinAmp o Windows Media Player o RealPlayer
o iTunes o Musicmatch Jukebox o Weiß nicht.
o Sonstiges: .....

Falls mit mobilen MP3-Playern:
Welche Art von mobilem MP3-Player besitzen Sie? (Mehrfachnennungen möglich)
o Festplatten-basierte Geräte:
o iPod o Creative Zen o Rio o iRiver H-Serie
o Sonstiges: .....

o Flash-Speicher basierte Geräte:
o iPod shuffle o iRiver i-Serie o USB-Stick o MP3-Handy
o Sonstiges: .....

o Autoradio mit MP3-Funktion:
o Becker o Pioneer o JVC o Sony o Blaupunkt
o Sonstiges: .....

o CD-basierte Geräte:
o MP3-CD-Player
o Sonstiges: .....

Wie hören Sie im Auto überwiegend Musik?
o Radio o Kassette o Audio-CD o MP3-Medien

Temic-SDS GmbH, Stefan Schulz, Söflinger Str. 100, 89077 Ulm
stefan.schulz@temic-sds.com

Zu welchen Gelegenheiten hören Sie Musik im Auto? (Mehrfachnennungen möglich)

o Immer	o Weg von/zur Uni/Arbeit	o Freizeit
o Im Stau	o Wenn alleine unterwegs	o Auf langen Strecken

o Sonstiges: .....

Kommt es manchmal vor, dass Sie Ihren Beifahrer bitten, die Musikanlage im Auto zu bedienen? Falls ja, worum bitten Sie ihn? (außer Radiofunktionen)

.....

.....

.....

Nach welchen Kriterien ist ihre MP3 Sammlung organisiert? (Mehrfachnennungen möglich)

o Albumname	o Künstlername	o Titelname	o persönliche Vorlieben
o Musikstil/Genre	o Erscheinungsjahr	o Stimmung	o Autobiografie

o Art der Tracks (Musik, Comedy, Hörbücher...)

o Sonstiges: .....

Wenn Sie Playlisten benutzen, wie viele persönliche Playlisten verwalten Sie dann?

o 1	o 2-4	o 5-10	o >10

Nach welchen Kriterien stellen Sie ihre Playlisten zusammen?

.....

.....

.....

Wie groß ist ungefähr ihre MP3-Sammlung?

..... Titel ..... Gigabyte

III. Sprachbedienung im Auto

Stellen Sie sich vor, es gäbe einen MP3-Player im Auto, der auch mittels Sprache bedient werden kann und Sie so versteht, wie ein Beifahrer Sie verstehen würde.

Unabhängig von der technischen Machbarkeit welche Funktion würden Sie am liebsten per Sprache steuern?

.....

.....

Welche zusätzliche Funktionalität wäre besonders cool?

.....

.....

Temic-SDS GmbH, Stefan Schulz, Söflinger Str. 100, 89077 Ulm
stefan.schulz@temic-sds.com

In dem folgenden Abschnitt werden Ihnen Funktionen vorgestellt, die ein solcher sprachgesteuerter MP3-Player besitzen könnte. Bitte bewerten Sie, wie wichtig folgende Funktionen für Sie wären:

