

User Concepts for In-Car Speech Dialogue Systems and their Integration into a Multimodal Human-Machine Interface

Von der Philosophisch-Historischen Fakultät der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Sandra Mann

aus Aalen

Hauptberichter: Prof. Dr. Grzegorz Dogil

Mitberichter: Apl. Prof. Dr. Bernd Möbius

Tag der mündlichen Prüfung: 02.02.2010

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart

2010

Acknowledgements

This dissertation developed during my work at Daimler AG Group Research and Advanced Engineering in Ulm, formerly DaimlerChrysler AG. Graduation was accomplished at the University of Stuttgart, Institute for Natural Language Processing (IMS) at the chair of Experimental Phonetics.

I would like to thank Prof. Dr. Grzegorz Dogil from the IMS of the University of Stuttgart for supervising this thesis and supporting me in scientific matters. At this point I would also like to thank Apl. Prof. Dr. Bernd Möbius for being secondary supervisor.

I also wish to particularly thank my mentors at Daimler AG, Dr. Ute Ehrlich and Dr. Susanne Kronenberg, for valuable advice on speech dialogue systems. They were always available to discuss matters concerning my research and gave constructive comments. Besides I would like to thank Paul Heisterkamp whose long-time experience in the field of human-machine interaction was very valuable to me.

Special thanks go to all colleagues from the speech dialogue team, the recognition as well as the acoustics team. I very much acknowledge the good atmosphere as well as fruitful discussions, advice and criticism contributing to this thesis. I would especially like to mention Dr. André Berton, Dr. Fritz Class, Thomas Jersak, Dr. Dirk Olszewski, Marcel Dausend, Dr. Harald Hüning, Dr. Alfred Kaltenmeier and Alexandros Philopoulos. In this context I would like to add the Institute of Software Engineering and Compiler Construction from Ulm University, in particular Prof. Dr. Helmuth Partsch, Ulrike Seiter, Dr. Alexander Raschke and Carolin Hürster.

Furthermore, I wish to thank the students having been involved into this thesis: Andreas Eberhardt, Tobias Staudenmaier and Steffen Rhinow.

I also owe special appreciation to my parents, Gert and Elisabeth Mann, who enabled this work through their upbringing by constant encouragement and support.

Above all, I want to thank my heavenly father who accompanied this work from start to finish.

Contents

Acknowledgements.....	3
Contents.....	5
Abbreviations.....	9
1 Introduction	13
1.1 Motivation	15
1.2 Goal	20
1.3 Outline	20
2 Multimodal dialogue systems	23
2.1 Speech in human-computer interaction	24
2.2 Modality and multimodal interface	25
2.3 Multimodal in-car dialogue systems.....	28
2.3.1 The control and display concept of the Mercedes-Benz S-Class	30
2.3.2 User study of the Linguatronic speech interface.....	33
2.4 The aspect of in-car human-computer interaction.....	34
2.5 Investigations on driver distraction	39
3 Communication.....	47
3.1 Dialogue.....	47
3.2 Discourse analysis	48
3.2.1 Conversational principles.....	49
3.2.2 Speech acts	51
3.2.3 Presuppositions	58
3.2.4 Deixis	59
4 Human-computer communication with in-car speech dialogue systems	63
4.1 Architecture and functions of a speech dialogue system.....	63
4.2 Constraints of speech dialogue systems	71
4.3 Collecting usability and speech data	73
4.4 Designing the interface for usable machines.....	76
4.4.1 Reusable dialogue components	76

4.4.2 Prompts	81
4.4.3 Vocabulary	83
4.4.4 Feedback from the system.....	88
4.4.5 Help.....	90
4.4.6 Spelling	92
4.4.7 Voice enrolments	93
4.4.8 Barge-in.....	97
4.4.9 Error-recovery	98
4.4.10 Initiating speech dialogue	101
4.4.11 Short cuts.....	103
4.4.12 Initiative	104
4.4.13 Combining spoken and manual interaction.....	105
5 Accessing large databases using in-car speech dialogue systems.....	109
5.1 State-of-the-art of in-car audio applications.....	111
5.1.1 Constraints.....	111
5.1.2 User needs	113
5.2 Interaction concepts for searching audio data	113
5.2.1 Category-based and category-free search	114
5.2.2 Fault-tolerant word-based search	115
5.3 General requirements for the user interface	118
5.4 Prototype architecture.....	118
5.5 Verifying generating rules	121
5.5.1 The data elicitation method.....	122
5.5.2 Results	126
5.6 Combined-category input	133
5.7 Application-independent approach.....	138
6 Conclusions.....	147
6.1 Summary.....	147
6.2 Future work.....	157
Zusammenfassung.....	159
References	173

A Usability Guidelines.....	185
A.1 Jakob Nielsen: Ten Usability Heuristics	185
A.2 Ben Shneiderman: Eight Golden Rules	186
A.3 Sharon Oviatt: Ten myths of multimodal interaction.....	187
B Speech recordings of audio file names.....	189
B.1 Questionnaires	189
B.1.1 General information	190
B.1.2 Questions on the experiment.....	193
B.2 Extract of generating rules.....	195

Abbreviations

ABNF	Augmented Backus-Naur Form
ASR	Automatic speech recognition
BMBF	German Ministry of Education and Research
BSI	British Standards Institution
CI	Contextual interpretation
COMAND	Cockpit management and data system
DM	Dialogue manager
DP	Determiner phrase
ESoP	European Statement of Principles
G2P	Grapheme-to-phoneme
GUI	Graphical user interface
HMI	Human-machine interaction
HMM	Hidden Markov Model
IP	Inflectional phrase
ISO	International Organization for Standardisation
JAMA	Japan Automobile Manufacturers Association

LM	Language model
MM	Media manager
NHTSA	National Highway Traffic Safety Administration
NLU	Natural language understanding
NP	Noun phrase
OEM	Original Equipment Manufacturer
POI	Point of interest
PP	Prepositional phrase
PTA	Push-to-activate
RDC	Reusable dialogue components
SAE	Society of Automotive Engineers International
SDS	Speech dialogue system
SISR	Semantic interpretation for speech recognition
SNR	Signal-to-noise ratio
SYNC	Synchronisation module
TDDM	Task-driven dialogue manager
TE	Text enrolment

TICS	Transport information and control systems
TTS	Text-to-speech
UMTRI	University of Michigan Transportation Research Institute
VE	Voice enrolment
VP	Verb phrase
VUI	Voice user interface
WOZ	Wizard-of-Oz

Chapter 1

Introduction

Language is the most powerful instrument of cooperative action that humankind has, that *any* species we know of has.

(Randy Allen Harris, Voice Interaction Design)

Language is a feature that human beings acquire without particular training in contrast to cognitive skills such as basic calculations. Very young children around the age of five can speak and understand a natural language nearly as proficient as their parents (Fromkin, 1993, p.4). Speech as a natural means of communication not only facilitates interpersonal communication. It has also been transferred to the interaction of humans and data-processing machines. While in the sixties of the last century it was assumed that man-machine communication could be realised just as simple as human-human communication (e.g. Weizenbaum's Eliza from 1966) speech communication systems nowadays are still far from approximating natural communication. Behavioural psychologist B.F. Skinner assumed (1957) he could extend the original model of conditioning to human linguistic behaviour, i.e. language (Crain, 1999, p.40). But Brown and Hanlon (1970) were among the first psycholinguists to show that the basic behaviourist laws of reinforcement and stimulus generalization are too simple to hold for the complexities of a person's verbal behaviour (Crain, 1999, p.46). Even for a simple conversation well-founded knowledge in various fields is required such as:

- Phonetics: the production of human speech sounds
- Phonology: the organisation and usage of sounds

- Morphology: the internal structure of words, i.e. the smallest meaningful units of a language
- Syntax: the rules according to which words are combined to form grammatical sentences
- Semantics: the language meaning, i.e. the ways in which sounds and meanings are related
- Pragmatics: the determination of both literal and nonliteral aspects of communicated linguistic meaning

Taking this diversity and complexity into account it is obvious that systems for understanding natural language need highly complex software components – whereby the interaction of these components easily gets very complex as well. Moreover, in case large speech databases are involved (e.g. from music or address book applications) not only a good recognition rate must be guaranteed but in addition a profound knowledge of the relation between data has to be established and elaborated.

Since this thesis concentrates on speech dialogue systems in the automotive environment a natural human-machine interaction is highly required to minimise the driver's focus on the system while pursuing the driving task. Over the years the equipment of vehicles with assisting systems as well as luxury features has permanently increased. On the one hand this means that drivers are offered a big variety of options – on the other hand, however, the number of systems and the handling of its extensive functions imply that drivers are distracted from operating their vehicle and monitoring the traffic which can lead to car accidents. Natural language (as opposed to using commands) should facilitate communication between man and machine and thus provide hands-free operation during the driving task and ensure that the driver can keep his eyes on the road and not on various displays. This also must be guaranteed if interaction gets more complex because large vocabulary is involved. For example, if a driver asks the on-board navigation system to bring him to 'Neustadt' the ambiguity problem occurs that 29 different cities are named 'Neustadt' in Germany. A navigation system must be capable of distinguishing between these 'Neustadts' in order to find out which 'Neustadt' the user is heading for. One possibility would be to ask the user for additional information to disambiguate, like for example, the name of a city nearby. Once the disambiguation has been successful it might be helpful for the system to keep the selected solution in mind and – in case the user aims for 'Neustadt' again – offer it to the user as the preferred town.

Many approaches for speech dialogue systems have been published in recent years. These approaches are still far from a natural-language-understanding approach between human and machine. Moreover, in case of an automotive environment where computing and memory capacity are financially restricted – let alone the necessity to synchronise language with a manual/visual component – establishing a natural conversation between human and machine becomes even more challenging.

The topic of this doctoral thesis is to show that a natural way of communication between human and machine can be established even when larger databases are involved and computing power and memory load are restricted as it is the case in the automotive environment. The next section provides some basic aspects of understanding and disambiguating large vocabulary.

1.1 Motivation

Daimler Group Research and Advanced Engineering in Ulm has been engaged in the field of natural language understanding between humans and vehicles for more than twenty years. In 1996 the first speech system (so-called Linguatronic) was integrated into an S-class Mercedes. This first system offers a command and control access for entering telephone numbers into the phone application. With respect to the very restricted application the system performed quite well.

Over the years the equipment of vehicles with telematic (composed of the terms telecommunication and informatics) applications has permanently increased. Simultaneously, the complexity within an application has increased as well. Nowadays it is required to have speech access to navigation data, address data, multimedia data which are all based on large and complex databases. But still in current systems natural language access to these data has not yet been established. Sentences like ‘I want to go to Einstein street in Munich’ or ‘I want to listen to some music by Herbert Grönemeyer’ are still not possible. If short cuts (command-like phrases directly leading from one application to another) are allowed, sentences like ‘I am searching for Bochum’ cannot be distinguished as ‘Bochum’ is a city in Germany and a song/album by Herbert Grönemeyer as well. Current systems only allow for dialogues like the following:

Example 1:

System: Beep (*signal tone*).

User: Navigation.

System: Which city?

User: Bochum.

Example 2:

System: Beep.

User: Music browser.

System: Which music?

User: Bochum.

Even in case of unambiguous data no direct access to the music browser from a different application is possible:

Example 3:

System: Which city? (*System is in the navigation application*)

User: <Any title>.

System: Please select a city. (*Presents a pick list of incorrectly understood city names (see Figure 1.1)*)



Figure 1.1: Pick list of city names after having entered the navigation input mode

For humans, however, it is quite natural to directly tell the system what they want without taking the system's actual application context into account:

Example 4:

System: Beep (*System is in the telephone application*).

User: I want to listen to 'No Need to Argue'.

System: *Plays the title.*

And in case of ambiguous data in human communication it is easy to figure out what the other communication partner wants.

Example 5:

Speaker 1: I am searching for Bochum.

Speaker 2: Do you mean the city Bochum or the album or title of Herbert Grönemeyer?

Speaker 1: The title of Herbert Grönemeyer.

Evaluations of in-car speech dialogue systems show that such systems can be quite frustrating for users. Without explicitly studying handbooks, tasks can only be achieved with extreme effort. For the majority of users the tasks are time-consuming because applications are not clearly laid out. Very often users do not know what to say to the system at all and in quite a few cases they are not even aware of the application/task they are actually in due to lacking transparency or repetitive misrecognitions. For large databases the amount of misrecognitions increases even more. This leads to user frustration and rather sooner than later the user abandons using speech as input modality. In order to find out what is necessary for building user-friendly interfaces it is first of all necessary to figure out what really happens in human-human communication. What exactly is going on when humans are involved in a dialogue like the one in example 5 – or rather how is this kind of dialogue, the exchange of information successfully completed within human-human communication (Kronenberg, 2001, p.3)?

1. Obviously speaker 1 has the capacity to produce sounds that signify certain meanings and speaker 2 is able to understand and interpret the sounds produced by speaker 1 – and vice versa.

2. The speakers' utterances are continued mutually in a cooperative way. Speaker 2 could also be talking about something completely different. Instead, he knows what to refer to within what has been said and reinterprets it in such a way that the final interpretation can be deduced.
3. Adequately to the topic of the dialogue the number of turns is small.
4. Both speakers must have access to a common base of additional sources of information. Although "Bochum" has not been particularly introduced by speaker 1, speaker 2 knows that "Bochum" is not clear without ambiguity.
5. The knowledge of speaker 2 comprises several topics. He knows that "Bochum" is either a term from geography, referring to a city or from music, referring to either a music title or an album by an artist called Herbert Grönemeyer.

This kind of dialogue is also desirable when talking to a system. Users would expect this disambiguation question rather than a long result list of possible city names as it is currently realised in in-car speech dialogue systems. The result list might not even fit to what has actually been said. Apart from that glancing at the display where the result list is shown in order to check the recognition results is not relieving the driving task. Therefore the above aspects about what is typical of human-human dialogues can be transferred to speech dialogue systems to see what requirements have to be met in order to simulate this kind of interaction. The following requirements are crucial for speech dialogue systems to achieve cooperative dialogue flow between man and machine (Kronenberg, 2001, p.4).

The analysis process – within this process the system has to be able to understand what has been uttered by the user. This implies that in addition to understanding sounds it also has to be capable of processing syntactic constructions. The latter ability implies that antecedent references have to be made, i.e. subsequent parts of an utterance need to be referred to what has been previously said in order to obtain the final interpretation of what a user wanted.

The interpretation process – in order to interpret spoken input the system needs to have access to semantic and pragmatic knowledge. Semantic knowledge implies that the "literal" meaning of spoken input has to be ascertained. The "literal" meaning of spoken input comprises three different aspects: the inherent meaning of individual utterances, the semantic relations between

utterances and the overall meaning of a sentence. To bridge the gap between sentence meaning and speaker's meaning it is also necessary to have access to pragmatic knowledge. Every analysed input needs to be embedded in a global context in order to draw conclusions about what was said. As soon as ambiguities occur they must be detected and resolved by either deducing the correct meaning from contextual interpretation or – if further clarification is necessary - by subsequent turns containing disambiguation strategies.

The generation process – it implies that syntactically correct and meaningful (vs. nonsense) continuations of preceding utterances can be made. Apart from that the utterance which is the most likely and plausible has to be produced. To achieve that, it is again necessary to possess knowledge that goes beyond linguistic meaning, i.e. pragmatic knowledge.

A dialogue history – that provides an organised way of storing system data, analysis results and the hypotheses of various components. Thus, results from previous turns can be retrieved by the system and reused in such a way that the dialogue flow becomes more natural, time-saving and less distracting. Once an ambiguous city, like in the 'Neustadt' case, has been entered and selected by the user there would be no need to go through the whole task and picklist time and again whenever heading for this city. Instead, the user could be offered the former selection as favoured choice.

Multimodal interaction – where speech and manual input, and graphical and speech output complement each other. The user is thus able to switch from one modality to another depending on external conditions, i.e. when hands and/or eyes are occupied. This clearly increases efficiency and user friendliness. Besides, adjusting speech and display on each other by following the general principle "what you see is what you can speak" particularly supports novice users who tend to have the problem of not knowing what to say when talking to a speech dialogue system.

A successful interaction of these components is the focus of this thesis – it considerably adds to user-acceptance and turns spoken interaction between man and machine into pleasure.

1.2 Goal

The goal of this thesis is to unify and simplify complex applications as well as the handling of large databases to provide an approach for in-car speech dialogue systems that allows for user-friendly human-computer interaction. The interaction is based on natural language. Being synchronised with one another, the integration of both visual display and manual control establishes a robust complement to the speech component.

As mentioned above, the concepts of this approach are based on human-human communication. This does not imply that a one-to-one transfer can take place from human-human dialogues to human-machine interaction to end up with *the* ultimate speech dialogue system. Instead, one objective is to analyse aspects of human dialogues and extract those features and rules of behaviour that are essential for a natural and cooperative interaction between man and machine, e.g. in the field of pragmatics. However, essentials might as well be strategies running counter to rules for human-human communication when well proven in a particular context.

As for in-vehicle information systems large databases occur within applications such as navigation, audio or address book. Nowadays the trend is towards rising complexity and in particular novice users are shaken off because complexity may easily turn into lack of understanding and transparency. The cognitive workload while driving increases. The main focus of this thesis is on user-friendly concepts that, despite the complexity, make these applications transparent and their large data accessible. Offering different search strategies, for example in order to meet a user's insufficient or even false input which might be due to bad recollection of personally stored data, reduces the strain of cognitive resources.

1.3 Outline

Chapter 2 gives an overview of state-of-the-art dialogue systems and within this field introduces the most important terms that are relevant for this thesis. The aim is to investigate the special status of in-car speech dialogue systems in contrast to other speech dialogue systems. Why do they have to be treated differently from other dialogue systems and what problems occur with regard to the car environment? It is furthermore examined how car manufacturers currently handle large databases within applications.

After the introductory description of dialogue systems **Chapter 3** turns to the aspect of spoken language as a natural means of communication. Being the basis for human-computer interaction, communication between humans is analysed with particular focus on the field of pragmatics such as communicative principles and discourse analysis. From that strategies with regard to cooperative user concepts for speech dialogue systems may be deduced. These strategies may also imply that communicative principles have to be violated deliberately. It is always important to bear in mind that human-human dialogues do not strictly follow rules either. Communication partners for example often produce mental leaps. Thus, for human-computer interaction it is essential to consider these aspects and accommodate the user with appropriate strategies.

Before developing cooperative user concepts and integrating them into a multimodal environment it is important to evaluate automotive applications to detect the difficulties that might arise when interacting with an in-car speech dialogue system. Part one of **Chapter 4** describes architecture, functionality and limitations of a multimodal speech dialogue system. The second part introduces different methods of evaluations that were carried out in context with this approach and depicts the findings for designing usable multimodal in-car speech interfaces. The results portrayed there are a blend of aspects from human-human communication and user needs having crystallised during the evaluations. Basic features enhancing voice-controlled in-car applications are for example to

- Enlarge the strict grammar of current command and control systems towards a speech approach that is more natural and intuitive
- Integrate short cuts allowing direct leaps from one application task to another in order to reduce the number of dialogue turns
- Accelerate dialogue flow by enabling the user to interrupt system utterances by speech
- Integrate a dialogue history to retrieve analysis results from previous turns
- Unify the structure of speech dialogue systems by providing dialogue components that are broadly reusable

Combining manual input and output – as it is realised in automotive HMI systems – and spoken input and output, creates further requirements for cooperative human-computer communication. Care must be taken that the modalities permanently exchange user input data to be on the same

level of knowledge. In so doing the manual interface may helpfully complement the speech interface, thus becoming a supportive means for human-computer dialogue. Also, it enables the user to switch from one input modality to the other at every step without having to start a task all over again. This makes human-computer interaction efficient and considerably adds to user acceptance.

In a further step concepts for cooperative speech dialogue systems are presented. **Chapter 5** combines an interaction of the above features with various search strategies to allow for accessing large databases like for example audio data, destination data and address book data. Due to the increasing number and complexity of electronic on-board devices over the past years current methods of navigating audio, navigation (points of interest) and address book data are no longer sufficient. They tend to have lost transparency. The presented concepts offer an in-car human-computer interaction that is user-friendly for both novice and expert user. They are based on a feature called text enrolment that provides speakable text entries by automatically transcribing dynamic data. Care is taken that spoken and manual interaction are synchronised.

Chapter 6 summarizes the findings of this study and contrasts human-human communication to human-computer interaction with regard to the concepts and strategies elaborated in Chapters 4 and 5. It is examined to what extent human communication principles and strategies have actually been transferred to human-computer interaction. Some suggestions for future research conclude this contribution.

Chapter 2

Multimodal dialogue systems

Areas applying human-computer interaction nowadays are manifold (Gibbon, 2000, p.116; McTear, 2004, p.22). Applications are for example

- Transactions & queries: electronic commerce (banking, online-shopping), call centers (travel information, weather forecasts, stock prices etc.)
- Data entry: PDAs, equipment maintenance, dictation, programming
- Controlling devices: cockpit (e.g. in-car, plane, motorbike), hospital (e.g. for medical operations), factories (e.g. for meat or bakery products)
- Education: intelligent tutoring systems (e.g. for schools or military service)
- Edutainment: chatterbots (e.g. Eliza, A.L.I.C.E)
- Entertainment: games, animations

Within the field of human-computer interaction speech has become a rapidly growing technology – either to replace the graphical user interface, as additional component or to completely substitute dialogues between humans. The origins of speech dialogue technology date back to artificial intelligence research in the 1950s (McTear, 2002, p.1).

This chapter conveys some basic knowledge in the field of human-computer interaction in order to pinpoint the type of dialogue system the thesis focuses on. Furthermore, to foreshadow the difficulties users encounter when applying multimodal in-car dialogue systems, a close look is taken at the special car environment.

2.1 Speech in human-computer interaction

Over the past years two major kinds of speech technology were brought on the market. Speech recognition systems like dictating machines allow for input of spoken language in context with large vocabulary which is transferred into written text. Products available on the market are for example Dragon NaturallySpeaking (Nuance, 2009), ViaVoice (IBM, 2009), SpeechMagic (Philips, 2009) – the latter is often used in medical practices. Due to the large vocabulary these systems comprise, they used to be speaker-dependent, i.e. before using such a system the user had to speak a text for about half an hour. In the meantime the time needed for training has shrunk to a few minutes only and some manufacturers already offer dictation systems that are speaker-independent (e.g. the latest version of Dragon NaturallySpeaking).

Speech dialogue systems not only recognise what has been spoken by the user but they are also able to interpret it. They are called speech understanding systems. Systems of the input type command and control have been on the market for several years now. They allow the user to control technical devices by using speech commands or fixed word sequences that can be entered fluently and without artificial pauses. These systems imply that the user has to be familiar with the tasks for he needs to know the commands that may be entered. They work speaker-independent. Due to the restriction of possible utterances speech recognition is usually robust. Speech dialogue systems providing less restricted input (i.e. conversational systems) do not force the user to use particular wording. He may apply continuous speech. Besides the system is capable of filtering relevant information from what has been uttered by the user. Like command and control systems this kind of system is speaker-independent. However, it is important to be aware that these systems are limited. The relevant software components are fairly complex and prone to errors – speech recognition for example nowadays is still lacking reliability, but also the natural language understanding unit may misinterpret the user's intention. In contrast to written language spontaneous speech is unpredictable because speakers make mistakes, hesitate or start sentences all over again. Therefore the idea of integrating speech into conversational (i.e. fully natural and interactive) systems is not yet realisable but remains the ultimate aim for the following decades. Chomsky's creative aspect of language use (Chomsky, 2006, p.6), however, clearly foreshadows that no matter how good recognition technology one day might be, language is infinite and subject to change, i.e. it will never be possible to compute *all* probabilities. What is required for successful interaction is good design based on usability testing.

2.2 Modality and multimodal interface

The type of speech dialogue system the thesis deals with is a multimodal interface. But when is an interface multimodal and what does the term modality refer to? The terminology has undergone numerous discussions. From a physiological point of view there exist six modalities, i.e. visual, auditive, tactile, olfactory, gustatory and vestibular (Schomaker, 1995, p.5; also see Silbernagel, 1979) – correspondingly to the six senses of human beings (see Table 2.1).

Sense organ	Sensory perception	Modality
Eyes	Sight	Visual
Ears	Hearing	Auditive
Skin	Touch	Tactile, haptic
Nose	Smell	Olfactory
Tongue	Taste	Gustatory
Organ of equilibrium	Balance	Vestibular

Table 2.1: Perceptive modalities derived from the human senses (Schomaker, 1995, p.5)

According to that modality is of perceptive nature only. In the field of human-computer interaction, however, it is not sufficient to restrict the term modality to sensory perception only. Interaction always implicates two sides, a perceptive *and* a productive side. It is therefore necessary to widen the definition such that a modality is

a particular way of perceiving *or* conveying information that is exchanged between dialogue participants¹.

To convey information it does not necessarily take many words – speech is one possible modality but in order to understand their counterpart people do not only use speech, they also

¹ In this thesis the role of a dialogue participant is not restricted to humans but can also be taken over by computers. Consequently the terms input and output modality hold for both human and computer.

make use of additional means of communication. Speech is usually accompanied by non-verbal cues like gestures, facial expressions or lip-reading (Gibbon, 2000, p.115). The more modalities are involved the more precise is the degree of perceiving the information that is actually conveyed. A human-computer interface is multimodal

when speech is combined with two or more modalities on either the input or output side of the interface.

It usually comprises a subset of the modalities presented in Figure 2.1.

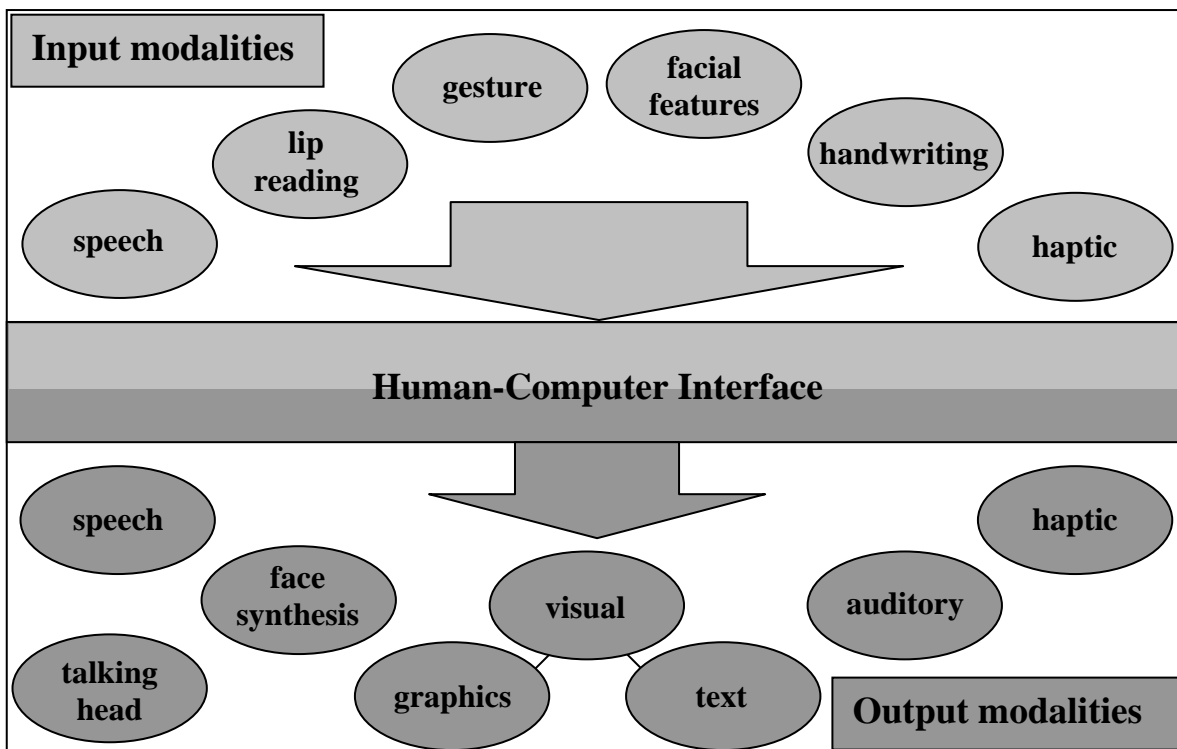


Figure 2.1: Input and output modalities in human-computer interaction (Gibbon, 2000, p.115)

One of the first multimodal prototypes was invented by Richard Bolt in 1980. In his “Put-that-there” experiment (Bolt, 1980) users could create and modify simple geometric shapes by combining speech input with gesture recognition. The vocabulary was very limited.

The research project SmartKom, funded by the German Ministry of Education and Research (BMBF), focused on combining a variety of modalities comprising speech in combination with

gestures, facial features and graphic control (Wahlster, 2006, p.4). What is more, these modalities are not only provided for on the input side of the system but also on the output side. This is achieved by means of a conversational agent called Smartakus (cf. Reithinger, 2006). He plays the role of various companions for three different scenarios (Wahlster, 2006, p.7):

- SmartKom-Public: a communication companion for phone, fax, email and authentication tasks
- SmartKom-Home: an infotainment companion for selecting media content and operating various TV applications
- SmartKom-Mobile: a mobile travel companion for navigation and point-of-interest information retrieval in location-based services

The aim of the SmartKom project was to improve interaction by transferring the natural character of human-human communication, i.e. the simultaneous combination of modalities, to human-computer interaction – with particular focus on novice users.

It is definitely a long-term aim to integrate as many modalities as possible into human-computer interaction. However, this is not the aim of this thesis. This work focuses on in-car information systems that are currently available on the market. Manufacturers of advanced cars enable the driver to operate various applications no longer via manual input only – with respect to current standards of technology (on telematic services for example) manual input would no longer be adequate and sufficient. Instead they offer a multimodal interface that enables controlling devices using speech as additional means of interaction. The modalities encountered in context with in-car human-computer interaction are haptic and speech as input modalities and speech, visual and auditory as output modalities (see Figure 2.2).

Why are non-verbal cues not made available for in-car human-computer interaction? Input modalities that require image processing are quite costly as it would be necessary to additionally install a camera inside the car. Apart from that modalities such as lip reading would be difficult to tackle as the driver's face is in movement while driving and therefore cannot be detected easily. Strongly varying weather conditions also have a negative impact on the degree of precision. As far as the output side is concerned an agent with gestures and facial expressions might be distracting the driver whose primary task is and should be to follow the traffic.

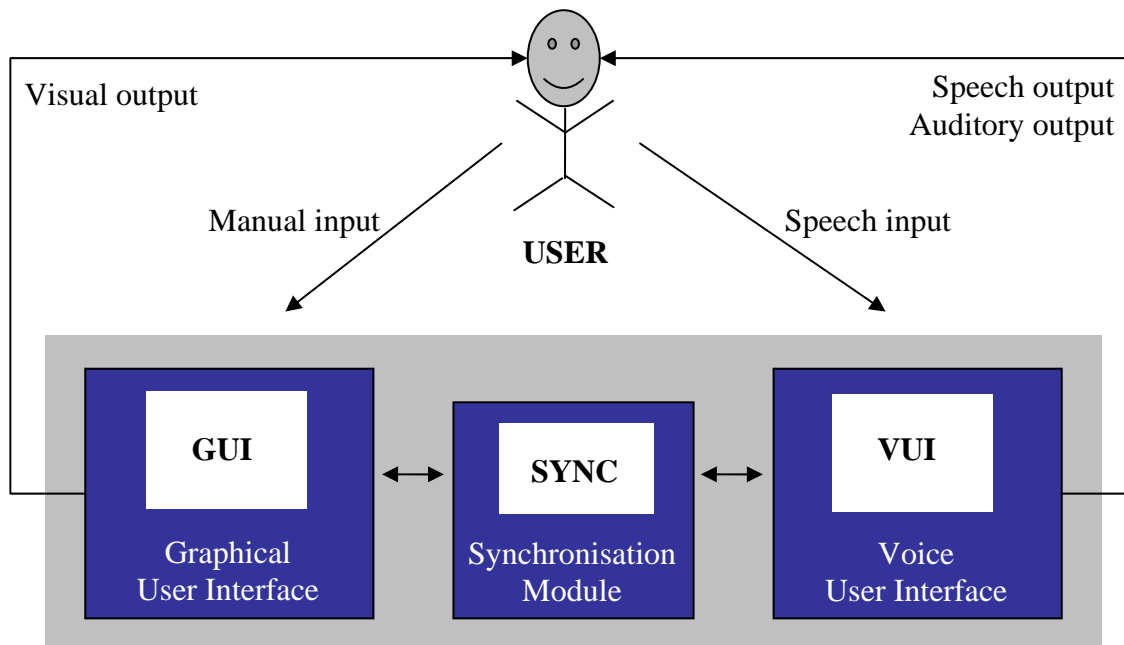


Figure 2.2: In-car human-computer interface

2.3 Multimodal in-car dialogue systems

Multimodality of in-car information systems has permanently increased in connection with the aspect of safety while pursuing the driving task. Due to the expansion of the telematics field into the automotive environment the amount of information systems has strongly increased. Heidingsfelder (2001, p.13) differentiates between two categories of information systems: one category directly related to the vehicle itself and one providing applications that had originally been restricted to the areas home and office (see Table 2.2, adapted from Heidingsfelder, 2001, p.14; Philopoulos, 2002, p.4).

Considering nowadays' in-car information systems the question arises to what extent users are able to interact with this variety while driving without decreasing driving safety. The point is to make the large amount of information systems accessible via different modalities. Speech interfaces in particular are welcome in situations where hands and/or eyes are busy (Weevers, 2004, p.11; Shneiderman, 2004). Combining them with a graphics/haptics interface unifies multiple modalities that are completely contrastive such that they may complement each other ideally. Depending on the driving situation the user is able to flexibly switch between modalities, i.e. from speech to manual input and vice versa.

Car-specific features	
<p>Navigation and traffic</p> <ul style="list-style-type: none"> - Basic navigation - Dynamic navigation - Traffic reports - Points of interest - Travel guide 	<p>Safety and emergency</p> <ul style="list-style-type: none"> - Automatic air bag notification - Emergency call - Roadside assistance - Vehicle tracking - Brake assistent
<p>Vehicle services</p> <ul style="list-style-type: none"> - Remote diagnostics - Warranty failure detection - Scheduled maintenance - Software updates 	<p>Intelligent transportation</p> <ul style="list-style-type: none"> - Adaptive cruise control
Non-car-specific features	
<p>Information and entertainment</p> <ul style="list-style-type: none"> - General or customized news - Audio or video downloads - Audio/video streaming - Interactive games 	<p>Mobile communication</p> <ul style="list-style-type: none"> - E-mail, text message, multi-media message - Internet, Intranet - Voice over IP - Mobile phones - PDAs
<p>M-Commerce</p> <ul style="list-style-type: none"> - Banking - Shopping - Concierge services - Electronic toll collection 	

Table 2.2: Categories of in-car information systems

Communication between humans is always multimodal (cf. Chapter 2, 2.2). Consequently the aspect of multimodality should make human-computer interaction more natural and familiar, not least since the user may transfer ‘interaction’ strategies from human-human communication to the computer. When it comes to contrasting multimodal interaction and unimodal interaction the results of various studies confirmed that combined usage of different modalities is strongly preferred by users and advantageous regarding error recovery and task completion rate (see e.g. Cohen, 2000; Oviatt, 1997).

2.3.1 The control and display concept of the Mercedes-Benz S-Class

Advanced cars, such as Audi (2005), BMW (2006) and Mercedes-Benz (2003) etc., offer the user a multimodal dialogue system to control a variety of the services presented in Table 2.2. As for spoken input most of these systems currently require input in form of commands or command sequences. The control and display concept of the Mercedes-Benz S-Class (Daimler AG, 2008) provides the user with a combination of graphical user interface (GUI) and voice user interface (VUI).

The graphical user interface comprises two displays: the high resolution multi-function display in the instrument cluster for vehicle functions that can be controlled by means of two buttons on the multi-function steering wheel and a swivelling COMAND (Cockpit Management and Data System) display to the right of the instrument cluster (see Figure 2.3).



Figure 2.3: View of operating devices of the Mercedes-Benz S-Class (Daimler AG, 2008)

The COMAND system comprises telematic and audio devices as well as vehicle functions, i.e. the user can control the applications navigation, telephone and address book, audio, video and vehicle (Daimler AG, 2008). The display (see Figure 2.4) contains five layers: status bar, main functions, main area, submenus and an air-conditioning function line.

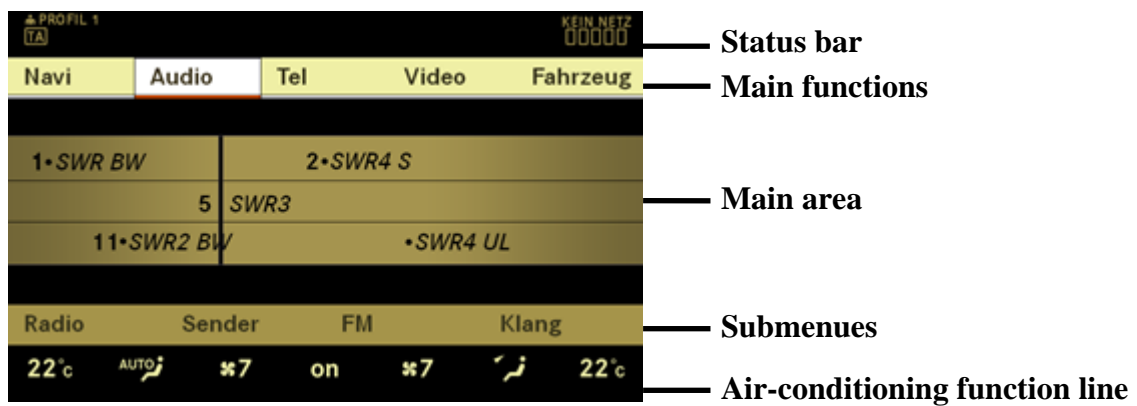


Figure 2.4: COMAND display (main menu) of the Mercedes Benz S-Class (Daimler AG, 2008)

To navigate the system (e.g. to scroll through menus or select particular items) a COMAND-controller situated in the lower section of the vehicle centre console can be pressed, turned and slid (see Figure 2.5). The currently selected function is always highlighted on the display.

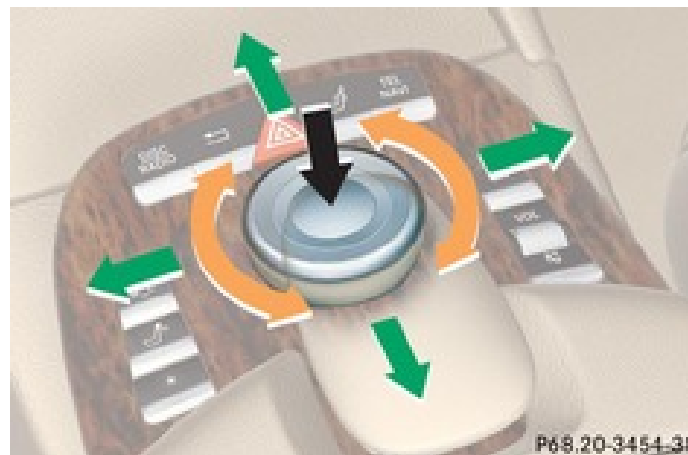


Figure 2.5: Mercedes-Benz S-Class – COMAND-controller and function buttons (Daimler AG, 2008)

For hands-free operation the voice user interface or Linguatronic (Heisterkamp, 2001) needs to be activated by pressing the so-called Push-to-Activate button (PTA) that is positioned on the steering wheel (Daimler AG, 2008). Having pressed this button, feedback via a signal tone is returned indicating the user that the system is now ready to accept spoken input. This is enabled via microphone arrays positioned in the overhead control panel. The type of input is command and control. In a command and control dialogue the user is “commanding” the device, so he must know both what commands to give and what to expect when the commands are properly

executed (Balentine, 2001, p.148). Table 2.3 presents an extract of speech commands of the applications relevant for this thesis: navigation, phone and audio.

Navigation	ComTel	Audio
Start route guidance	Phone on	Radio on
Stop route guidance	Address book on	CD on
Guidance instructions on	Enter pin code	Audio DVD on
Guidance instructions off	Dial number	Traffic information on
Switch to map	Redial number	Traffic information off
Spell country	Store name	Next station
Enter town	Read out address book	Previous station
Spell town	Read out phone book	Frequency selection
Spell street	Read out destination memory	Waveband selection
Enter house number	Delete address book	Store station <voice name>
Spell town center	Dial name	Select station
Enter destination	Dial <voice name>	Listen to station list
Store destination	Delete name	Next medium
Enter POI (point of interest)	Navigate <voice name>	Previous medium
Last destinations	Open entry	Next track
Map smaller	Spell entry	Previous track
Map bigger	Create <voice name>	Next directory
Map minimal size	Change <voice name>	Previous directory
Map maximal size	Delete <voice name>	etc.

Table 2.3: Extract of possible speech commands of the Linguatronic

Commands such as “map smaller”, “next track”, “previous track” are straightforward requiring no further interaction between user and system whereas commands like for example “enter destination” initiate a dialogue that might take one or more subtasks (e.g. enter country, city, post code, street and house number) as example (6) illustrates.

Example 6:

User: (Beep) Navigation.

System: Navigation menu.

User: Enter destination.

System: Please speak the city. (Subtask city)

User: Munich.

System: (*Presents picklist*) Please select an item.

User: Number two.

System: Is that entry correct?

User: Yes.

System: Do you want to enter a street name? (Subtask street)

User: Yes.

System: Please speak the street name.

User: Main Street.

System: (*Presents picklist*) Please select an item.

User: Number one.

System: Is that entry correct?

User: Yes.

System: Do you want to enter a house number? (Subtask house number)

User: Start route guidance.

When speech is used as input modality the system provides feedback as spoken output, i.e. via speech synthesis but also always through visual feedback. In case the user does not know what to say a teleprompter displays all commands that may be currently spoken – a feature that can be distracting for the driver, especially if the list of commands occurring on the display is fairly long. Additionally the voice user interface provides a help component for the available functions. It can be selected at any step returning a list of possible commands via spoken output.

2.3.2 User study of the Linguatronic speech interface

An evaluation of the Linguatronic interface was conducted on a set of 46 subjects at an average age of 52 years (Enigk, 2004; Mann, 2006). 74% of the participants were male and 26% female. The aim of the study was to investigate what kind of mental models users have and how these develop while interacting with the speech dialogue system. It was also focused on how far user acceptance changes during the testing period.

The test setting comprised 6 different tasks on the applications audio, telephone and navigation:

1. Audio: selecting a favourite radio channel
2. Telephone: dialling a phone number
3. Navigation: entering a particular destination and start navigation
4. Navigation: changing map size on the display
5. Navigation: entering new destination and start navigation
6. Telephone: calling a particular person using her mobile phone number

The findings showed that general acceptance of the Linguatronic interface is very high, i.e. the innovative character of speech control. The participants were highly motivated in using spoken interaction. The given tasks were fulfilled effectively. However, they were not efficient as the time needed for completion was too long. Operating the system is considered difficult and cumbersome:

- The logic of the PTA button is not intuitive
- Manual and speech control are not uniform and consistent
- An abundance of choice leads to lacking transparency
- The integrated help function is hardly used and if so, it is not regarded helpful

The subjects' mental model was rudimentary only, leading to an enormous effort required for getting acquainted with the system. Conceptual design as opposed to technical mistakes is regarded a major error source for inefficient and inadequate use of the system. The thesis focuses on methods that, despite the complexity and variety, enable both novice and expert users to keep an overview of the system and accomplish tasks in a shorter period of time than with current systems.