Aktion	**Beispieläußerung**	völlig unwichtig / eher unwichtig / weiß nicht / eher wichtig / sehr wichtig
Auswahl Album	„Spiele Madonna – The Immaculate Collection“	o-------o-------o-------o-------o
Auswahl Titel	„Ich möchte Männer von Grönemeyer hören.“	o-------o-------o-------o-------o
Auswahl des Künstlers	„Spiele etwas von Madonna.“	o-------o-------o-------o-------o
Auswahl Genre	„Ich will etwas Heavy Metal hören!“	o-------o-------o-------o-------o
Zufälligen Titel spielen	„Spiele irgendwas!“	o-------o-------o-------o-------o
Abspielen von Playlisten	„Meine TopTen, bitte.“	o-------o-------o-------o-------o
Speichern als Playliste	„Aktuelle Auswahl als ‚Meine Lieblingshits’ speichern“	o-------o-------o-------o-------o
Speichern in vorhandene Playliste	„Füge dies zur Playliste ‚Meine Lieblingshits’ hinzu“	o-------o-------o-------o-------o
Steuerkommandos	„Nächster Titel!” „Stopp.“ „Pause.“	o-------o-------o-------o-------o
Wiederhole Einzeltitel	„Wiederhole dieses Lied”	o-------o-------o-------o-------o
Wiederhole alles	„Wiederhole alles!“	o-------o-------o-------o-------o
Shuffle/Random	„Shuffle”	o-------o-------o-------o-------o
Inform. Zum aktuellen Titel	„Was ist das denn?”	o-------o-------o-------o-------o
Inform. zum gesamten Bestand	„Welche Genres sind vorhanden?“	o-------o-------o-------o-------o
Ähnlichen Titel spielen	„Ein ähnliches Lied, bitte“	o-------o-------o-------o-------o
Auswahl durch Summen Melodie	„Ich will <Titel ansummen>“	o-------o-------o-------o-------o
Sonstiges 1:	.....	o-------o-------o-------o-------o
Sonstiges 2:	.....	o-------o-------o-------o-------o

Stellen Sie sich nun vor, Sie haben gesagt: „Ich möchte Männer von Grönemeyer hören“, und Sie hören nun diesen Titel. Was sollte ihrer Meinung nach passieren, wenn dieser Titel zu Ende gespielt ist? Das System soll...

o keinen weiteren Titel spielen und warten, bis ich eine neue Auswahl treffe.
o mich zu einer neuer Auswahl auffordern.
o den nächsten Titel des Albums, das den Titel enthält, abspielen.
o den nächsten Titel aus der gesamten Titelliste abspielen.
o den nächsten Titel im Verzeichnis abspielen.
o einen möglichst ähnlichen Titel abspielen.
o irgendeinen anderen Titel abspielen.
o Sonstiges:.....

Nachdem Sie nun einen ersten Einblick in die Möglichkeiten eines sprachgesteuerten MP3-Players bekommen haben: Wären Sie bereit, für die zusätzliche Funktionalität der Sprachsteuerung eines MP3-Players im Auto einen Aufpreis zu zahlen?
o Ja o Nein o Weiß nicht.

Bitte überprüfen Sie noch einmal, ob Sie auch alle Felder bearbeitet haben. Danke!

Temic-SDS GmbH, Stefan Schulz, Söflinger Str. 100, 89077 Ulm
stefan.schulz@temic-sds.com

B.1.2 WOZ

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nur teilweise zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

Vorbefragung (WOZ)

:: Vorbefragung ::

I. Allgemeine Daten

Alter o 18-21 o 22-25 o 26-30 o 31-40 o >40

Geschlecht o männlich o weiblich

Studiengang/Beruf .....

Mein Interesse für technische Dinge ist

sehr groß eher groß mittel eher gering sehr gering
o----------o----------o----------o----------o

II. MP3

Wie häufig hören sie MP3s?

mehrfach täglich (fast) täglich mehrmals wöchentlich mehrmals monatlich seltener
o----------o----------o----------o----------o

Wie hören Sie häufiger MP3s? o am Computer o mit mobilen MP3-Playern

Falls am Computer:
Welches Programm verwenden Sie hauptsächlich zum Abspielen ihrer MP3s?
o WinAmp o Windows Media Player o RealPlayer
o iTunes o Musicmatch Jukebox o Weiß nicht.
o Sonstiges: .....

Falls mit mobilen MP3-Playern:
Welche Art von mobilem MP3-Player besitzen Sie? (Mehrfachnennungen möglich)
o Festplatten-basierte Geräte:
o iPod o Creative Zen o Rio o iRiver H-Serie
o Sonstiges: .....

o Flash-Speicher basierte Geräte:
o iPod shuffle o iRiver i-Serie o USB-Stick o MP3-Handy
o Sonstiges: .....

o Autoradio mit MP3-Funktion:
o Becker o Pioneer o JVC o Sony o Blaupunkt
o Sonstiges: .....

o CD-basierte Geräte:
o MP3-CD-Player
o Sonstiges: .....