2.4 The aspect of in-car human-computer interaction

Dialogue systems that 'understand' spoken language inside the car encounter more difficulties than interfaces such as telephony user interfaces (e.g. for information enquiries or travel reservations). These difficulties are due to the special communicative situation of in-car speech dialogue systems on the one hand and to their design and implementation on the other.

The *acoustic environment* of in-car speech dialogue systems deviates from systems employed outside the car. The narrow environment inside the car leads to stronger reflection of speech signals. This results in stronger echoes coming from the loudspeakers. If the option of barge-in² is provided for, these echoes interfere with user input and decrease the recognition rate. Using microphones to enable hands-free spoken interaction also impairs the incoming speech signal. The longer the distance between driver and microphone the weaker the speech signal gets.

Noise factors inside the car are manifold. The main factors are engine noise, wind noise, tire noise and noises from devices such as climate control or audio (Schmidt, 2006, p.550; also see Puder, 2006). Thus the ambient noise level is far more aggressive (see Figure 2.6) compared to applications carried out in a home or an office environment, in particular if cars are driven at medium and high speed. Recognition rate of both interaction partners sinks.

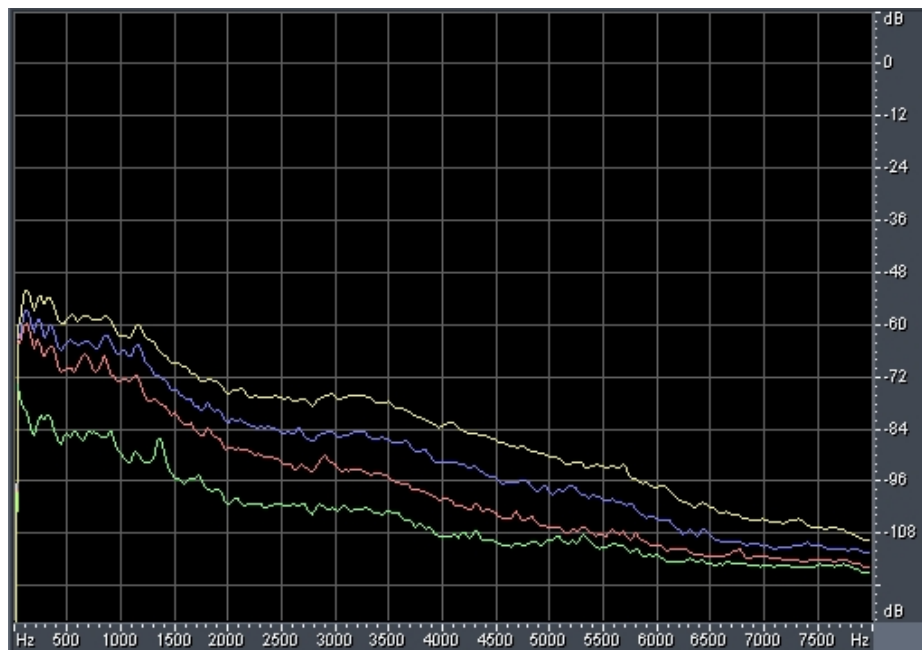


Figure 2.6: Noise measured in a Mercedes-Benz S-Class at a speed of 0 km/h (green), 70 km/h (red), 100 km/h (blue) and 140 km/h (yellow)

² Barge-in implies that system output, i.e. speech synthesis, may be verbally interrupted by the user any time.

A phenomenon observed in noisy environments is that the more the noise level increases the louder the speaker's voice gets, i.e. formant amplitudes increase. The effect of speaker adaptation is called Lombard effect (Lombard, 1911). Parallel to loudness other speech characteristics change as well (Schleß, 2000, p.42; Schmidt, 2006, p.550):

- Pitch frequency increases
- Duration of vowels is longer
- Formant frequencies of vowels are increased

Adapting speech characteristics, however, only partially compensate for the negative impact of a high environmental noise level. To be more precise, it is the increase of the signal-to-noise ratio (SNR) arising from the increasing utterance power that contributes to better recognition (Ogawa, 2007). The other changes within speech spectra caused by the Lombard effect are negligible.

Methods for reducing background noise and acoustic echoes are manifold (Hänsler, 2006; Vaseghi, 2006) but unfortunately none of them is one hundred per cent. What is currently employed in the automobile industry is for example a combination of microphone arrays, spectral subtraction and echo cancellation. Microphone arrays can be grouped effectively such that information coming from the driver is intensified while simultaneously the diffuse ambient noise level can be partially suppressed (Schleß, 2000, p.13). Spectral subtraction estimates the noise level (e.g. in speech pauses) in every given frequency channel of the absolute signal spectrum and subtracts these values from the spectrum (Vaseghi, 2006, p.417; Heute, 2006, p.344). To encounter the strongly reflecting in-car environment echo cancellation is applied. It is based on synthesizing a copy of the echo that is then subtracted from the initial signal (Schleß, 2000, p.14; also see Hänsler, 2006).

Turning from the acoustic environment to the driver, i.e. user, there is an additional factor having an enormous impact on the success rate of a system's performance. User characteristics such as speaking mode vary significantly in an automobile environment. What leads to a change in speaking mode are so-called behavioural barriers (Gellatly, 1997, p.15): psychological stress (e.g. fatigue, anxiety, changes in levels of stress or workload) as well as physical stress (e.g. acceleration, heat or vibration).

Operating in-car devices while simultaneously pursuing the driving task creates a *multitasking environment*. The primary driving task itself presents a complex task where coordinating cognitive, perceptual and physical skills is obligatory. Information exchanged between driver and interface is secondary. Secondary tasks can be distracting for the driver as soon as they interfere with any of the driving-related skills. The National Highway Traffic Safety Administration (NHTSA) splits distraction into four major categories: visual distraction (e.g. focusing the display of the graphical user interface), auditory distraction (e.g. answering an incoming telephone call), physical distraction (e.g. manually entering a destination) and cognitive distraction (e.g. getting involved in a discussion with the co-driver) (Ranney, 2000, p.1). Accomplishing a secondary task may involve more than one of these categories simultaneously. Consequently the driver's own safety and the safety of the motoring public may decrease. Concerning user interfaces outside the car environment, there usually is just one particular task the user wants to get accomplished. Besides, he also has the freedom to concentrate on it, i.e. there is nothing else that keeps drawing his attention from the actual task.

In-car interfaces enabling voice input are an effort to minimise visual and physical distraction that comes along with manually operating technical devices. However, they do not solve the problem of cognitive distraction. Numerous studies have been carried out examining the effect of technical devices (e.g. in-car navigation systems, mobile phones, entertainment systems as well as internet and email facilities) on driver distraction (Young, 2003). Research conducted by Burch (2002, p.7 et seqq.) confirmed that both hand-held and hands-free mobile phones negatively impact braking response time of the driver. The greatest impact occurred while using a hand-held mobile phone: on average users took half a second longer to react compared to using no phone at all. Taking a speed of 70 miles per hour, this difference would yield travelling an additional 14 metres before reacting to a hazard on the road. For hands-free mobile phone users it took an additional 8 metres to normal response. Burch (2002, p.8) attributes this to the fact that mobile phone conversation alone is a distracting factor for driving safety. However, research on different tasks such as internet and email access also shows that speech-based interaction undermines driving safety (Burns, 2002; Lee, 2001).

In a driving simulator Tsimhoni (2004) examined the effect of entering an address into a navigation system. Users had to accomplish tasks via three different methods: word-based

speech recognition, character-based speech recognition and typing on a touch-screen keyboard. The findings (Tsimhoni, 2004, p.603) showed that the tasks solved by means of word-based speech recognition were outstanding with a shortest total task time of 15.3 seconds while driving on sharp curves versus 14.2 seconds while parked. Total task time for typing increased significantly from 32 seconds while parked to 63 seconds on straight sections to 84 seconds on sharp curves. Young (2003, p.VI) confirms that route guidance systems with voice recognition technology are a more ergonomic and safer option than systems that require visual-manual entry. Cognitive load always increases with secondary tasks. Therefore careful attention needs to be paid to the design of speech dialogue systems and their underlying complexity in order to lower the increase.

Concerning the *design and implementation phase* in-car speech dialogue systems encounter limitations. They are systems where the graphical user interface is designed and developed first, detached from the voice user interface. This implies that the voice user interface is added to a self-contained system. Many theories are based on the fact that there is only 'speech' as modality. But with regard to in-car speech dialogue systems, the world between human and computer has the GUI as basis. And everything is driven through this graphics/haptics world. This means that speech is merely an attachment. This method is not desirable, instead the two approaches should undergo a permanent process of merging together – because care needs to be taken when two approaches that are so complementary are unified. In contrast to in-car speech dialogue systems where the user can choose between manual or spoken input, telephony applications mainly focus on speech as input modality – manual input is also enabled but it is more or less restricted to selecting digits. But what is returned to the user is only speech, thus there is no need to synchronise a visual and auditory world as it is the case for in-car speech applications.

When developing a new in-car speech dialogue system for a particular series, it is necessary to provide technology that is downwards compatible. When then buying a new model of a car series the users that are already familiar with a system do not have to go through completely new approaches over and over again. However, this process can be very restrictive for the designer of in-car speech dialogue systems.

In order to keep the costs of in-car speech dialogue systems for the end user as low as possible it is aimed at using as little hardware as possible. Consequently the final product has limited memory capacity which in turn demands compromises when it comes to designing the interface.

2.5 Investigations on driver distraction

Considering the risk of driver distraction in-car dialogue systems might become object to legislative activities rather sooner than later. There are no precise data to what extent inattention while driving is an influential factor on accidents. Estimations from experts range from 20-50% (Transport Canada, 2003, p.3; also see Wang, 1996). The NHTSA (2006) analysed the effect of driver inattention in a naturalistic driving study. The experiment comprised data recorded from 109 cars over a period of 12 to 13 months per car, representing everyday driving including the stress encountered in a metropolitan area (NHTSA, 2006, p.1). These real-world observations of drivers' behaviour enable to directly relate driving inattention and crash and near-crash involvement (NHTSA, 2006, p.VII). Driver inattention was split into four different types: secondary tasks, driving-related inattention to the forward roadway, drowsiness and non-specific eye glance away from the forward roadway (NHTSA, 2006, p.21). Findings showed that engaging in secondary tasks (e.g. talking to a co-driver, dialling hand-held mobile phone, dialling hands-free mobile phone using voice activated software, operating PDA, eating, drinking out of open cup, smoking cigarette, applying make-up, etc.) contributed to over 22% of all crashes and near-crashes (NHTSA, 2006, p.117). Analysing eye glance behaviour showed that total eyes-off-road duration of more than two seconds significantly increased crash and near-crash risk (NHTSA, 2006, p.118). Whereas systematic eye glances related to the primary driving task enhance safety - provided they do not exceed 2 seconds – looking at objects inside the car undermines safety.

To contribute to road safety numerous standards, guidelines and recommendations on in-car devices have been developed over the past years (Transport Canada, 2003), e.g. by the

- European Commission (EC)
- International Organization for Standardisation (ISO)
- Society of Automotive Engineers International (SAE)
- British Standards Institution (BSI)

- Japan Automobile Manufacturers Association (JAMA)
- UMTRI Guidelines (University of Michigan Transportation Research Institute)

The European Commission for example invented the European Statement of Principles (ESoP) on the components and functions that constitute the interface and interaction between system and driver (Transport Canada, 2003, p.23). The set of principles covers the design, installation and information presentation, interaction with displays and controls, system behaviour and information about the system (EC, 2006). The following principles present an extract thereof.

Overall design principles (ibid., p. 7):

- The system supports the driver and does not give rise to potentially hazardous behaviour by the driver or other road users.
- The allocation of driver attention while interacting with system displays and controls remains compatible with the attentional demand of the driving situation.
- The system does not distract or visually entertain the driver.
- The system does not present information to the driver which results in potentially hazardous behaviour by the driver or other road users.
- Interfaces and interface with systems intended to be used in combination by the driver while the vehicle is in motion are consistent and compatible.

Interaction with displays and controls (ibid., p.18):

- The driver should always be able to keep at least one hand on the steering wheel while interacting with the system.
- The system should not require long and uninterrupted sequences of manual-visual interfaces. If the sequence is short, it may be uninterrupted.
- The driver should be able to resume an interrupted sequence of interfaces with the system at the point of interruption or at another logical point.
- The driver should be able to control the pace of interface with the system. In particular the system should not require the driver to make time-critical responses when providing inputs to the system.

- System controls should be designed such that they can be operated without adverse impact on the primary driving controls.
- The driver should have control of the loudness of auditory information where there is likelihood of distraction.
- The system's response (e.g. feedback, confirmation) following driver input should be timely and clearly perceptible.
- Systems providing non-safety related dynamic visual information should be capable of being switched into a mode where that information is not provided to the driver.

System behaviour principles (ibid., p.24):

- While the vehicle is in motion, visual information not related to driving that is likely to distract the driver significantly should be automatically disabled, or presented in such a way that the driver cannot see it.
- The behaviour of the system should not adversely interfere with displays or controls required for the primary driving task and for road safety.
- System functions not intended to be used by the driver while driving should be made impossible to interact with while the vehicle is in motion, or, as a less preferred option, clear warnings should be provided against the unintended use.
- Information should be presented to the driver about current status, and any malfunction within the system that is likely to have an impact on safety.

Within the International Organization for Standardization (ISO) standards are developed by expert committees from 157 countries (ISO, 2009). These standards are intended to ensure desirable characteristics of products and services such as quality, environmental friendliness, safety, reliability, efficiency and interchangeability and thus provide governments with a technical base for health, safety and environmental legislation.

Concerning vehicles and their technical devices they have developed the following standards:

ISO 15005:2002, Road vehicles -- Ergonomic aspects of transport information and control systems -- Dialogue management principles and compliance procedures.

This International Standard presents ergonomic principles for the design of the dialogues that take place between the driver of a road vehicle and the vehicle's transport information and control systems (TICS) while the vehicle is in motion. It also specifies compliance verification conditions for the requirements related to these principles.

This International Standard is applicable to TICSs consisting of either single or multiple devices, which can be either independent or interconnected. It is not applicable to TICSs without dialogues, TICS failures or malfunctions, or controls or displays used for non-TICS functions.

ISO/TS 16951:2004, Road vehicles -- Ergonomic aspects of transport information and control systems (TICS) -- Procedures for determining priority of on-board messages presented to drivers.

ISO/TS 16951:2004 provides formal procedures and two alternative methods for determining the priority of on-board messages presented to drivers of road vehicles by transport information and control systems (TICS), and other systems. It is applicable to the whole range of TICS in-vehicle messages, including traveller information, navigation, travel and traffic advisories, "yellow pages" information, warnings, systems status, emergency calling system information, and electronic toll/fee collection, as well as to messages from non-TICS sources such as telephone, warnings and telltales.

ISO 15006:2004, Road vehicles -- Ergonomic aspects of transport information and control systems -- Specifications and compliance procedures for in-vehicle auditory presentation.

ISO 15006:2004 establishes ergonomic specifications for the presentation of auditory information related to transport information and control systems (TICS) through speech or sounds. It is applicable only to the use of auditory displays when the vehicle is in motion. It presents a set of requirements and recommendations for in-vehicle auditory messages from TICS, and provides message characteristics and functional factors for maximizing message intelligibility and utility while helping prevent auditory or mental overload.

ISO 15008:2009, Road vehicles -- Ergonomic aspects of transport information and control systems -- Specifications and test procedures for in-vehicle visual presentation.

ISO 15008:2009 specifies minimum requirements for the image quality and legibility of displays containing dynamic (changeable) visual information presented to the driver of a road vehicle by on-board transport information and control systems (TICS) used while the vehicle is in motion. These requirements are intended to be independent of display technologies, while reference to test methods and measurements for assessing compliance with them have been included where necessary.

ISO 15008:2009 is applicable to mainly perceptual, and some basic cognitive, components of the visual information, including character legibility and colour recognition. It is not applicable to other factors affecting performance and comfort such as coding, format and dialogue characteristics, or to displays using

- Characters presented as a part of a symbol or pictorial information
- Superimposed information on the external field (e.g. head-up displays)
- Pictorial images (e.g. rear view camera)
- Maps and topographic representations (e.g. those for setting navigation systems) or
- Quasi-static information

ISO 17287:2003, Road vehicles -- Ergonomic aspects of transport information and control systems -- Procedure for assessing suitability for use while driving.

ISO 17287:2002 specifies a procedure for assessing whether specific TICS, or a combination of TICS with other in-vehicle systems, are suitable for use by drivers while driving. It addresses user-oriented TICS description and context of use, TICS task description and analysis, assessment process, and documentation.

The TICS description and context of use includes consideration of improper use, reasonably foreseeable misuse and TICS failure. The TICS description, analysis and assessment include a process for identifying and addressing suitability issues.

ISO 17287:2002 does not recommend specific variables for assessing suitability nor does it define criteria for establishing the suitability of use of a TICS Table while driving.

ISO 15007-1:2002, Road vehicles -- Measurement of driver visual behaviour with respect to transport information and control systems -- Part 1: Definitions and parameters

ISO/TS 15007-2:2001, Road vehicles -- Measurement of driver visual behaviour with respect to transport information and control systems -- Part 2: Equipment and procedures

This Technical Specification gives guidelines on equipment and procedures for analyzing driver visual behaviour, intended to enable assessors of Transport Information and Control Systems (TICS) to

- Plan evaluation trials
- Specify (and install) data capture equipment and
- Analyse, interpret and report visual-behaviour metrics (standards of measurement)

It is applicable to both road trials and simulated driving environments. It is not applicable to the assessment of head-up displays.

ISO/TR 16352:2005, Road vehicles -- Ergonomic aspects of in-vehicle presentation for transport information and control systems -- Warning systems

ISO/TR 16352:2005 provides a literature survey about the human-machine interface of warning systems in vehicles. It covers the experimental experiences about the efficiency and acceptance of different modalities and combinations of warnings, and the design of the sensorial, code and organizational parameters of visual, auditory and tactile warnings.

ISO 16673:2007, Road vehicles -- Ergonomic aspects of transport information and control systems -- Occlusion method to assess visual demand due to the use of in-vehicle systems

ISO 16673:2007 provides a procedure for measuring visual demand due to the use of visual or visual-manual interfaces accessible to the driver while the vehicle is in motion. It applies to both Original Equipment Manufacturer (OEM) and After-Market in-vehicle systems. It applies to both permanently installed and portable systems. It applies to any means of visual occlusion and is not dependent on one specific physical implementation.

Standards for road vehicles and limitation of driver distraction are also developed by the Society of Automotive Engineers International (SAE). Their technical reports play a key role in market access, safety, reducing costs, increasing productivity, improving market position and advancing new technologies (SAE, 2007).

Turning to guidelines, the Japan Automobile Manufacturers Association (JAMA) guideline for in-vehicle display systems is one of the strictest. While the vehicle is in motion several functions/ features concerning visual information are prohibited (JAMA, 2004, p.5 et seq.):

- Displaying navigation maps if the driver is confused when the maps are automatically scrolled in keeping with the speed of the vehicle
- Showing minor roads in urban areas on maps displayed for navigation purposes
- Displaying addresses and telephone numbers as guiding information
- Displaying information describing restaurants, hotels and other similar facilities
- Motion pictures (TV, video and DVD)
- Scrolling of letters
- Text exceeding an amount of 31 letters displayed at a time

Regulations differ from country to country, but all of them aim at decimating the distracting potential of in-car telematics devices. Technology nowadays becomes more and more complex. Address books can be compiled from different sources leading to an immense amount of address book entries and soon the same will hold for audio data in cars. But does a trend towards rising complexity necessarily mean that handling in-car dialogue systems has to become more complex and frustrating, too? The point is to design a system that unifies both the big variety of features and input data as well as user-friendliness. A user-friendly system implies that different requirements have to be met for two kinds of users: the novice user needs to be prevented from getting lost and frustrated while using in-car human-computer interaction while simultaneously the expert user needs to be offered additional options that enable him to accomplish a task as effectively and efficiently as possible. These aspects are focused in this thesis.

Chapter 3

Communication

The best automatic speech recognition, the ultimate microphone array and the best speaker would not be sufficient to achieve good system performance if dialogue design is poor. It is crucial for successful interaction. Literature provides designers of speech dialogues with plenty of guidelines, yet there exists no established theory on how these systems can be specified and evaluated which makes designing speech dialogue systems that are usable an art (Hulstijn, 2000, p.1). In order to understand the challenges dialogue design for human-computer interaction has to meet, it is indispensable to examine the nature of human-human dialogue:

- What is a dialogue?
- What kind of knowledge is necessary to cooperate successfully?
- What are the underlying principles of communication?

Having considered its complexity and variety of influential factors it can then be checked which mechanisms of human-human dialogue may be replicated for human-computer dialogue. Besides, it can also be verified if the crucial aims speech dialogue research pursues, i.e. naturalness and freedom of communication, are indispensable to user-friendly human-computer interaction.

3.1 Dialogue

As the term dialogue is crucial for this thesis, it is necessary to first of all give a definition of it, including the terms that constitute a dialogue. According to Bußmann (1983; also see Kölzer, 2002, p.37) dialogue is a

Wechselrede zwischen mindestens zwei Sprechern. Ursprünglich war der Dialog vor allem von literaturwissenschaftlichen Interessen, dagegen richtet sich in neueren Untersuchungen zur Sprechakttheorie, Argumentation und Konversationsanalyse das Interesse einer an Pragmatik, Textlinguistik u.ä. orientierten Sprachwissenschaft auf den Dialog als Grundeinheit jeder Form von Gespräch.

Communication is a dialogic event based on an alternation between two or more parties³, i.e. dialogue participants. This event is referred to as turn-taking. What constitutes a turn is

the talk of one party bounded by the talk of others [...], with turn-taking being the process through which the party doing the talk of the moment is changed (Goodwin, 1981, p.2).

Number and length of turns are variable. Parallel to taking turns communication undergoes a constant process of providing feedback. Feedback from the hearer that he has understood what the speaker uttered. This process of achieving mutual understanding is referred to as grounding (McTear, 2004, p.54). Feedback may occur prior to taking turns implicitly (e.g. by nodding the head or keeping eye contact) or explicitly (e.g. by saying yes, ok etc.) as long as the hearer is listening to the speaker or once turns have been taken in a more extended form.

3.2 Discourse analysis

Vladimir: On attend Godot.

Estragon: C'est vrai. (*Un temps.*) Tu es sûr que c'est ici?

Vladimir: Quoi?

Estragon: Qu'il faut attendre.

Estragon: Tu es sûr que c'était ce soir?

Vladimir: Quoi?

³ In this thesis the term party may refer to humans as well as machines. Consequently the terms dialogue and communication are valid for both human-human interaction and human-computer interaction.

Estragon: Qu'il fallait attendre?

Vladimir: Attendons voir ce qu'il va nous dire.

Estragon: Qui?

Vladimir: Godot.

Vladimir: J'ai cru que c'était lui.

Estragon: Qui?

Vladimir: Godot.

(Samuel Beckett, En attendant Godot)

What makes a dialogue a dialogue? As it has already been shown in Chapter 1, speech as the most natural form of interpersonal communication is not only restricted to combining phonemes into morphemes, morphemes into words, and words into sentences. What makes it an ideal means of communication is the ability of combining sentences together to larger entities in order to convey complex ideas, thoughts and make inferences, the interpretation of which needs to be deduced by integrating dialogue context (cf. McTear, 2004, p.46). To enable successful interpretation dialogue context must be coherent (also see Bellert, 1970; Harris, 1952, Chapter 6). This means that sentences and their components have to relate to each other on a grammatical and semantic level (Bußmann, 1990, p.389). In case coherent context is not available, dialogue crashes as can be seen in the above passages from Samuel Beckett's theatre of the absurd 'En attendant Godot' (Müller, 2003, p.195): contextual elements such as anaphors can no longer be used to referring back to something that has already been introduced (also see Fox, 1993). Instead they need to be focussed anew time and time again. Repetitions, pro-forms, ellipsis, causal or temporal connexions also lose their function of creating coherence. Coherence is not only important on the level of connected speech but also when it comes to interacting on a multimodal level. Care must be taken that all modalities are synchronised with each other.

3.2.1 Conversational principles

When it comes to analysing communication, i.e. tasks that are completed in a dialogue, it is not sufficient to focus on the 'speaking party' solely. Instead, what must be focused are the efforts two kinds of communication partners (i.e. speaker *and* hearer, sender *and* recipient etc.)

undertake to interact successfully, namely to be cooperative (Grice, 1989, p.26). Following a particular purpose they take turns, provide feedback and refer to what they said in previous statements. Grice (1989, p.26) subsumes this cooperative effort under the cooperative principle:

- Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

The cooperative principle is the underlying principle of a set of guidelines Grice considered indispensable for efficient and cooperative interaction: the maxims of quantity, quality, relation and manner (Grice, 1989, pp.26-27; also cf. Harris, 2005, Chapter 4).

The Maxims of Quantity:

- Make your contribution as informative as is required (for the current purposes of the exchange).
- Do not make your contribution more informative than is required.

The Maxims of Quality: “Try to make your contribution one that is true”

- Do not say what you believe to be false.
- Do not say that for which you lack adequate evidence.

The Maxim of Relation: “Be relevant”.

The Maxims of Manner: “Be perspicuous”.

- Avoid obscurity of expression.
- Avoid ambiguity.
- Be brief.
- Be orderly.

None of them is sufficient when applied solely; instead they need to interact according to the situation. If someone asks “Is there a train from Stuttgart to Paris tomorrow at 4 pm?” the reply “no” might be correct. Having checked the time-table, the person being in charge of the reply has adequate evidence for saying no. The maxim of quality is fulfilled. But is the reply as

informative as is required? Taking a closer look at the time-table shows there is a direct train from Stuttgart to Paris at 4.30 pm. So the reply is not sufficient as the alternative train has been withheld from the asking person. The maxim of quantity has been violated.

Grice's maxims are a theoretical construct and it is clear that people do not exclusively follow these guidelines while communicating (Levinson, 1983, p.102). Nevertheless they are guidelines that help people orientate when communication partners make unexpected utterances (that at first sight might even seem to have violated the one or the other principle) to the effect that proper inferences, i.e. implicatures, can be made (Grice, 1989, Chapter 2, cited in McTear, 2004, p.51; Levinson, 1983, p.102). This orientation takes place subconsciously. Grice's maxims therefore present indispensable rules when it comes to designing dialogue to take place between human and machine:

Cooperation in terms of quantity, involves coordinating the amount of information (not too much, not too little, "j-u-s-t right"); in terms of quality, cooperation involves trafficking only in reliable information; in terms of relevance, cooperation is a matter of sticking to the topic; in terms of manner, cooperation concerns the clarity of the message, how easily the hearer can figure out what it means (Harris, 2005, p.121).

3.2.2 Speech acts

Als ob mit dem Akt des Benennens schon das, was wir weiter tun, gegeben wäre. Als ob es nur Eines gäbe, was heißt: 'von Dingen reden.' Während wir doch das Verschiedenartigste mit unseren Sätzen tun.

(Ludwig Wittgenstein, Philosophische Untersuchungen)

Grice's conversational maxims focus on how two parties interact cooperatively. But what is people's intention when communicating via speech? As Wittgenstein (Schulte, 2003, p.28) already put it in his *Philosophical Investigations* from 1953 the intention is not simply to convey meaningful utterances. Instead people communicate for a certain purpose. The sentence "It is cold in here!" uttered by speaker A might lead speaker B to confirm it by saying "Yes!" which would be an inappropriate response though. Most likely however it will be a trigger for speaker

B to *act* such that speaker A's actual state of being cold is changed, for example either by closing any open windows or by turning on the heating. Speaker A could have chosen different syntactic forms such as questions (e.g. "Is it always that cold in here?") or imperatives (e.g. "Turn on the heating, please!") – even a non-verbal chattering of teeth – but the request to change something about the state of being cold remains the same (also cf. Harris, 2005, p.94; McTear, 2004, p.49). Thus, when people communicate they do not necessarily respond to the grammatical structure but to what was *intended* by the utterance. An act like the above request is referred to as dialogue act⁴. Dialogue acts represent

the smallest functional units of dialogues, and are utterances corresponding to speech acts such as 'greeting', 'request', 'suggestion', 'accept', 'confirm', 'reject', 'thank', 'feedback'. When considering the overall communicative function of dialogs, it is as well to bear in mind that for annotation as well as for processing purposes, they are seen as decomposable into such basic communicative units (Gibbon, 2000, p.6).

The term speech act was invented by Austin (1962) and Searle (1969). Based on the theories of the late Wittgenstein they developed the speech act theory, analysing what people actually *do* in saying something (Bußmann, 1990, p.726). In "How to Do Things with Words" Austin's theory first of all concentrated on a particular group of utterances, so-called performatives (Austin, 1962, p.4). Verbs like promise, bet, name and many more belong to the class of performative verbs. To give a few examples:

- I command light to shine.
- I name this boat the 'Mary Ann'.
- I nominate him vice chancellor.
- I promise to walk the dog three times a day.
- I give in.

⁴ The term dialogue act goes back to Bunt (1979) (McTear, 2004, p.60). The more common term is speech act by Austin and Searle (Harris, 2005, p.94). During the last century linguists and philosophers have come up with a variety of alternative terms such as communicative act, conversational move or dialogue move (McTear, 2004, p.60).

What is special about these utterances is not simply the fact of saying something but instead the performance of an action changing circumstances in the world (Austin, 1962, p.6; also cf. Levinson 1983, p.228). According to Austin performative sentences

- (a) [...] do not ‘describe’ or ‘report’ or constate anything at all, are not ‘true or false’; and
- (b) the uttering of the sentence is, or is a part of, the doing of an action, which again would not *normally* be described as, or as ‘just’, saying something (Austin, 1962, p.5).

Austin opposed performative utterances to constative utterances such as statements and assertions. In the course of his work however, Austin extended what he had initially attributed to performative verbs to all kinds of utterances (Austin, 1962, p.94; Levinson, 1983, p.231, 235). This resulted in what is referred to as speech act theory according to which *each* utterance is regarded as some kind of speech act. A speech act in turn was found to be dividable into three different senses, i.e. acts (see Table 3.1).

The locutionary act is executed when producing an utterance. According to Austin it includes uttering certain noises (phonetic act), uttering certain words according to a language-specific grammar (phatic act) and uttering them with a certain ‘meaning’ (rhetic act), i.e. with a certain sense and with a certain reference (Austin, 1962, p.95).

The illocutionary act concerns the communicative meaning of an utterance. It is performed additionally to a locutionary act, like for example by asking a question, expressing a warning, criticizing someone or something or announcing an intention (Austin, 1962, p.98). Austin also refers to it as illocutionary force, performing an act *in* saying something as opposed to performing a locutionary act *of* saying something (Austin, 1962, p.99).

To infer the proper illocutionary force an utterance may contain various indications (Austin, 1962, p.73 et seqq.). Apparent indications within an utterance in written form may be its grammatical mood, e.g. whether it is indicative, conditional, imperative etc; adverbs and adverbial phrases such as ‘you would do well to keep that in mind’; and connecting particles such as ‘therefore’ to ‘conclude that’.

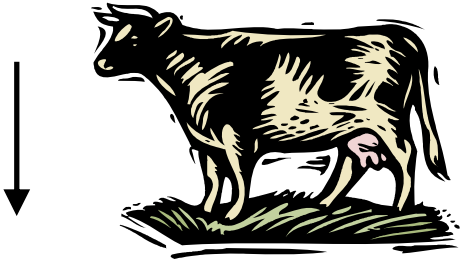

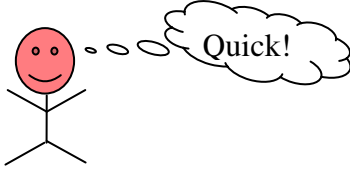
Locutionary act	Act of saying something: He said that ...
- Phonetic act (articulatory level)	/kau/, /maedou/, /Iz/
- Phatic act (syntactic level)	The cow is standing in the meadow.
- Rhetic act (semantic level)	He said that the cow was standing in the meadow.
	
*The meadow is standing on the cow.	
Illocutionary act	Communicative meaning: He warned you that ...
	The cow is standing in the meadow!
Perlocutionary act	Speaker's intention: He roused me.
	Causes the hearer to hurry to get the cow back into the cow barn. Obviously he must have forgotten to lock it properly.

Table 3.1: Structure of speech acts according to Austin (1962)

What may only be reflected in spoken language are indications through suprasegmental features such as tone of voice and emphasis to signalize whether an utterance is a request or a warning; as well as the non-verbal part going along with an utterance, i.e. gestures. These are indications concerning the occasion of an utterance, but in addition an essential aid is the context an

utterance is embedded in (Austin, 1962, p.100). The utterance ‘The cow is standing in the meadow’ may then be interpreted accordingly, whether it is an imperative warning to ‘hurry and get it back into the cow barn’ or an exclamation of relief as the animal has been missing.

Whether the speaker was actually reassuring or rousing the hearer remains unanswered up to that point. This is where the third speech act comes in, the so-called perlocutionary act.

Saying something will often, or even normally, produce certain consequential effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons: and it may be done with the design, intention, or purpose of producing them (Austin, 1962, p.101).

Illocutionary forces can be made explicit by the performative formula: one can say ‘I argue that’ or ‘I warn you that’ (Austin, 1962, p.103). Austin speaks of the ‘use of “language” for arguing or warning’. This formula does not hold for perlocutionary acts, ‘the use of “language” for persuading, rousing, alarming’. It is not possible to say ‘I convince you that’ or ‘I rouse you that’ (Austin, 1962, p.104). So there is the speaker’s intention on the one hand and the achievement of his intention on the other.

In contrast to Austin, Searle differentiates between four kinds of simultaneous speech acts (Searle, 1969). What Austin subsumes under the term locutionary act, is split in two separate speech acts by Searle: the locutionary act and the propositional act (Bußmann, 1990, p.465, 616, 727). The locutionary act is performed with the utterance of words (phonetic act) according to language-specific rules (phatic act). The propositional act is a common denominator concerning the variety of illocutionary types. It is identical, regardless of whether the speaker utters ‘Sam smokes habitually.’, ‘Does Sam smoke habitually?’, ‘Sam, smoke habitually!’ or ‘Would that Sam smoked habitually.’ (Searle, 1969, p.22).

In uttering any of these the speaker *refers to* or mentions or designates a certain object Sam, and he predicates the expression “smokes habitually” [...] of the object referred to. Thus we shall say that in the utterance of all four the reference and predication are the same, though in each case the same reference and predication occur as part of a complete speech act which is different from any of the other three (Searle, 1969, p.23).

With regard to speech dialogue systems, interaction between human and machine may comprise three types of syntactic constructions (Harris, 2005, p.95):

1. Assertions:

- The phone number is 0 7 3 1 5 6 7 2 3 4.
- Address book entry could not be found.

2. Imperatives:

- Read out points of interest.
- Enter destination.

3. Interrogatives:

- Your destination is Frankfurt am Main, Lange Straße. Is this correct?
- What is the estimated time of arrival?

Taking a look at the examples concerning interrogatives it becomes clear that it is crucial to take into account the illocutionary force (or dialogue act) of an utterance when designing dialogue.

Dialogue acts

exploit the core taxonomy of interrogative, imperative, and assertive, but do so for a range of communicative activities far broader than that shallow and formal taxonomy suggests. Interrogatives are always interrogatives, imperatives are always imperatives, but their syntactic needs are often satisfied tacitly while the conversants respond more directly to higher-level dialogue acts (Harris, 2005, p.100).

Consider the example “Your destination is Frankfurt am Main, Lange Straße. Is this correct?”. Syntactically, the final sentence is a yes/no question. So at this stage a dialogue system theoretically could be designed such that it only allows yes or no as possible commands. Going one step further a likely user reply instead of ‘yes’ could be “start navigation”. In this case the user would ignore the response to the form of the interrogative and respond directly to the intention behind it (Harris, 2005, p.100). Thus the number of interaction steps can be reduced, saving time required for accomplishing a task.

In human-human communication dialogue acts are manifold. Searle (1979, p.20; also see Levinson, 1983, p.240; Searle, 1969, p.66) splits them up into five kinds of illocutionary forces:

1. Assertives, committing the speaker S to the truth of the proposition p, e.g. assert, state, affirm
2. Directives, with which the speaker S tries to get the hearer H to do something, e.g. to request, question, command
3. Commissive, committing the speaker S to a future course of action, e.g. to promise, offer
4. Expressives, expressing a psychological state, e.g. to thank, greet, congratulate
5. Declarations, changing the state of an entity within extra-linguistic institutions, e.g. to baptize, espouse, declare ceasefire

Dialogue act	Examples for in-car speech dialogue systems
Assertives	<ul style="list-style-type: none"> - State, describe, assume, guess, claim, opine, announce, insist, notify - Answer, confirm, correct, negate
Directives	<ul style="list-style-type: none"> - Constituent question, yes/no question, ask-if, inquiry, request - Command, direct, instruct - Advise, recommend, suggest, propose - Present - Caution, warn - Accept-of - Halt, pause - Resume - Repeat - Skip - Jump, jump-back - Configure
Commissives	<ul style="list-style-type: none"> - Offer - Book - Accept-to, agree-to - Reject

Table 3.2: Overview of dialogue acts for SDS (adapted from Harris, 2005, p.104)

To be able to successfully accomplish tasks in speech dialogue systems three kinds of illocutionary acts are obligatory, in accordance with the above three types of syntactic constructions by Harris. These are assertives, directives and commissives. Optionally speech

dialogue systems may be extended by the factor of politeness (André, 2004). This would imply integrating expressive speech acts such as greet, welcome, thank, excuse, congratulate etc. Table 3.2 presents a minimum overview of dialogue acts required for speech dialogue systems.

3.2.3 Presuppositions

Presuppositions also play an important role in discourse analysis. Presuppositions are implicit assumptions made in context with utterances that must be mutually known by speaker or hearer (Bußmann, 1990, p.600; SIL International, 2004). Take a look at the following sentences:

1. Who took the last piece of chocolate cake?
2. Have you finished doing the washing up?
3. Rachel can no longer play the guitar.

In the first sentence the speaker presupposes that there once was the final piece of a chocolate cake. The speaker of the second sentence presupposes that at some time in the past the hearer has started washing up. In the third sentence the speaker states that Rachel once used to play the guitar (cf. Fromkin, 1993, p. 161).

Presuppositions can generally be associated with a specific lexical item or grammatical feature in the utterance (SIL International, 2004). The presupposition “Rachel used to play the guitar” is tied to “no longer”. The presupposition “The chocolate cake is all gone” arises from “last piece”. Levinson (1983, p.179) refers to these items as presupposition triggers.

Now the question arises how presuppositions may be differentiated from implications. The presupposition of an utterance is what the speaker considers to be true with respect to this utterance (Herbst, 1991, p.169). It remains true even in case an utterance is negated. This observation goes back to Strawson (1952). Strawson states the concept of presupposition as follows (cited in Levinson, 1983, p.175):

Statement S presupposes another statement P iff:

- (a) if S is true, then P is true
- (b) if S is false, then P is still true

The sentences S(a) “It is sad that nowadays many fish are threatened by extinction” and S(b) “It is *not* sad that nowadays many fish are threatened by extinction” for example both presuppose P “Nowadays many fish are threatened by extinction” (cf. Herbst, 1991, p.169). Independent of the aspect whether it is sad or not, the speaker’s uttering of S(a) or S(b) presupposes that P is true.

3.2.4 Deixis

In dialogue speakers articulate strings of letters (signifié) to *refer* to objects of the world (signifiant) – the relation between sound and meaning is arbitrary (Saussure, 1975; also see Schwarz, 1996, p.22). However, dialogues contain numerous elements the references of which may not merely be deduced by means of their inherent meaning (denotation) but by taking into account coherent context (Ehrlich, 1990, p.11). In the example ‘look at that tree’ the reference of the object tree may only be ascertained by its denotation *and* a particular context in which the focus is directed towards one special tree with particular features (see Figure 3.1). The utterance ‘look at that tree’ might be an utterance of pleasure, astonishment or shock and accordingly the object ‘tree’ might be beautiful, gigantic, rotten, etc.

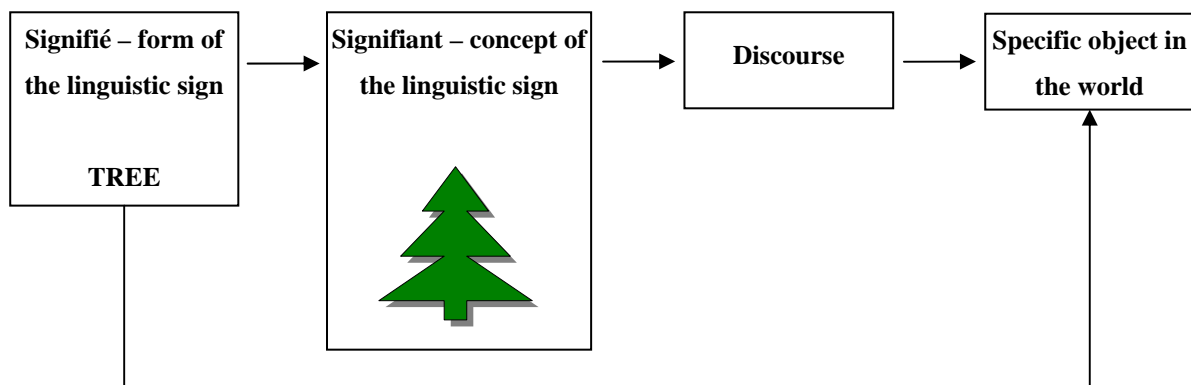


Figure 3.1: Reference of linguistic signs

References can be divided into deictic and non-deictic expressions (Levinson, 1983, p.68). Deictic references are words or expressions establishing reference to the situational context (Ehrlich, 1990, p.12; Herbst, 1991, p.182). They have pointing character, going back to Bühler (1934) and may be divided into:

- Person deixis: in the course of a dialogue speaker and hearer do not remain static but alternate (Levinson, 1983, p.68). To distinguish between speaker and hearer of an utterance pronouns are used in first (e.g. I, my, we, our), second (e.g. you, your) or third (e.g. he, she, they) person.
- Time deixis: to correctly deduce the meaning of deictic time adverbs such as now, yesterday, tomorrow, last week etc., it is necessary to know when an utterance was made (Fromkin, 1993, p.163). Accordingly, time deixis refers to the speaker (Levinson, 1983, p.73). The adverb 'now' for example refers to the point of time the speaker produces the utterance containing 'now'.
- Place deixis: adverbs such as here, there or over there refer to locations relative to anchorage points in the utterance (Levinson, 1983, p.79). Objects can be referred to either by describing or naming them or by locating them in relation to others.

Demonstrative articles such as this, that, these, those can be used for all three types of deixis, e.g. this woman, that day, these buildings.