Wie hören Sie im Auto überwiegend Musik?
o Radio o Kassette o Audio-CD o MP3-Medien

Zu welchen Gelegenheiten hören Sie Musik im Auto? (Mehrfachnennungen möglich)

o Immer	o Weg von/zur Uni/Arbeit	o Freizeit
o Im Stau	o Wenn alleine unterwegs	o Auf langen Strecken

o Sonstiges: .....

Kommt es manchmal vor, dass Sie Ihren Beifahrer bitten, die Musikanlage im Auto zu bedienen? Falls ja, worum bitten Sie ihn? (außer Radiofunktionen)

.....

.....

.....

Nach welchen Kriterien ist ihre MP3 Sammlung organisiert? (Mehrfachnennungen möglich)

o Albumname	o Künstlername	o Titelname	o persönliche Vorlieben
o Musikstil/Genre	o Erscheinungsjahr	o Stimmung	o Autobiografie

o Art der Tracks (Musik, Comedy, Hörbücher...)

o Sonstiges: .....

Wenn Sie Playlisten benutzen, wie viele persönliche Playlisten verwalten Sie dann?

o 1 o 2-4 o 5-10 o >10

Nach welchen Kriterien stellen Sie ihre Playlisten zusammen?

.....

.....

.....

Wie groß ist ungefähr ihre MP3-Sammlung?

..... Titel ..... Gigabyte

III. Vorerfahrung Sprachdialogsysteme

Haben Sie bereits Erfahrung mit Sprachdialogsystemen?

o Ja o Nein o Weiß nicht

Wenn ja, mit welchen (z.B. telefonische Bahnauskunft, Postbank, Infotainmentgeräte im Auto)?

.....

.....

.....

schriftliche Nachbefragung (WOZ)

:: Teil 2b – Befragung zu Teil 1 ::

Sie haben nun grundlegende Funktionen des Systems kennengelernt. Wir möchten Sie nun bitten, uns einige Fragen zu beantworten.

Bitte zeichnen Sie in jeder Zeile einen Kreis um den zutreffenden Wert. Wenn Sie sich nicht sicher sind, markieren Sie die „3“. Bitte denken Sie nicht lange über eine Antwort nach, sondern antworten Sie ganz spontan.

	lehne stark ab	lehne eher ab	weiß nicht	stimme eher zu	stimme sehr zu
Ich glaube, ich würde dieses System gerne häufig benutzen	1	2	3	4	5
Ich fand das System unnötig komplex	1	2	3	4	5
Ich fand, das System war einfach zu benutzen	1	2	3	4	5
Ich glaube, dass ich die Unterstützung eines technisch versierten Menschen bräuchte, um dieses System benutzen zu können	1	2	3	4	5
Ich fand, dass die verschiedenen Funktionen in dieses System gut integriert wurden	1	2	3	4	5
Mir kam es so vor, als wäre das System sehr inkonsistent	1	2	3	4	5
Ich könnte mir vorstellen, dass die meisten Menschen innerhalb kurzer Zeit lernen würden, dieses System zu benutzen	1	2	3	4	5
Ich fand, dass das System sehr umständlich zu benutzen war	1	2	3	4	5
Ich hatte das Gefühl, das System im Griff zu haben	1	2	3	4	5
Ich musste eine Menge lernen, bevor ich mit diesem System loslegen konnte	1	2	3	4	5

A) Fühlten Sie Sich vom Fahren abgelenkt, während Sie die Spracheingabe benutzten?	nein (1) ja, etwas (2) ja, ziemlich (3) ja, stark (4)
- Falls ja: Was hat Sie abgelenkt (mehrere Antworten möglich)?	o sich an das Sprachkommando zu erinnern o auf eine Reaktion des Systems zu warten o auf das Display zu schauen o sich im Menü zurechtzufinden o der Sprachausgabe zuzuhören o Sonstiges: _____
B) Welche Schulnote würden Sie dem Sprachbediensystem geben?	o 1 (sehr gut) o 2 (gut) o 3 (befriedigend) o 4 (ausreichend) o 5 (mangelhaft) o 6 (ungenügend)

B.1.3 Fragebogen & WOZ Ergebnisse

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nicht zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

B.1.4 Abschlussevaluation

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nur teilweise zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

Vorbefragung (Evaluation Dorothy)

:: Vorbefragung ::

I. Allgemeine Daten

Alter o 18-21 o 22-25 o 26-30 o 31-40 o >40

Geschlecht o männlich o weiblich

Studiengang/Beruf .....