Non-deictic references are words or expressions that refer to other linguistic signs within a dialogue (Ehrlich, 1990, p.12). In this context it is necessary to differentiate between anaphoric and cataphoric usage. According to Bußmann (1990, p.82) anaphors refer to the linguistic sign (antecedent) of the previous context. They are pronouns (e.g. Peter has bought a house. *He* thinks it was a bargain.), ellipsis (e.g. Mary likes cooking, Andy [likes cooking] too.) and the empty category PRO (e.g. Peter promised his boss [PRO] to be punctual.).

Unlike Bußmann, Roberts (1997, p.149) clearly differentiates between anaphors and pronouns. Anaphors are reflexives (like for example himself, myself, themselves) and reciprocals (each other). Pronouns are personal pronouns such as I, me, you, he, she, him, her, they etc. Both types, i.e. anaphors and pronouns stand for another determiner phrase (DP). However, in contrast to anaphors, pronouns do not necessarily require antecedents, although they can have them. In the above example "Peter has bought a house. He thinks it was a bargain." the pronoun *he* may either refer to the subject Peter introduced in the first sentence or to another person derived from contextual interpretation. Anaphors always require antecedents, i.e. they are always bound whereas pronouns are unbound or free (also cf. Haegeman, 1998).

Cataphors refer to information following in the course of the dialogue (Bußmann, 1990, p.372). They are determiners (e.g. *This* is my favourite song.) and pronouns (e.g. At the time *she* arrived, Vanessa was very ill.).

Chapter 4

Human-computer communication with in-car speech dialogue systems

The word *glamour* comes from the word *grammar*, and since the Chomskyan revolution the etymology has been fitting. Who could not be dazzled by the creative power of the mental grammar, by its ability to convey an infinite number of thoughts with a finite set of rules?

(Steven Pinker, The Language Instinct)

Having taken a look at human-human communication in the previous chapter this chapter now focuses on what is required to transfer speech communication between humans to man-machine interaction: what is a speech dialogue system (SDS) and what are the components comprised by state-of-the-art technology? Various user studies on automotive applications (see Chapter 4, 4.3) have been made to analyse in how far current SDS are limited and how they can be improved taking into account relevant aspects from human-human communication.

4.1 Architecture and functions of a speech dialogue system

A speech dialogue system is an interface that allows communication between a human being and a machine (Wahlster, 2006; Reithinger, 2005; Wahlster, 2007). According to the definition given in Chapter 2, 2.2 a usual in-car speech dialogue system is a multimodal interface, with two input

and two output modalities (at least): manual and speech input and graphical and speech output. In order to accept spoken input, understand and process it and answer appropriately while simultaneously synchronising spoken interaction with graphical output, several components have to interact successfully: speech input and output combined with a dialogue management system as well as a synchronisation component to exchange parameters between speech control and the graphics/haptics side of the user interface. Figure 4.1 presents the typical architecture of a speech dialogue system.

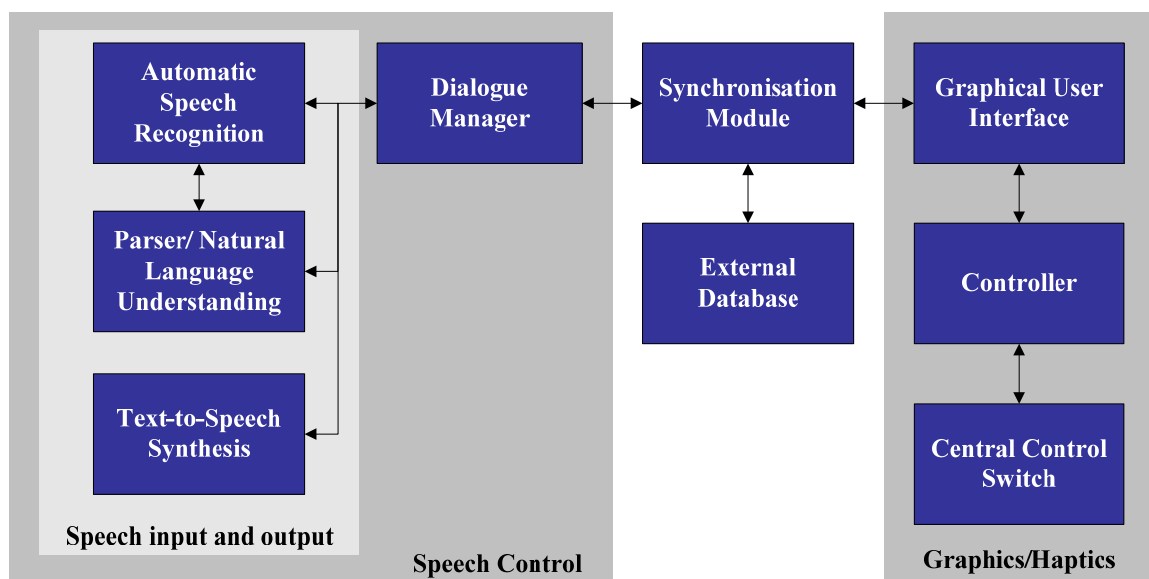


Figure 4.1: Architecture of a multimodal in-car speech dialogue system

Speech control

Speech control consists of a speech input and output module and a dialogue manager. The module for speech input and output is responsible for recognising spoken input such as commands, digits and spelling and also, for turning system output into speech. Accordingly the module comprises two components: a first module for automatic speech recognition, the results of which are passed on to a parser or a unit for natural language understanding, and a second module for text-to-speech synthesis.

Automatic speech recognition

To be able to recognise spoken input, an automatic speech recogniser (ASR) contains a lexicon and a language model. The lexicon comprises all words including their acoustic models (Hidden-

Markov Models, see e.g. Schmandt, 1994; Gibbon, 1997) the user is allowed to say (see Figure 4.2). In case of natural language input the lexicon also includes morphosyntactic features, e.g. form, category, subcategorisation, case, number and gender.

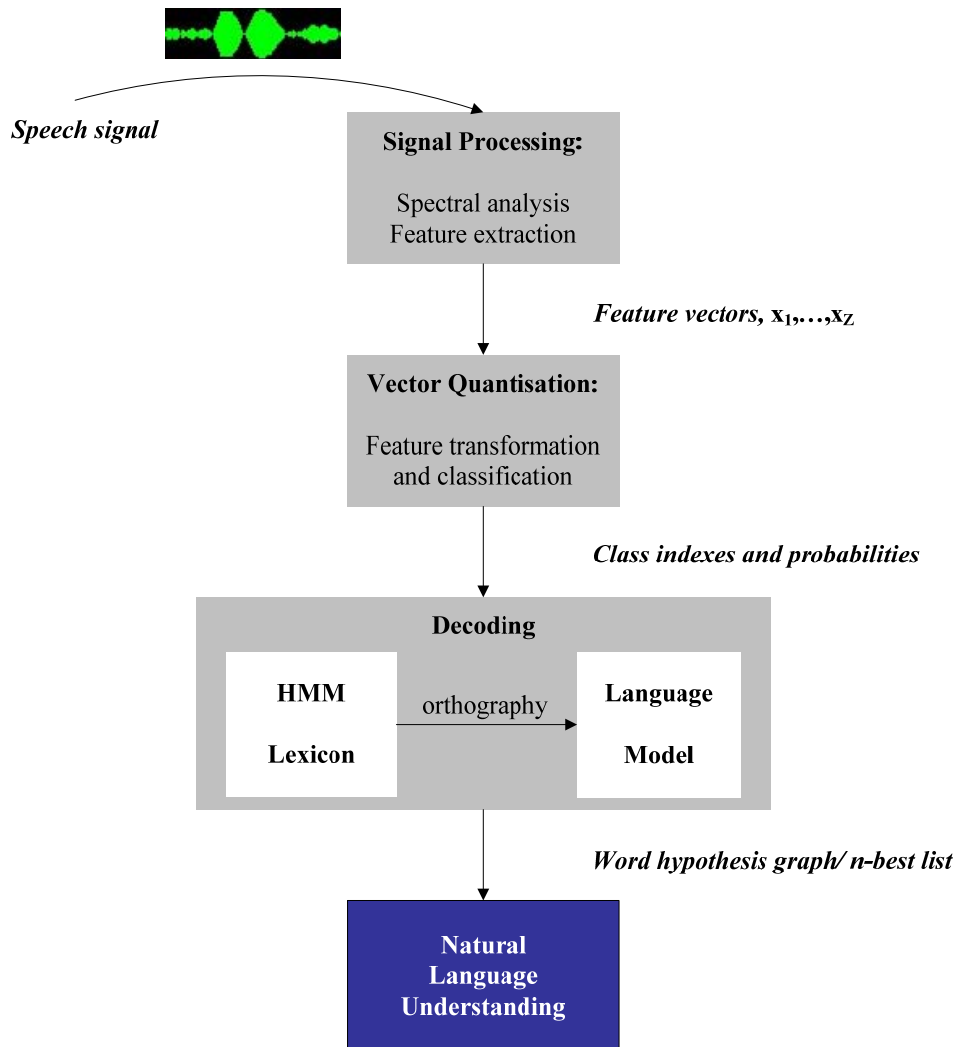


Figure 4.2: Process of automatic speech recognition (adapted from Berton, 2004, p.14)

The language model (LM) can either be a grammar comprising all possible word sequences or, in case of less restricted input, a statistical model (McTear, 2004, p.86).

The grammar describes how lexical entries may be combined to phrases and sentences. The syntax of the grammar format is presented in Augmented Backus-Naur Form (ABNF) (Hunt, 2004; Hunt, 2000). To account for the productivity of language, so-called rules in ABNF are applied in a grammar. ABNF rules are structured hierarchically like a tree diagram.

The statistical model is based on N-grams, reducing context to a maximum number of N words. Trigrams as in the example “Ich möchte in die Olgastraße“ with the probabilities $P(\text{in} \mid \text{ich möchte})$, $P(\text{die} \mid \text{möchte in})$ and $P(\text{Olgastraße} \mid \text{in die})$ are not precise but robust with regard to spoken language input, e.g. when it comes to recognising hesitations, restarts, ungrammatical utterances etc.

To recognise spoken input automatic speech recognition applies a complex search algorithm (see Figure 4.2). During the first step called signal processing the incoming digitalised speech signal is split into e.g. 10ms frames (McTear, 2004, p.83). By means of Fourier-Transformation each frame can be analysed according to particular features that describe the frame’s spectral energy distribution. These features in turn are subsumed under a feature vector. To reduce complexity of the represented signal, feature vectors x_1, \dots, x_Z undergo an additional process of feature transformation and classification, resulting in class indexes and probabilities (Berton, 2004, p.14). The results of this process are passed to a decoder. In the decoder the use of a statistical language model aims at determining the most probable transcriptions (e.g. Hamburg, Homburg, Humburg) given a spoken utterance, i.e. a word hypothesis graph (Berton, 2004, p.14; also see Kuhn, 1996 and Jelinek, 1990). Alternatively, the use of a grammar results in an n-best list that is passed on to the component for Parsing/ Natural Language Understanding.

Natural language understanding

The module for natural language understanding (NLU) gets the result of the speech recogniser. Its task is to analyse the recognition result such that the system is able to ‘understand’ what has been said by the user and deal with it accordingly. The analysis process is two-stage (Fromkin, 1993, p.483 et seqq.; McTear, 2004, p.91 et seqq.). In stage one syntactic analysis or parsing splits a user utterance into its constituents. A constituent analysis of the sentence “I want to go to Munich” for example would result in a phrase marker as presented in Figure 4.3.

To ascertain phrase markers the module requires access to two knowledge bases:

- The lexicon, comprising speakable words including morphosyntactic features, semantic features (e.g. [+ direction] or [+ location]) (Schwarz, 1996) and rules for combining meanings on the basis of thematic roles (e.g. whether a noun in a particular context has to be an agent or theme).

- The grammar knowledge base, describing possible combinations of lexical entries resulting in phrases and sentences. The rules could for example be phrase structure rules (e.g. $IP \rightarrow NP\ I'$; $I' \rightarrow I\ PP$; $PP \rightarrow P'\ VP$ etc.) (cf. Radford, 1988; Crain, 1999). Phrase structure rules reflect the domination relation within a sentence, categorising nodes that dominate other nodes, until the lowest nodes, i.e. the final nodes of a tree diagram, are reached (Crain, 1999, p.91). They are recursive rules from which phrase markers like the following can be deduced.

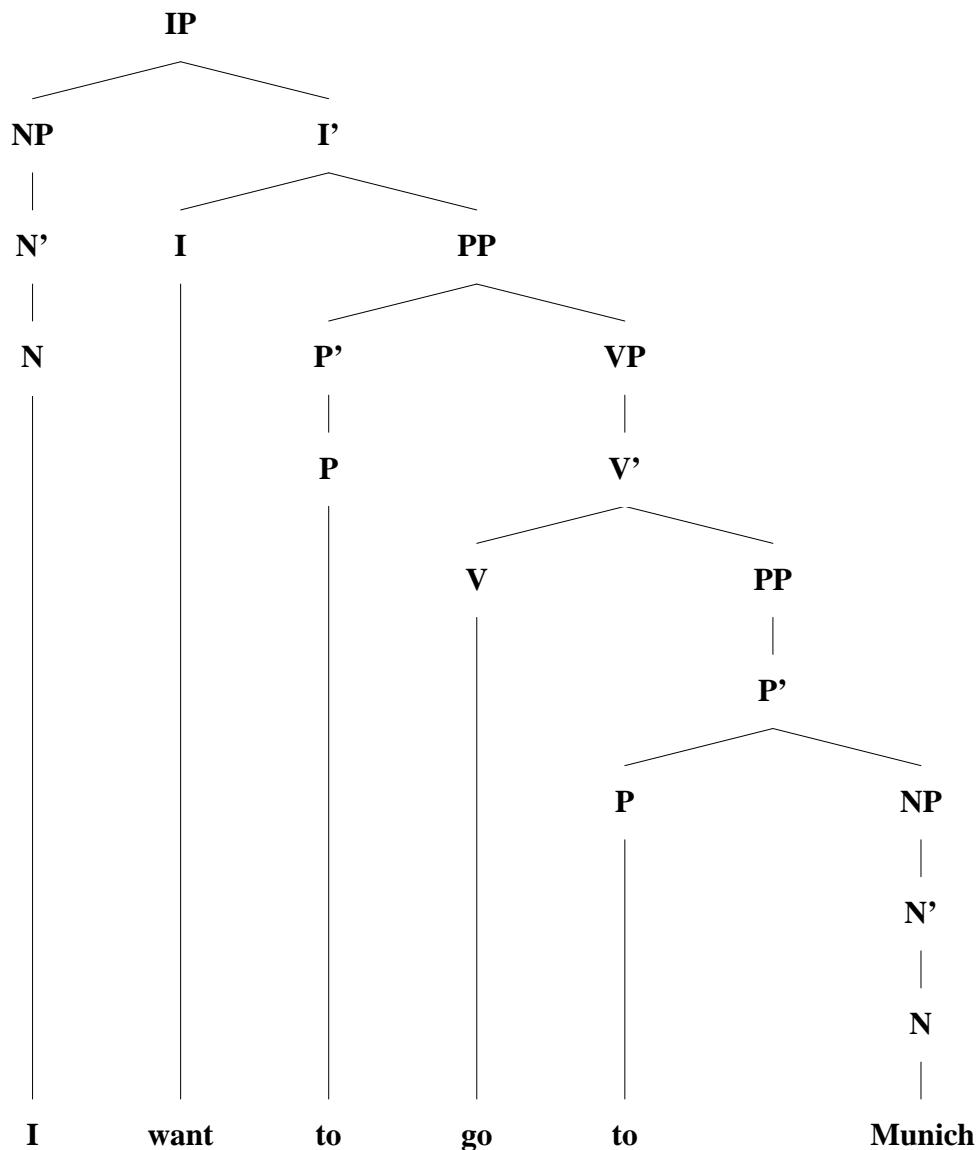


Figure 4.3: Phrase marker of the sentence “I want to go to Munich.”

The syntactic structure of an utterance is a prerequisite for deducing the overall meaning of an utterance. For example the word ‘play’ has two different functions and meanings in the verb

phrase “play music” and in the noun phrase “Shakespeare play”. Considering these utterances in the context of a speech dialogue system the latter ‘play’ is not a direct command to the system for playing any kind of music. Instead it represents a part of an audio item the user wants to select. As soon as an utterance has been parsed, the structural combination of words becomes subject to semantic analysis. This process requires the lexicon for accessing the semantic features of words and their defined combinations. Having located the verb ‘go’ in a sentence like for example “I want to go to Munich” a semantic representation as presented in Figure 4.4 could be ascertained.

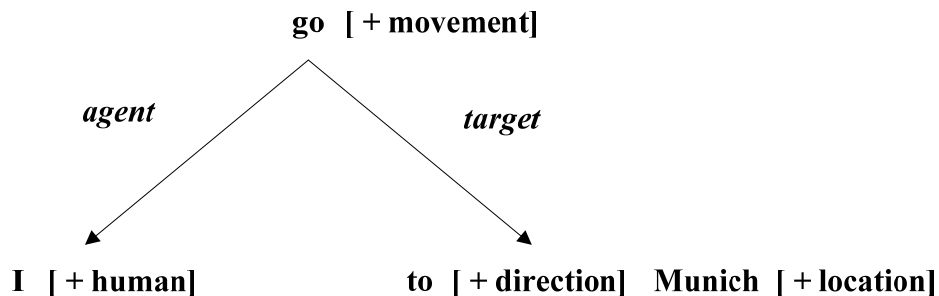


Figure 4.4: Semantic representation of the sentence “I want to go to Munich.”

Dialogue manager

As its name already implicates, the dialogue manager (DM) is in charge of controlling the dialogue. Depending on what was entered by the user this module determines a corresponding system reaction and/or system prompt and is responsible for interacting with external modules from the outside world (see Figure 4.5). Examples may be to

- Change applications, e.g. from navigation to audio (4,1)
- Request the user to enter information, e.g. an artist or point of interest (speakable database entries) (3, 1, 6)
- Access a database, e.g. to check all titles (2)
- Request the user to give additional information, e.g. “Which artist?” in case the title is ambiguous (3,1)
- Perform an action after user input has been completed and all information necessary is available, e.g. to play a particular title (4)
- Process barge-in and timeout (5)

The digits in brackets refer to the interaction steps illustrated in Figure 4.5.

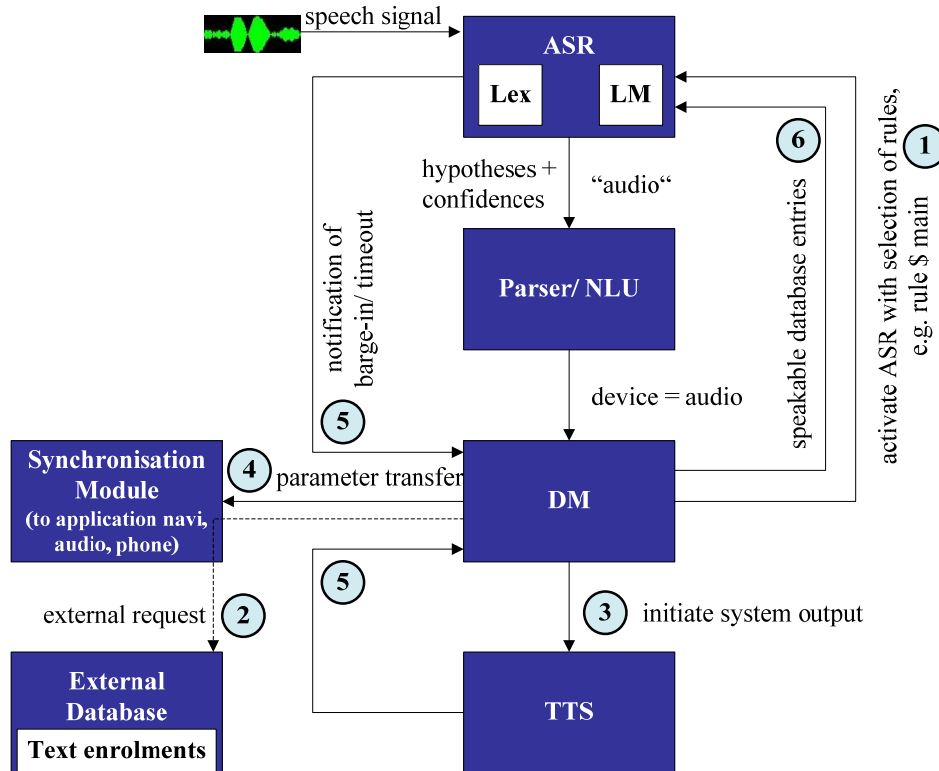


Figure 4.5: Sample tasks of a dialogue manager

Text-to-speech synthesis

Text-to-speech (TTS) synthesis technology synthesises speech from utterances created by a response generator (McTear, 2004, p.102), e.g. to return system prompts or give feedback to the user. The technology is recommended for applications that contain unpredictable data, such as audio or telephone applications. First, the text from the response generator is analysed comprising four steps (McTear, 2004, p.103):

1. Text segmentation and normalisation:

- Splits text into reasonable units such as paragraphs and sentences
- Solves ambiguous markers like for example a full stop that can be used as sentence marker or component of a date or acronym

2. Morphological analysis:

- Reduces amount of words to be stored by unifying morphological variants
- Assists with pronunciation by applying morphological rules

3. Syntactic tagging and parsing:
 - Determines the parts of speech of the words in the text
 - Permits a limited syntactic analysis
4. Modelling of continuous speech effects to achieve naturally sounding speech:
 - Adjusts weak forms and coarticulation effects
 - Generates prosody, i.e. pitch, loudness, tempo, rhythm and pauses

The second step involves generating continuous speech from the above text analysis (McTear, 2004, p.104). Over the past years enormous advances have been made in the field of TTS synthesis. This is due to a process called concatenative speech synthesis (Cohen, 2004, p.25). According to that a database of recorded sentences is cut into syllables and words. When it comes to outputting speech the corresponding system utterance is produced by concatenating a sequence of these prerecorded segments. Boundaries between segments are smoothed out to make concatenation splices inaudible. As soon as dynamic data from applications such as audio are involved comprising various languages, speech synthesis needs to be supplemented by G2P (grapheme-to-phoneme) conversion (cf. Chapter 5).

Synchronisation module

The synchronisation module (Sync) turns a speech dialogue system into a multimodal system by connecting and synchronising spoken and graphics-haptics world. It stores data coming from the display and hands over the corresponding parameters to the dialogue manager. These parameters comprise the contents of buttons and lists displayed on the screen, the current state of the Push-To-Activate (PTA) button and actions performed by the user (e.g. change of application, abort etc.). The dialogue manager is then able to initiate a particular dialogue the results of which are, after successful recognition, returned to the display via the synchronisation module.

Graphics-haptics interface

The graphics-haptics control follows the model-view-controller paradigm (Reenskaug, 1979). Models (state charts) and views (widgets) are described in the graphical user interface (GUI) module. The controller module contains the event management and the interface (CAN bus) to the central control switch, which can be pressed, pushed and turned. Such a control switch is the typical control element in advanced cars, such as Audi, BMW and Mercedes-Benz.

4.2 Constraints of speech dialogue systems

In-car speech dialogue technology that is currently available on the market is generally based on the architecture described in section 4.1 and offers spoken interaction via command and control. For novice users who are completely unfamiliar with the technology this is a hard-to-learn means of interaction. State-of-the-art in-car speech dialogue technology encounters several problems. One major problem is speech recognition. Recognition errors may occur any time, an uncertainty that makes interaction cumbersome as it is necessary to persistently confirm (or reject) what has been said and understood. Imagine a dialogue like the following:

Example 7:

User: Store name in address book.
System: The number please.
User: 7 – 8 – 4 – 3 – 6 – 9 – 6
System: 2 – 8 – 4 – 3 – 6 – 9 – 6?
User: No, that’s wrong. It’s 7 – 8 – 4.
System: 7 – 8 – 4, please continue.
User: 3 – 6 – 9 – 6.
System: 3 – 6 – 0 – 6?
User: No, it’s 3 – 6 – 9 – 6.
System: 3 – 6 – 9 – 6?
User: That’s correct.
System: The number is deleted, please continue...

To the user this kind of interaction can become quite confusing and frustrating. And novice users in particular tend to think they did something amiss. Taking a look at the number of interaction steps also clearly indicates that the dialogue can become very time-consuming, time that is not available while simultaneously performing the driving task. Schmandt (1994, p.160) splits recognition errors into three different types:

1. Rejection errors occur when the user says something that is valid according to the recogniser’s vocabulary (e.g. the command phrase “store name”) but recognition is unsuccessful.

2. Substitution errors imply that a valid expression entered by the user is assigned to a different expression in the vocabulary. In the above dialogue for example the system interprets the user's consent "that's correct" as "correction" and consequently deletes the number.
3. Insertion errors happen when stimuli other than speech input are recognised. Perhaps when co-drivers are talking or environmental noise increases due to speeding up.

What makes human-machine interaction even more susceptible to recognition errors are additional problems regarding the special in-car situation. They have already been explained in Chapter 2, 2.4, but are subsumed here for simplicity's sake: the noisy environment inside the car, the factor that it is a multitasking environment as well as strongly varying user characteristics due to psychological and physical stress.

The second major problem in-car speech dialogue technology has to face is the nature of speech. The limitations of its characteristics have already been discussed many times (Schmandt, 1994; Balentine, 2001). First, speech is more difficult to process for human beings than written language. Speaking rates range from 175 to 225 words per minute whereas the rate for reading text comprises the double amount of words (350 to 500 words) within the same time (Schmandt, 1994, p.101). Second, speech is temporal. Once an utterance has been made, it is gone (Schmandt, 1994, p.102; Gibbon, 1997, p.82). This is referred to as the so-called persistence problem (Balentine, 2001, p.11). Being focused on the traffic situation or being disturbed by environmental noise, it may easily happen that given information is missed. In an on-going task this might spoil interaction completely, leading the user to do it all over again. In contrast to that a visual display is persistent, giving the user the freedom to decide when to interrupt and when to continue. Also, opposed to speech, visual display states remain in a certain context or discourse that is transparent to the user. He is able to make his decisions within this context. Let us take a list of artist names, for example: the visual list enables the user to quickly process all artist names presented on the screen and to make his choice in the context of all items. Presenting the items by speech, i.e. by reading out item per item, may lead the user to forget what items were spoken at the beginning once the list has been finished. This leads to the third property of speech, the problem of sequential presentation (Schmandt, 1994, p.102; Balentine, 2001, p.11). It is bit by bit, word by word that the user is given information and not in form of a complete chunk. This can make spoken interaction time-consuming and cognitively demanding.

4.3 Collecting usability and speech data

When it comes to designing and developing in-car speech dialogue systems the constraints outlined in section 4.2 have to be taken into account and balanced accordingly. Understanding technology, speech and the nature of their potential problems is helpful for developing dialogue strategies on error recovery and to avoid problems in the first place (Cohen, 2004, p.15). Therefore various advanced cars providing the above architecture combined with command and control were tested and evaluated in the context of this thesis. To find out where usability problems occur the method of traditional user testing was chosen. Jakob Nielsen (1993, p.165) describes this method as follows:

User testing with real users is the most fundamental usability method and is in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the concrete interface being tested.

Users required for the testing consisted of two categories: the novice user on the one hand, who is minimally experienced with in-car speech applications; and the expert user on the other hand, who has already been frequently employing one or more speech dialogue systems inside the car. The aim was to cover both types of customers purchasing the technology to see to what extent they get along with the systems and what needs to be adapted such that a speech dialogue system is compatible and usable for both user types.

The tasks they were given covered all important interaction tasks available in a common in-car speech dialogue system, i.e tasks within the applications navigation, audio, telephone and address book. Tasks were described by means of various scenarios that were read out to the user. Having them read out brings the advantage that the user is unable to stick to a particular text passage when about to fulfill the task. Table 4.1 presents sample tasks including the corresponding instructions⁵.

⁵ The language of the instructions is German. The sample tasks presented here have been translated into English.

Navigation	
<i>User's task is to store a particular destination.</i>	You are out on business to the company ,Halle' in Ulm, Frauenstraße 39. As this is just the beginning of a series of meetings add this address to the system.
<i>User wants the system to navigate to a point of interest (POI).</i>	Having arrived at Ulm Frauenstraße you find nowhere to put your car. To find a solution, please ask the system.
Audio	
<i>User is requested to change the current radio station.</i>	You do not like the music that is being played on the radio. Check out what else is on the radio that is according to your taste.
<i>User's task is to store a particular radio station.</i>	You found a radio station you particularly like. Therefore you want to make sure you can quickly access it any time.
Telephone/ Address book	
<i>User wants to make a phone call.</i>	You want to get hold of your business partner Mr. Sieger from Pecker enterprise. You start with the area code 0711 by mistake. Then change to Mr. Sieger's correct number which is: 0731 505 4121 ⁶ .
<i>User is requested to redial the number.</i>	Mr. Sieger's phone is engaged. Please try again.
<i>User has to store a phone number in the address book.</i>	As you have to call Mr. Sieger regularly make sure that the system remembers his phone number at any time.

Table 4.1: Sample tasks used during in-car SDS evaluation

Many scientists have already come up with requirements speech interfaces should fulfil, for example Nielsen's (2005) ten usability heuristics, Oviatt's (1999) myths of multimodal interaction or Shneiderman's (2004) eight golden rules (see appendix B). The aim of our user testing however was to see what guidelines can be established that explicitly hold for speech dialogue systems in the automotive area. Compared to telephone applications for example, in-car

⁶ Note that complex contents such as addresses or phone numbers were additionally handed over to the subjects in written form.

speech dialogue systems are far more complex as they comprise several applications. These applications in turn have to be integrated into a multimodal system. Rules holding for voice interfaces may have to be re-adjusted when speech and manual interface are combined.

In view of human-human communication the aim is to examine where aspects and guidelines thereof can be transferred to human-computer interaction: where do they make sense? What needs to be derived from human communication to make current speech dialogue technology usable? On the other hand, where do these principles have to be replaced by different guidelines? Is natural dialogue a prerequisite for successful interaction between human and machine? To what extent do users want to communicate with a system in a less restricted way than what is currently offered on the market by means of short commands? It was also verified how users express themselves and put their wishes into phrases or sentences by means of a Wizard-of-Oz (WOZ) experiment. The basic idea behind a WOZ test is

to simulate the behavior of a working system by having a human (the “wizard”) act as the system, performing virtual speech recognition and understanding and generating appropriate responses and prompts (Cohen, 2004, p.111).

The advantage of human recognition is that whatever users might say, their input can be easily understood and processed accordingly. During the experiment the subject was seated in front of a display inside a parked car with running engine. In a separate area the human wizard used a computer with wizard software to simulate the system the subject expects to be interacting with. Figure 4.6 shows the experimental setup.

The wizard software enabled controlling the graphical output of the system as well as the acoustic output of the system (synthesised speech prompts). The human wizard controlled dialogue flow such that there were hardly any differences between a real dialogue system and the simulation. In the same area the test administrator was giving instructions to the subject via microphone. The subject had to accomplish the tasks (again from the applications navigation, audio, telephone and address book) by means of spoken interaction. To activate the “system” the subject had to press a push-to-activate button. The following recommendations also include evidence drawn from this WOZ experiment.

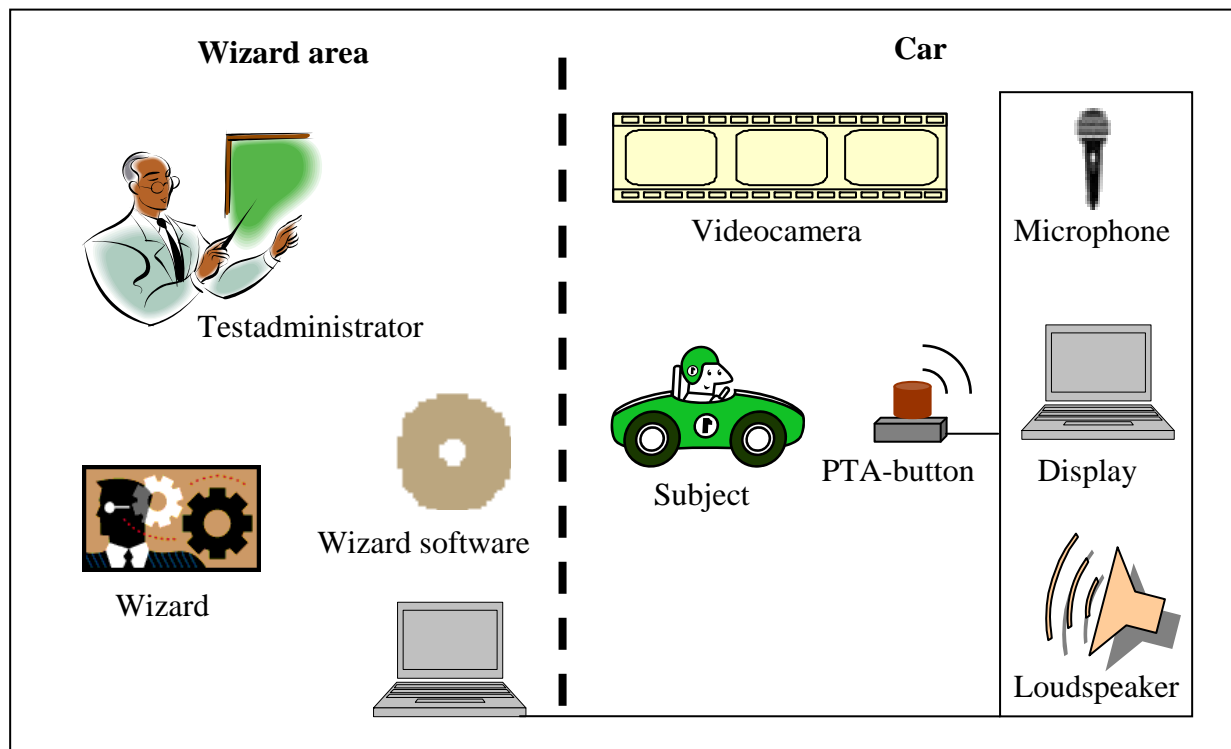


Figure 4.6: Experimental setup of Wizard-of-Oz test

4.4 Designing the interface for usable machines

Chapter 4.4 presents a set of recommendations for designing and developing multimodal dialogue systems in cars. The set was established on basis of the experiences and conclusions drawn from the experimental studies described in the previous chapter. It considers both expert and novice user. Where useful, features from human-human communication presented in Chapter 3 were integrated into the set as well.

4.4.1 Reusable dialogue components

In order to create in-car speech interfaces that are user-friendly it is necessary to provide a system the applications of which are well-structured and consistent. Reusable dialogue components (RDC) enable to realise actions resembling each other in a similar way (cf. Mann, 2003). They are subroutines that are complex enough to provide useful functionality, but small enough to be broadly reusable (Burnett, 2000, Chapter 2, 2.2). Taking a complex in-car speech dialogue system for several applications, the variety of actions may be subsumed by a small number of dialogue components (cf. Figure 4.7).

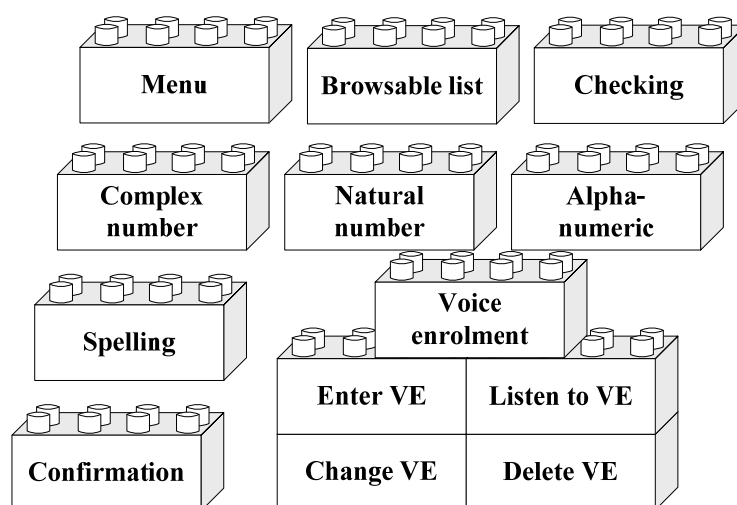


Figure 4.7: Examples of reusable dialogue components (Mann, 2003)

The category address book for example should provide the following functions, i.e. to store

- Phone numbers: 0731/5052331
- Street names: Main Street, King's Road
- House numbers: 48, 19/A
- Cities: Hamburg, Ulm
- Post codes: 89075
- And also to select arbitrarily chosen sound patterns (i.e. voice enrolments, see Chapter 4, 4.4.7) the user has linked with specific information: Home, Rachel and Colin, Mum

To cover this functionality the dialogue components presented in Table 4.2 would have to be subsumed. Dialogues may also be subsumed to RDC across applications. The dialogue component 'alphanumeric' for example covers both dialogues for entering a house number and selecting a radio station by frequency.

Imagine storing data within the applications address book and navigation. The novice user wants to store a new entry in his address book. Using the command phrase 'store name' he is first of all prompted for the number and, in a second step, to assign a spoken name (i.e. voice enrolment) to the corresponding number. The same user now enters the application navigation to store a new destination by speaking 'store destination'. In this case the system offers two alternatives. Alternative one already contains a destination on the display from previous interaction, so the

system prompts the user to speak a name and then stores this very destination. The process is completed. The novice user most likely will not be aware of what has been stored in this case. Taking the process of storing a name in his address book, the novice user would most naturally expect a request for entering a destination after having entered a voice tag. Alternative two does not show a particular destination on the display. After the user has said ‘store destination’, the system informs him that there is no destination available and ends the dialogue, leaving the user in total confusion. In a system applying reusable dialogue components the system would, analogously to ‘store name’, reply to the user’s request (i.e. ‘store destination’) by prompting to enter destination and afterwards assign a name to it. Novice users will feel more confident when coming across a process they have already accomplished successfully in a different task. Unnecessary complexity makes interaction inadequate and inefficient. By means of RDC and the consistency going along with them, efforts for learning an interface as well as the number of errors can be reduced.

RDC also help application developers to the effect that they have to deal less with time-consuming problems (due to lacking transparency) when defining, designing and implementing an interface.

Reusable Dialogue Components	Application Address Book
Voice enrolment	Spoken name/ nickname
Complex number	Phone number; post code
Spoken name	City; street name
Alphanumeric	House number

Table 4.2: Functions within address books and their corresponding RDC

Specification Tool

Nowadays various tools are used for specifying in-car speech dialogue systems. Flowcharts for example consist of sequences of actions and decisions. They are linked by directed arrows to describe the dialogue flow. A flowchart of the above alphanumeric dialogue component for example could be realised as shown in Figure 4.8. Actions can for example be speech commands, system output or concatenation with dialogue components described separately in the specification (cf. IBM, 2002). Different types of action are represented by means of different

shapes. Decisions depend on variables that store recognition data or internal data for verification.

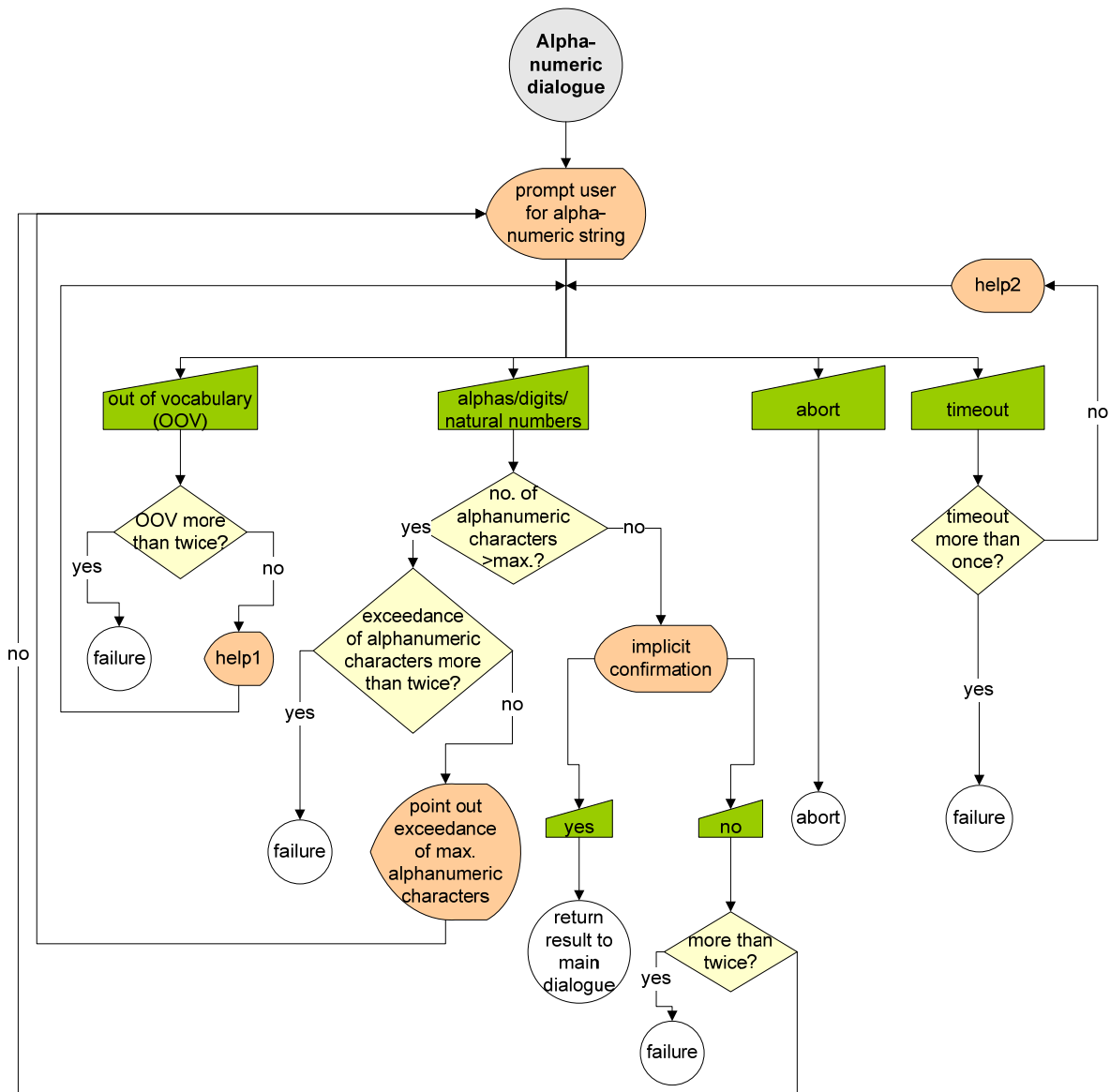


Figure 4.8: Sample dialogue flow for alphanumeric dialogue component (Mann, 2003)

As long as applications remain simple, flowcharts can be easily captured. However, if spoken and manual interaction need to be compared this may already comprise dealing with 2 separate specification documents as they are difficult to be unified within one. If each document in turn comprises several hundred pages to cover all interaction steps, consistent design and implementation becomes a difficult and time-consuming task. Consistency between transitions or among terms for example can only be verified manually. Beyond that, the functionality of in-car

applications becomes more and more complex, demanding and manifold. Speech input is heading for less restricted input. This implies that input may contain more than one parameter, e.g. ‘play me Laundry Service by Shakira’. Specifying all possibilities within flowcharts is likely to diminish clarity within this tool.