Mein Interesse für technische Dinge ist

sehr groß eher groß mittel eher gering sehr gering
o----------o----------o----------o----------o

II. MP3

Wie häufig hören sie MP3s?

mehrfach täglich (fast) täglich mehrmals wöchentlich mehrmals monatlich seltener
o----------o----------o----------o----------o

Wie hören Sie häufiger MP3s? o am Computer o mit mobilen MP3-Playern

Falls am Computer:
Welches Programm verwenden Sie hauptsächlich zum Abspielen ihrer MP3s?
o WinAmp o Windows Media Player o RealPlayer
o iTunes o Musicmatch Jukebox o Weiß nicht.
o Sonstiges: .....

Falls mit mobilen MP3-Playern:
Welche Art von mobilem MP3-Player besitzen Sie? (Mehrfachnennungen möglich)
o Festplatten-basierte Geräte:
o iPod o Creative Zen o Rio o iRiver H-Serie
o Sonstiges: .....

o Flash-Speicher basierte Geräte:
o iPod shuffle o iRiver i-Serie o USB-Stick o MP3-Handy
o Sonstiges: .....

o Autoradio mit MP3-Funktion:
o Becker o Pioneer o JVC o Sony o Blaupunkt
o Sonstiges: .....

o CD-basierte Geräte:
o MP3-CD-Player
o Sonstiges: .....

Wie hören Sie im Auto überwiegend Musik?
o Radio o Kassette o Audio-CD o MP3-Medien

Zu welchen Gelegenheiten hören Sie Musik im Auto? (Mehrfachnennungen möglich)

o Immer o Weg von/zur Uni/Arbeit o Freizeit
o Im Stau o Wenn alleine unterwegs o Auf langen Strecken
o Sonstiges: .....

Kommt es manchmal vor, dass Sie Ihren Beifahrer bitten, die Musikanlage im Auto zu bedienen? Falls ja, worum bitten Sie ihn? (außer Radiofunktionen)

.....

.....

.....

Nach welchen Kriterien ist ihre MP3 Sammlung organisiert? (Mehrfachnennungen möglich)

o Albumname o Künstlername o Titelname o persönliche Vorlieben
o Musikstil/Genre o Erscheinungsjahr o Stimmung o Autobiografie
o Art der Tracks (Musik, Comedy, Hörbücher...)
o Sonstiges: .....

Wenn Sie Playlisten benutzen, wie viele persönliche Playlisten verwalten Sie dann?

o 1 o 2-4 o 5-10 o >10

Nach welchen Kriterien stellen Sie ihre Playlisten zusammen?

.....

.....

.....

Wie groß ist ungefähr ihre MP3-Sammlung?

..... Titel ..... Gigabyte

III. Vorerfahrung Sprachdialogsysteme

Haben Sie bereits Erfahrung mit Sprachdialogsystemen?

o Ja o Nein o Weiß nicht

Wenn ja, mit welchen (z.B. telefonische Bahnauskunft, Postbank, Infotainmentgeräte im Auto)?

.....

.....

.....

B.2 Systemdokumentation

B.2.1 Vorarbeiten

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nicht zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

B.2.2 WOZ

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nicht zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

B.2.3 Prototyp „Dorothy“

Die Materialien zu diesem Abschnitt sind in dieser öffentlichen Version nicht zugänglich, da sie der Geheimhaltung durch Harman/Becker Automotive Systems unterliegen.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und nur die erwähnten Hilfsmittel und Quellen verwendet habe.

Dresden, den 9. Januar 2006

Stefan Schulz