Statecharts as opposed to flowcharts are used to describe the behaviour of a system by means of diagrams and transitions (see e.g. Goronzy, 2005; Heinrich, 2007). They describe all of the possible states of an object as events occur (Braun, 2008). The basic idea of statecharts dates back to Harel (2007, p.1) in 1983.

The main novelty of the language [of statecharts] is in being a fully executable visual formalism intended for capturing the behaviour of complex real-world systems, and an interesting aspect of its history is that it illustrates the advantages of theoreticians venturing out into the trenches of the real world, “dirtying their hands” and working closely with the system’s engineers.

Independent of the representation formalism, tools for specifying multimodal speech dialogue systems should include the following aspects:

1. Multimodal specification - specifications for spoken and manual interaction have to be unified to get a better overview of both modalities. Being able to compare them comfortably ensures that both modalities can be adjusted to each other. This considerably increases consistency across modalities.
2. Consistency checks for
 - Transitions, automatically ensuring that each interaction step has a logical successor. This successor in turn has to be directly accessible to avoid extensive search.
 - Terms, to ensure that they remain consistent within a system but also throughout various design cycles. For example, this is relevant for dialogue component names as well as active vocabulary within and across modalities. It is necessary to make sure that vocabulary used for the graphical interface is also available for spoken interaction. Also expressions used within one particular task have to be reused in tasks that are similar.
 - Dialogue flow, to make sure that similar tasks from different applications as well as across different modalities are realised in a similar way.

3. Changes in menu structures, templates and for wording etc. have to be adjusted automatically within the specification to avoid errors. Changes that are not allowed have to be forbidden and should be indicated to the developer accordingly.
4. Integration of text passages is well-established in flowchart specifications. It is sometimes the easiest way to describe particular system features. This option is recommended for any type of specification tool.
5. Widgets and templates need to be provided to simplify the specification process. This also implies that changes on one particular widget or template automatically have to be transferred to wherever the corresponding widget or template is used.
6. Compatibility for multiple users – it is necessary that two or more persons may work on a specification document simultaneously. Care needs to be taken that changes made by one person do not collide with those of another person. In this context a history protocolling changes automatically is strongly recommendable to be able to comprehend who has changed what and when.
7. Specification of several displays within one document is recommendable to guarantee consistency between them.
8. Integration of variants within one specification. Variants occur because different car series comprise different functionalities and also due to different languages and regional differences on the European, Asian and North American market.

4.4.2 Prompts

Prompt design or speech output is the crucial modality when interacting with a dialogue system while driving. Interacting with the display should only be of minimal concern and need.

+ Prompts should be as informative and as short as possible.

Taking into account the persistence problem of speech, long prompts cannot be processed adequately. The user most likely will not (be able to) listen to it and in case he does, he will not be able to memorise what has been spoken. When offering menu items to the user, they should therefore not exceed the number of three items. In case a system provides for barge-in, the number of items may be extended accordingly.

+ Ensure awareness of context as if speech was the only modality.

When designing prompts care needs to be taken that context is always transparent to the user. Prompts need to reflect the state the system is in and have to be formulated accordingly, e.g.:

- The user is in the application audio and says ‘enter city name’. Due to potential recognition errors it is important that the system confirms the change of application by saying ‘navigation – please speak the city name’. In case the user already is in the application navigation, the first part of the prompt may be omitted.
- Similar tasks within different applications (e.g. for storing city name, radio station, name in address book etc.) clearly need to indicate to the user where he is. A prompt such as ‘please speak the name’ would be too vague and needs to be differentiated. The user thus becomes aware if he is about to store a radio station or an address book entry - in particular if he got into this dialogue accidentally.
- When entering a destination the user may be asked for additional information such as city nearby, river, Bundesland etc. This might be necessary for disambiguation. The questions should be adjusted to each other taking into account context: ‘do you know a city nearby?’ followed by ‘do you also remember the zip code?’. Simply stringing together identical prompts should be avoided as it runs counter to natural dialogue flow. It gives the impression that the system does not consider what the user says.

The dialogue system on the other hand is constantly aware of the current state it is in and knows what information needs to be requested from the user. It is therefore crucial for the system to

+ Consider context when interpreting user utterances.

Many utterances can only be interpreted correctly if the system takes into account the context they were made in. For example, if the user says ‘number 3’ in the application navigation he refers to the third item presented in a list of city or street names. When in the address book, the user intends to select the third entry of his address book.

4.4.3 Vocabulary

Evidence from collected user speech data in context of a Wizard-of-Oz experiment (Hüning et al., 2003) has shown that the length of user utterances goes beyond what can be covered by simple command & control grammars. It is therefore recommended to

+ Use less restricted input for spoken user interaction.

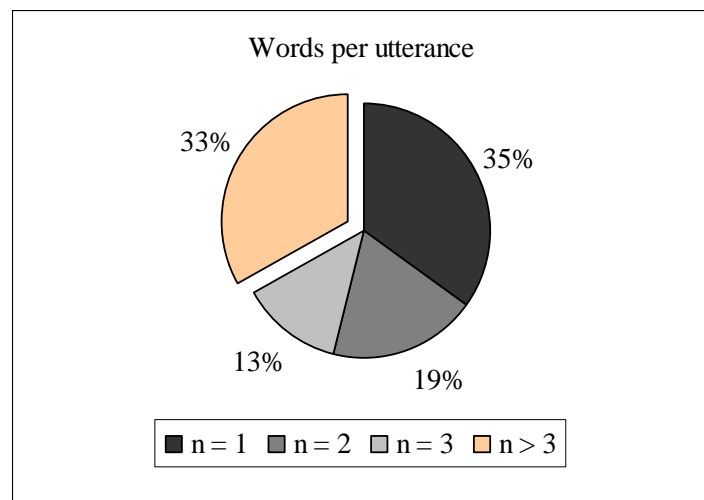


Figure 4.9: Number of words per user utterance (adopted from Hüning et al., 2003, p.24)

Figure 4.9 shows the number of words per utterance from the data set. One third of the utterances, i.e. where the number of words is bigger than three, could be considered as natural language input – input that is not yet covered by current speech dialogue systems.

+ Use synonyms and filler words the user is likely to come up with.

To evaluate the vocabulary used by participants when interacting with the WOZ system described in section 4.3 the data were split into five types of words:

- User words: different words used by the participants
- Prompt words: words used in the system's prompt
- Task words: words used in given tasks
- GUI words: words identical with the graphical user interface
- Grammar words: words available in a command & control grammar

Comparing the user words with the other four types of words shows that most of the words the subjects use (62%) are not covered by any of the sources (see Figure 4.10). This means that despite externally influencing factors (i.e. display, prompts, task description) users feel free to use their own wording. When further dividing the uncovered vocabulary into words relevant for applications or not, it can be seen that 38% are synonyms and 24% are so-called filler words, coming from natural input.

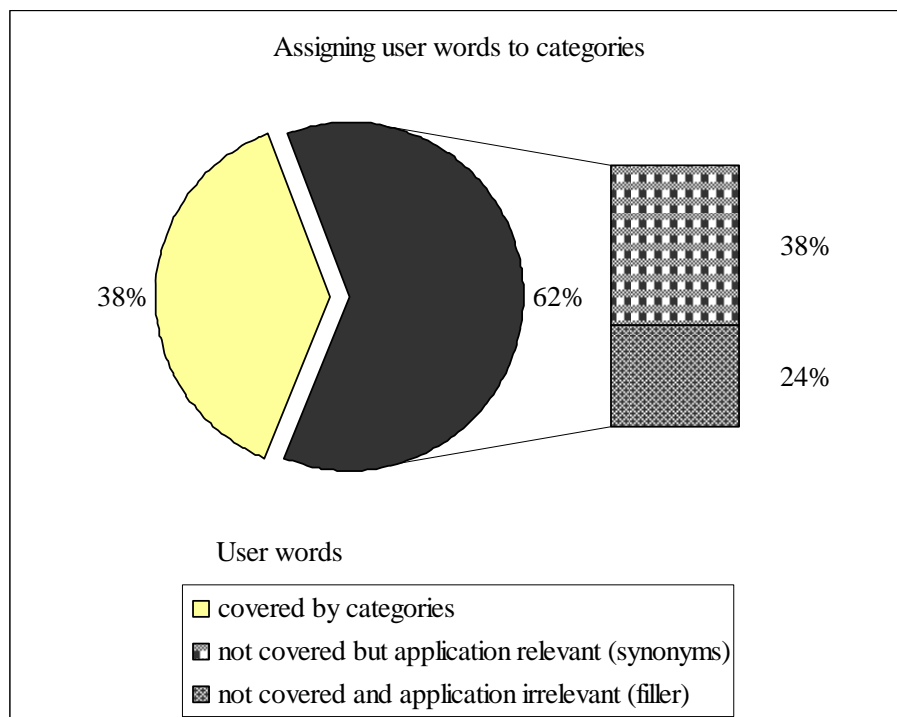


Figure 4.10: Coverage of user words (adopted from Hüning et al., 2003, p.26)

+ Ensure contextual interpretation and backtracking of former user behaviour to resolve linguistic phenomena.

To correctly interpret linguistic phenomena occurring when humans interact with computers, the system needs to be able to backtrack former user behaviour and to recover context. These linguistic phenomena are for example ambiguities, deictic expressions, ellipsis and extrapositions.

Let us take the ambiguous city name Neustadt again (cf. Chapter 1), standing for 29 different cities in Germany called Neustadt. In case the user tells his navigation system to go to Neustadt

(probably not aware of the ambiguity) the system will most likely reply “which Neustadt?” while simultaneously displaying a picklist containing all alternatives. Having gone through the list (that might be several pages long due to the large number of items), the user finally selects “Neustadt an der Weinstraße”. In case the user wants to go there again some time later, it would be cumbersome if he had to go through exactly the same procedure again. Taking into account former user behaviour by means of a dialogue history, the system could offer the user “Neustadt an der Weinstraße?” as default value.

Deictic expressions are linguistic items of pointing character (cf. Chapter 3, 3.2.4). They may refer to the situational context, i.e. situations taking into account speaker-hearer constellation as well as place and time relators (Herbst, 1991, p.183). In example (8)

Example 8:

System: Your destination is Frankfurt am Main, Amselweg.

User: Yes. I want to book a hotel room *there*.

the adverbial pronoun *there* refers to the final destination Frankfurt am Main. *Now* as in “I need a break now” refers to the driver’s current position. Both *there* and *now* refer to elements of a concrete situation (Herbst, 1991, p.183).

Non-deictic expressions have specific reference the identity of which is supplied by the linguistic context or context of situation (Herbst, 1991, p.182; also see Greenbaum, 1990). When establishing references to the linguistic context, they may function as anaphors and cataphors. An anaphor is a syntactic element referring to another syntactic element (antecedent) of the preceding context, i.e. its reference can only be deduced by means of an antecedent (Bußmann, 1990, p.82).

Example 9:

System: Internet – how can I help you?

User: I want the latest news. Please read *them* out.

Cataphors in turn are syntactic elements the identity of which is established by what follows (Greenbaum, 1990, p.79).

Example 10:

User: Start navigation to *the following* destination: Frankfurt am Main, Amselweg.

Ellipsis is a sentence where grammatical constituents are omitted (Bußmann, 1990, p.207; Greenbaum, 1990, p.255). The missing word or words occur in the remaining context and can be precisely deduced (Herbst, 1991, p.183). Bußmann (1990, p.207) distinguishes five types:

1. In lexical ellipsis constituents are left out the interpretation of which depends on the linguistic context.

Example 11:

System: Frankfurt am Main, Amselweg.

User: Store (destination).

2. Imperative sentences go along with an obligatory ellipsis of the subject, e.g. “redial number” or “delete entry”.
3. Ellipsis in question-answering pairs omits identical constituents that have already been mentioned.

Example 12:

System: Which music title?

User: (Music title) number one.

4. Coordination or gapping constructions may go along with reduction of identical constituents, e.g. “add title number 7 and (title number) 9 to playlist”.
5. In infinitive constructions the subject is obligatorily left out, e.g. “I have been trying (...) to get hold of you several times”.

Lexical ellipsis, ellipsis in imperative sentences and ellipsis in question-answering pairs are already covered by command and control applications (Hüning et al., 2003, p.27). Coordination or gapping constructions however may be interesting to consider for new dialogue strategies when less restricted input is allowed, as for example “dial phone number and then store (phone number)” or “add title number 8 to playlist and then play (title number 8)”.

Extrapositions are constructions where a clause having the function of either subject or object is shifted to the end of a sentence; the actual position may be substituted by a placeholder such as

‘it’ (Herbst, 1991, p.102; Bußmann, 1990, p.233). E.g. “It should be stored under Miller, the number I have just dialled” or “I want to store a phone number, the one just dialled”.

As far as the collected user speech data are concerned extrapositions and gapping constructions only occur rarely (Hüning et al., 2003, p.28). Consequently, rather than developing a deep syntactic processing strategy for these phenomena, a grammar providing a less restricted word order would be sufficient.

+ Expect occurrence of disfluencies in spontaneous speech.

If input for spoken user interaction aims at input that is less restricted than command & control systems, it is necessary to consider spontaneous speech and its peculiarities. For of course dialogue is not a mere rational cooperation, assuming that the speaker’s utterances be well-formed sentences but also a social interaction in which phenomena such as disfluencies, abrupt shift of focus, etc. occur (Cole, 1996, Chapter 6.4; also see Levinson, 1983). Types of disfluencies may include the following (Kronenberg, 2001, p.12 et seqq.):

- Ungrammatical syntactic constructions with regard to case, number or gender etc. In the sentence “I want to listen to titles number 7” a discrepancy occurs in number. The user only selects one song, the word titles however is plural. Inconsistencies in syntactic constructions occur more often in languages such as German, French or Italian as they are inflectionally rich languages.
- Utterances may be interrupted at any position. Interruptions often go along with sentence new starts, e.g. as in “I want to make a. Store phone number please” or “Store phone number under. Correct phone number”.
- Substitutions occurring due to recognition problems. The title “Folsom Prison Blues” for example is likely to cause problems with regard to phonetic segmentation. Instead of “Folsom Prison Blues” the system might return a title called “False Imprisoned Blues”.
- Deletions resulting from corrections of the preceding utterance or recognition errors.

Example 13:

User: Navigation – store destination.

System: Navigation – which function please?

- Hesitations such as ‘um’, ‘uh’ and ‘err’, often accompanied by pauses.

Studies on the correlation between cognitive load and the occurrence of disfluencies are manifold (Corley, 2008; Oviatt, 1995; Oomen, 2001). However, it is controversial whether increasing cognitive load causes an increase in the number of disfluencies.

4.4.4 Feedback from the system

At any time during human-machine interaction the user needs to be aware of the state the system is in (cf. Nielsen, 1993, p.134). This implies that both speech and display consistently have to reflect the same state.

+ Avoid that speech and display represent different system states.

Imagine the user is in the application navigation and, using speech as input mode, changes to the application telephone. Instead of getting the same feedback both verbally and visually, only the speech mode changes to telephone allowing the user to verbally input a phone number; the visual mode, however, remains in the application navigation displaying the current route. The fact that the display does not change the system state is likely to confuse the user as he cannot be sure whether the system has actually carried out changing states. Apart from that the user is not able to verify the entered phone number by throwing a quick glance at the display.

Whenever applications are changed the aim is to inform the user about where he currently is. Otherwise it may easily happen – in particular in context with recognition errors or free input – that he does not know in which state he is actually in. The change should be confirmed by speech as well as through corresponding visual reaction. Thus speech and haptic interface are adjusted to each other and the user is not induced to focus on the display while driving.

+ Provide additional feedback for similar tasks.

To ensure that the user is aware of the system state he is in, tasks with similar dialogue flow need to be differentiated by disambiguating system prompts. As already mentioned in 4.4.3 storing an entry is a procedure occurring in several applications such as navigation, telephone or audio (e.g. to store a destination, phone number or radio station). Requesting the user to “please speak the name” has been proven to be misleading many times. Extended prompting such as “store destination – please speak the name” resolves ambiguities and avoids potential errors.

+ Provide feedback for unspecific input.

It is necessary to ensure that unspecific user input (e.g. utterances such as “destination“, “store“ or “delete”) does not cause an abort. Instead it is necessary that the system communicates what has been recognised and reacts accordingly to keep interaction going into the intended direction. The command “delete” for example could imply deleting an address book entry, a destination or a list of favourite music titles etc. In case context cannot clarify the procedure intended with the respective command, possible options should be offered in a menu: “Delete an address book entry? – *pause* – a destination? – *pause* – favourite title list? – etc.”. As far as the sequence of items is concerned care needs to be taken that the options of the active application are offered first. Alternatively the system could prompt the user “what would you like to delete?”.

Utterances that can only be recognised partially also need to be interpreted by the system. In case interpretation is unclear the system needs to reprompt the user.

+ Ensure additional user feedback if input has not been recognised properly.

In case of reprompting the user, confidences of recognition results are useful to consider. If given input is on a predefined way through the current task, confidence for confirmation should be higher. On the other hand, confidence for reprompting should be lower if recognition results deviate from this predefined way. Consider entering a phone number: the user changes from the application navigation to telephone and tells the system he wants to enter a phone number. In this case it is irrelevant whether confidence is high or low because the chosen command “enter phone number” is on a predefined way of the telephone application. Thus, in the next step, the system may directly prompt the user to enter a phone number without additionally requiring confirmation. However, in case the user is in the application navigation speaking “enter phone number” a higher confidence is necessary for the system to change applications. In case of low confidence, the system could reprompt the user if he wants to make a telephone call.

+ Impede changing applications if dialogue flow has advanced considerably.

Whenever dialogue flow has already advanced considerably within a task, utterances implying a change of applications should be weakened concerning their confidence. Instead results having

worse confidence should be taken into account in case they fit into the dialogue flow. Take “dial phone number” versus “enter house number” for example: in the application navigation the user is about to store a new destination. He has already entered city and street name – in the following interaction step, however, the system determines high confidence for “dial phone number” and low confidence for the actually following subtask (i.e. enter house number). As the user has already proceeded considerably in his navigation task the recognition result with higher confidence, i.e. “dial phone number”, will be rejected. Instead, the command with lower confidence will be accepted as it fits into the current dialogue flow. As a precaution the system could reprompt the user whether he wants to enter a house number.

+ Confirm selected list elements.

Once a user has selected a particular item from a list it is important that feedback is given of what has actually been recognised and selected by the system. Not till then is it apparent to the user that the respective element of his choice has been selected. Imagine a list of city names for example: once the user has entered a city name the system in most cases returns a list with several alternatives. The user is then asked to select an item, in general by speaking the corresponding line number. In case the command “number 3” leads to a system prompt “city has been selected” the user does not get feedback which city has been chosen. If the system chooses line number 2 instead of 3 due to recognition errors the user can only find this out by looking at the display. Alternatively the system should explicitly confirm city names as in “city Hamburg has been selected”. Thus the user directly notices if something went wrong during interaction.

4.4.5 Help

It is crucial to design human-machine dialogues in such a way that help modes can be largely avoided. Whenever changing to a different task, the user should be informed about the options available in this context. This information should be conveyed prior to leading the user to activate a separate help mode.

+ Inform the user about possible options prior to leading him into a help mode.

Example 14:

User: Address book.

System: Address book – which function?

Being familiar with the system, the expert user can easily continue this dialogue without having to bother about extended system prompts. The novice user who might not be aware of the options he has at this state will be additionally prompted in form of a menu: “do you want to store an entry – *pause* – search for an entry?”. This kind of adaptive help prevents the user from explicitly having to activate the help function. Besides it keeps tracking the task flow.

+ Provide help that is short and precise.

Help providing an overload of information makes it impossible for the user to keep in mind all details. Stringing together a large number of possible commands should be prevented as at the end of the prompt the user most likely will have forgotten what was said at the beginning. In case the user is offered help with several menu items its number should not exceed more than three unless the system provides barge-in.

+ Make sure the user knows help is available at any time (also see Balentine, 2001, p.62).

This can be provided for best if the system takes initiative to offer it. For example if the user verbally changes to the telephone application the system should not simply react by adjusting the display and then stop interacting. The aim at this point should be to also confirm this change of application by speech. The user may then directly select the intended function. In case the user says something unknown to the system or does not give any input at all, the system could successively offer available functions within the application. Should the user in turn not select any of these functions the system could explicitly offer help.

+ Ensure that context-sensitive help is available.

Information provided for in a general help should additionally be retrievable in form of context-sensitive help. Being about to adjust map settings in the application navigation for example the user might directly want to know the corresponding commands to do this. No more and no less.

A general help providing all possible functions within navigation would force the user to listen to a lot of redundant information before getting what he was looking for.

+ Provide exit points for help prompts (also see Balentine, 2001, p.64).

Extensive help prompts that are hierarchically structured (implying that the user is lead through several menus) need to provide exit points, e.g. questions like “do you want more help on ...?”.

+ Continue dialogue subsequent to help activation.

After the user has activated a particular help it is important not to abort interaction. Instead, the system could keep interaction going by directly changing to the corresponding function that has been explained in the latter help prompt.

+ Clarify complicated instructions with examples (also see Balentine, 2001, p.62).

Help instructions that at a first glance might not be transparent to the user should be accompanied by examples if possible. Requesting the user to “please spell the street” is often misleading and rather than spelling the street name the user simply enters one word. In this case the system could alternatively prompt “please spell the street – for Stuttgarter Straße for example say S-T-U-T-T”. This would be more helpful than simply reprompting the user to please spell the street name.

4.4.6 Spelling

When it comes to requesting names such as cities, streets or address book entries, the option of whole-word input should always be favoured to the spelling mode. As already mentioned in section 4.4.5 the request to spell a city or street name is often ignored or even misunderstood by users. The consequence is that users apply whole-word input and only after several trials (with correction and help instruction) do they realise the corresponding name needs to be input letter by letter. Spelling recognition in turn is prone to errors as very often the input number of letters is either too small or too big. If in this context the system reprompts the user saying “sorry, what did you say?” the user gets the impression the system does not understand what he says. He starts thinking what he did wrong, thus causing timeouts.

+ Use spelling mode as fallback solution only.

Whenever a task specification provides spelling mode the duration of timeouts needs to be extended. Having dealt with several in-car speech dialogue systems it was found that users fairly often managed to enter one letter of a city only and then recognition stopped. The same happened in context with phone numbers. Consequently the impression arose that digits or letters need to be input one by one, causing the user to continue interaction that way.

4.4.7 Voice enrolments

Voice enrolments are retrieval indices for address book entries, radio stations, destinations etc. in form of voice templates (Mann, 2006). These enrolments are entered by the user and generally need to be spoken twice before being stored with the correspondingly linked data (Audi, 2005; BMW, 2006; Mercedes-Benz, 2003). The necessity of entering a name twice becomes problematic if the recogniser is not able to correctly match the two utterances. Because then the user will be prompted another two times to speak the name. Repeating the same name four times is irritating to the user to the effect that he pronounces the name differently, hoping to be finally understood. The probability of successfully storing an entry by voice enrolment is thus fairly low and cumbersome. It should therefore be sufficient to speak a voice enrolment once in order for it to be added to the lexicon (Saab, 2003).

+ Speaking a voice enrolment once must be sufficient.

Taking a closer look at in-car address books it can be found that they are not conform to address books in printed form. Whereas the latter are most often sorted alphabetically by the initial letter of the retrieval index, this quasi-natural approach (Heisterkamp, 2003) is not always possible for speech-enabled address books.

Entries of state-of-the-art speech dialogue systems can either have indices in form of textual fields or voice templates, or both. This means that some retrieval indices do not exist in text form, only as speech templates. In order to sort the voice enrolments in address books in the same fashion as textual indices, and at the position the user expects them to be, it would at least be necessary to know the initial letter. However, state-of-the-art speech recognition can neither reliably transfer voice enrolments into written form nor can they reliably determine the initial

letter of the template. Voice enrolments are therefore usually put at the end of an address book. The voice enrolments are presented in a quasi-random fashion and cannot be accessed systematically. This means that the user has to tediously listen through all voice enrolments when looking for a particular entry.

As long as on-board-only address books were still fairly small, they could be read out to the user item by item, beginning to end. However, the number and complexity of electronic on-board devices in vehicles has permanently increased over the past years. As a consequence of the increased functionality, also parts of in-vehicle dialogue systems tend to have lost transparency. Address books available in cars today can be compiled from several different sources (e.g. from the on-board address book, an organizer, mobile phone or Personal Digital Assistant PDA). The number of address book entries available in the vehicle thus rises sharply. This in turn aggravates the user's possibility of accessing the data that have been stored, in particular if he does not correctly remember the retrieval index of an address book entry.

+ Provide methods for accessing personal data that include lacking recollection rate.

In a user study DaimlerChrysler analysed the structure and usage of address book data (Enigk, 2004; Mann, 2006). 21 subjects aged 28 to 62 years submitted their electronic address books. Datasets from 24 devices (20 mobile phones and 4 PDAs) were read out to analyse how personal address books are used. The first finding was that the subjects do not structure their address books in a uniform way. Address book entries are strongly idiosyncratic. Data fields for names, in particular, contain different components that can be combined in numerous ways:

- First and last name
- First name only
- Last name only
- Title
- Organisation/ institution
- Position/ function/ department

This variety of combinations causes problems – already for the users themselves. During the study it was measured how well users recollect the entries they have stored in their address book

(see Figure 4.11). By means of 420 scenarios (20 per subject) on different contact groups such as business, private etc., speech commands were recorded and subsequently compared to the data that had actually been stored by the user.

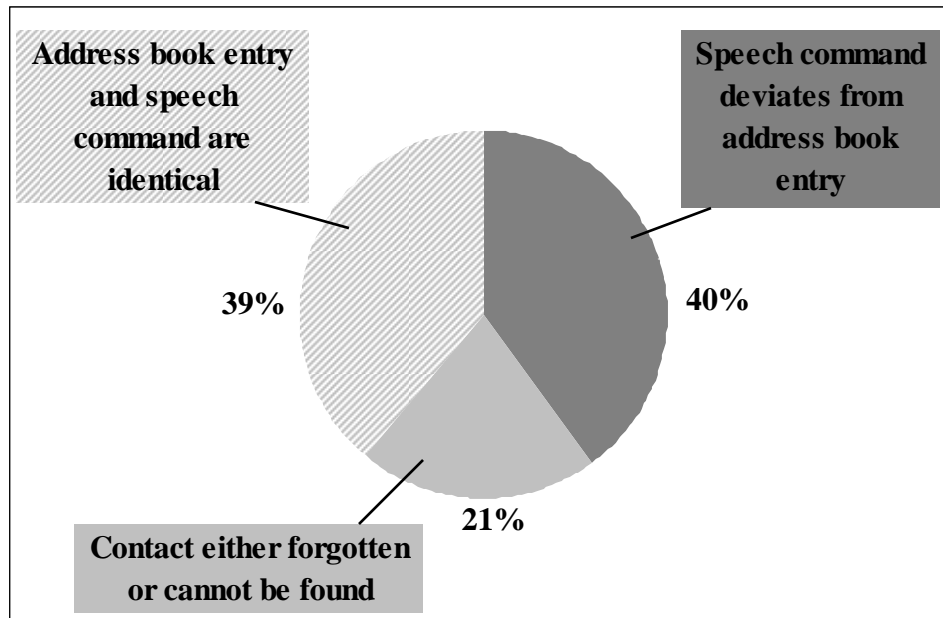


Figure 4.11: User knowledge of personal address book data

It was found that only in 39% of the cases did the subjects correctly recall the address book entries they had originally chosen. In 61% of the cases, there was no direct match between the commands the users used and what they had stored. To retrieve these entries, users need to have other means of retrieval rather than the direct access through voice enrolment.

+ Enable alphabetical sorting for voice-enrolled entries.

To allow for direct, structured and user-friendly access to address book entries the user could be prompted to assign an alphanumeric character to each voice enrolment entered. It can either be spoken using one of several spelling alphabets (MindSpring Enterprises, 1997), e.g. “A as in Alpha”, “B as in Baltimore”, etc. or by manual input. This approach allows structuring both voice enrolments and text entries in a uniform way. When then searching for a certain voice enrolment, the user can restrict the search to only read out the entries under a particular letter, getting both the textual entries and the voice enrolments (see Figure 4.12). Note, however, that

this only applies to first-letter sorting, i.e. all the voice enrolments under one initial letter will be ranked equal at either beginning or end of the letter-sorted group. But: The approach also allows repeating the letter-allocation process for subsequent letters if this proves necessary.

The figure shows two side-by-side boxes, each representing a user interface for a letter 'J'.
 The left box has a large 'J' at the top left. Below it are two input fields: 'NAME' containing 'John Q. Public' and 'PHONE NUMBER' containing '+49 (0)432 56789'.
 The right box also has a large 'J' at the top left. Below it are two input fields: 'NAME' containing a dashed line '-----' and 'PHONE NUMBER' containing '+49 (0)345 68294'. A speaker icon is positioned to the right of the 'NAME' field.

Figure 4.12: Example of combining text entries and voice enrolments under one letter

Second, for those voice enrolments the user did not (yet) assign an initial letter to, there remains a catch-all category (see Figure 4.13). It is a separate category additionally to the 26 letters in German (29 including vowels with umlaut). The entries of this new category could for example be accessed by “names without letters” or “spoken names”.

The figure shows two side-by-side boxes representing alternative categories for voice enrolments.
 The left box features a speaker icon at the top left. Below it are two input fields: 'NAME' containing a dashed line '-----' and 'PHONE NUMBER' containing '+49 (0)123 45678'. A second speaker icon is located to the right of the 'NAME' field.
 The right box is titled 'Spoken Names' at the top. Below the title are two input fields: 'NAME' containing a dashed line '-----' and 'PHONE NUMBER' containing '+49 (0)123 45678'. A speaker icon is positioned to the right of the 'NAME' field.

Figure 4.13: Two alternative examples for storing address book entries (voice enrolments) without alphanumeric characters

The user is thereby given the option to retrieve all entries that could not be alphabetically sorted.

Spelling alphabets might be difficult to learn. Alternatively, if spelling alphabets prove cumbersome, the user could speak a city name from the currently active city name vocabulary. For Germany, e.g. about 58,000 location names are directly speakable and are available in textual form as well. Instead of “B as in Baltimore” the user could then for example say “B as in Berlin” or “B as in Bonn”.

4.4.8 Barge-in

In human-human communication various turn-taking strategies are applied when making contributions to a dialogue (Jersak, 2006): Taking turns when rising to speak, interrupting turns of dialogue partners, taking another turn when pauses or hesitations occur within a conversation, as well as backchanneling (utterances such as yeah, right, o.k. etc.) to confirm the turn of the dialogue partner. The more of these strategies a dialogue system supports, the more natural dialogue control seems to be. To a large extent, however, acceptance of a speech-driven dialogue system depends on how efficient it is. Interacting with a speech dialogue system while driving should distract the driver as little as possible. In this situation an interaction with a low number of turns is the most efficient one.

+ **Replace sequential recognition by barge-in recognition.**

When supporting a more natural dialogue flow with turn-taking strategies the efficiency of dialogue control may decrease. Two factors that might increase the number of necessary turns are talk-over (Kaspar, 1997) and interrupting system utterances by error recognition. As for sequential recognition it is not until the end of a system prompt that the user is able to start speech input. In case the user starts speaking too early (talk-over) only the part after recognition has started will be recognised. At first sight the user does not notice this recognition problem at all. Recognition continues even though the initial part of the user input is missing. Consequently the parser may either return a wrong result or no result at all.

Example 15:

System: Navigation – **which function?**
User: **Search entry in** address book.
System: Address book – which function?
User: Search entry.
System: Which name?

In example (15) the initial part of the utterance “search entry in” would get lost. The parsing result “address book” would lead the system to change to the address book and ask the user what he would like to do – although in fact the user has already precisely uttered his request.

In case speech input lies completely outside recognition (interruption), the recogniser returns a timeout error and the system “acts” as if no input had taken place at all.

Example 16:

System: Your destination is: Hamburg, Poststraße 25. Do you want to start navigation?

User: **Start navigation.**

System: Start navigation?

User: Yes.

In both cases, talk-over and interruption, additional system-user turns are necessary for clarifying dialogue, unnecessarily delaying task completion.

Barge-in recognition instead allows interrupting system utterances. Speech synthesis and recognition start simultaneously, i.e. user input may more or less start simultaneously with speech synthesis. This makes interaction more user-friendly, reducing the number of system-user turns and accelerating dialogue flow.

As far as interruptions are concerned, care needs to be taken that unnecessary queries from the system are avoided (Jersak, 2004). This means that recognition should not automatically stop whenever an utterance was made by the user. Instead, while recognition still remains active the system should simultaneously verify if the utterance is a valid expression or not. In case it is not, recognition should simply continue, ignoring what has been spoken by the user. Users will then automatically repeat or correct their input, without explicitly being prompted by the system (Sacks, 1974).

4.4.9 Error-recovery

In general, errors can be divided into three major categories (Balentine, 2001, p.273):

- Recognition errors
- Environmental errors (noise and echo)
- Behavioural errors

User testing of various speech dialogue systems shows that fairly often behavioural errors are on the majority, for example not knowing when to activate the push-to-activate button, spoke-too-soon, timeout, use of synonyms etc.

+ Provide a “back”-option at any step during dialogue.

Avoiding errors should have top priority during dialogue design. Nevertheless it is necessary to provide adequate means of recovering from potential errors. It is essential for the user to have a “back”-option at any step to prevent him from being stuck in a dialogue as follows:

Example 17:

System: Which point of interest?
User: Petrol station.
System: Please select a petrol station. (*Presenting list of possible items*)
User: Back.
System: You have reached the beginning of the list. Which petrol station?
User: Back.
System: You have reached ...

With his command “back” the user clearly intends to get out of the list selection. This option should be possible as fallback solution in case the user either does not know how to proceed or accidentally gets into a particular dialogue due to error recognition or simply because he has changed his mind. Besides, getting out of the list to the previous interaction step would be analogue to what people daily experience and expect when manually interacting with the world-wide web.

As far as list selection is concerned it is also important to let the user know he is able to verbally scroll a list. Experience from several studies showed that this is not necessarily apparent to the user. Corresponding information could be conveyed to the user by simply adding “previous” and “next” icons to the display. He is thus prevented from unnecessary timeouts or aborts when not knowing how to continue if the required item does not appear on the display.

+ Avoid dialogue states that cannot be exited verbally.

Dialogue states that cannot be exited by speech should be avoided. State-of-the-art training of voice enrolments to store an entry generally requires speaking a name twice (cf. Chapter 4, 4.4.7). Once the user is prompted by the system to please speak the name, anything spoken by the user will be interpreted as speech tag for the new entry. In case the user got into this state accidentally it is difficult to exit.

Example 18:

System: Please speak the name.

User: Back.

System: Please repeat the name.

User: Go back please.

System: The names are too different. Please speak the name again.

User: ...

Due to error-proneness it is advisable to keep the active number of commands low at this stage. Nevertheless crucial commands such as “back” or for main applications need to be processable by the system.

+ Discard recognition results that have already been rejected by the user.

Whenever the user corrects a list of two or more items (e.g. a result list of cities, street names or audio titles) because the item of his choice is not among them, it is recommended to discard previous recognition results for the following attempt as they have already been rejected (also see Balentine, 2001, p.277).

+ Determine useful entry points once dialogue has been cancelled.

If dialogues are cancelled by the user care needs to be taken to determine useful entry points for when the user restarts interaction. If for example the input of a pin number leads to an abort after several trials the display should not remain in the state of presenting the digits (or asterisks) that have been entered. Instead, the digits should be deleted automatically to facilitate re-entering dialogue.

+ Consider various types of correction strategies.

When something goes wrong during interaction users make use of different types of corrections. In the Wizard-of-Oz experiment by Hüning et al. (2003, p.32) the most common type is complete repetition, i.e. in nearly 50% of the cases the subjects made a one-to-one repetition of what they had previously uttered. Other major corrections are partial repetition of previous utterances, simplification as well as keyword use such as 'wrong' or 'no'. These four strategies cover almost 95% of all corrections. Whereas complete and partial repetition as well as simplification frequently occur in context with an initial keyword, in particular partial repetition is often combined with keywords (84%). It was also found that users apply these strategies depending on the type of input. In context with digit sequences used for entering phone numbers or zip codes corrections were mostly made using a complete repetition (59%). Functionality words (expressions used for controlling functionalities of address books, radio etc.) on the contrary are corrected most frequently using just a keyword (58%) and hardly ever using a repetition type (<5%). Depending on ASR (Automatic Speech Recognition) performance and the complexity of a speech dialogue system it would be advisable to enable more than just one of these correction strategies.

4.4.10 Initiating speech dialogue

Systems that do not use barge-in require a button for activating speech dialogue. In general this button is positioned on the steering wheel. Experience from various studies has shown that the logic of a so-called push-to-activate (PTA) button must be intuitively understandable. Otherwise this button may cause errors and uncertainty on the user-side, turning into an obstacle for acquiring a system.

+ Avoid sporadic use of a push-to-activate button.

Dialogue flow should be designed such that the user only needs to activate the PTA button once when about to begin a task (or after dialogue has been cancelled). Alternatively the user could be requested to consequently press this button whenever he wants to speak. Sporadic necessity of pressing a PTA button merely prevents the user from actually learning when to press it. After several trials he is likely to be so confused that he starts pressing the PTA button permanently.

This in turn gets even more problematic if the PTA button is given additional functionality (e.g. selecting highlighted items from a list by pressing the PTA button, ending phone calls etc.) as can be seen in the following example.

Example 19:

System: Which point of interest?

User: *(Pressing PTA button)* Hotel.

System: *(Displaying a list of hotels with the first item being highlighted as default value).*

Please select a hotel.

User: *(Pressing PTA button)* Number 3, please.

System: Hotel number one.

In this case the system prompt “hotel number one” is not a recognition error as assumed by the user. Having requested the user to select a hotel, the system interprets the first incoming event of the user, i.e. pressing the PTA button. The actual utterance “number 3, please” remains unheard. This is due to the ambiguous character of the PTA button. At the time the user pressed the PTA button, the recogniser was already open waiting for user input. This is when the additional functionality of the PTA button comes in: once the recogniser is active, pressing the PTA button within a list implies selecting the highlighted item, i.e. item number one. To avoid this confusion it is recommended to use a PTA button for one function only, i.e. for activating dialogue.

+ Avoid multiple functions for PTA buttons.

If then, during dialogue the user unnecessarily activates this button, it may simply be ignored by the system, thus remaining without consequences for the user.

Adding a prompting tone going along with pressing the PTA button is a matter of taste (Balentine, 2001, p.133). But given the fact that the display visualises an active recogniser by means of a loudspeaker symbol or else, it would be consistent to insert an acoustic counterpart to synchronise spoken and manual interaction (cf. Chapter 4, 4.4.13).

4.4.11 Short cuts

Strictly adhering to hierarchical menu structures of speech interfaces is not common for human beings. It is important for novice users to be lead through a task step by step. But as soon as they have acquired a system (slowly turning from novice to expert users), directly jumping from one task to another gets common. They know how the system works, so they want to be able to diminish the number of interaction steps to get their tasks accomplished as quickly as possible. In the Wizard-of-Oz study by Hüning et al. (2003, p.29) it was examined how users jump from solving one task to another without sticking to the menu structure of the interface. Two types of task changes were identified: short cuts, implying a task change within a particular application, e.g. changing from CD to radio which both belong to the audio application; and long cuts, implying a task change from one application to another, e.g. changing from radio to dialling a phone number which belong to the applications audio and telephone respectively. Whereas there is a strong tendency of one user group to extensively use long and short cuts (roughly every fifth utterance), the other user group mainly stayed within their given tasks, hardly using long or short cuts at all. It is obvious that the users frequently changing tasks by long and short cuts are those being experienced with speech dialogue systems. This behaviour seems to indicate that the idea of cutting long tasks short is a desirable feature for man-machine interaction.

+ Provide short cuts for advanced users.

It was also found that short cuts and long cuts are not equally used (Hüning et al., 2003, p.30). In the collected data there were about five times more short cuts than long cuts. This leads to the conclusion that users seem to be quite willing to switch tasks within one application rather than two applications. Obviously they respect the high-level structure of a speech interface. This aspect is advantageous when designing a speech interface as it is not necessary to keep a large vocabulary active extending over several applications. Instead it would be sufficient to extend active vocabulary within an application by a small number of short cuts. Taking a music application, short cuts such as “search artist”, “play current title list” or “search Rock” would considerably decrease the number of interaction steps and thus the time necessary for accomplishing a task.

4.4.12 Initiative

Harris (2005, p.153) states three logical types of initiative in a dialogue between two agents:

- System initiative: the system has all the control over the flow of the dialogue
- User initiative: the user has all the control over the flow of the dialogue
- Mixed initiative: both agents share control over the flow of the dialogue, each able to assert (or relinquish) that control at any given point

Mixed initiative naturally occurs in human-human communication. In the WOZ experiment by Hüning et al. (2003, p.30) it was therefore investigated to what extent subjects are willing to stick to a system-initiated interaction. To do so, the data were analysed according to occurrences where the subject took initiative to provide input without having been prompted to do so. The results clearly indicate that, just as in human dialogue, the subjects deviate from a directed dialogue. They applied mixed initiative in about 43% of the cases.

+ Provide speech interfaces with mixed initiative.

Mixed initiative means that initiative may shift as the task-collaboration requires, e.g. to initiate a different task or to ask questions (Harris, 2005, p.155; McTear, 2004, p.110). This could be achieved by allowing the user to input two or more command phrases within one utterance.

Example 20:

User: Navigation.
System: Yes, please?
User: Search point of interest.
System: Which point of interest?
User: Petrol station.
System: (*Displaying a list of various petrol stations.*) Please select a petrol station.
User: Guide me to the nearest Aral petrol station, please.

Accordingly, dialogue strategy must be flexible to support an interaction by which users are free to choose their own way through the dialogue (Hüning et al., 2003, p.38).

4.4.13 Combining spoken and manual interaction

Beyond developing concepts for voice control the logic of both speech and manual interaction has to be integrated into a concept that is uniform and consistent. This requires that both kinds of interaction mode have to be synchronised and adjusted to each other. It is important to make the user feel and experience he interacts with one system only. A manual interface where the speech component is merely attached is cumbersome for the user to acquire: identical tasks might have to be dealt with differently across modalities which results in having to learn things twice. In addition, the user has to be capable of assigning the correct procedure to the corresponding interaction mode.

+ Ensure the wording of the graphical user interface is speakable.

The display of in-car information systems consists of several function bars, e.g. for main functions, main area and submenus (cf. Chapter 2). These bars for example contain terms for available applications (e.g. navigation, audio, telephone, video, vehicle) as well as corresponding submenus available in these applications (e.g. guide, position, destination for the application navigation) that can be selected manually (see Figure 4.14). It would be consistent to follow the principle “what you see is what you can say”, i.e. to select terms for the graphical interface that are at the same time speakable. This does not mean that spoken interaction is restricted to what is manually possible but it provides a first basis for users who are unfamiliar with the system.



Figure 4.14: Display state of a prototype in the application navigation

In case the display contains terms representing unspecific input for the speech interface (e.g. destination (Ziel)), the system should be able to process them and offer the user corresponding options to keep interaction going (cf. Chapter 4, 4.4).

For functions that cannot be controlled by speech (e.g. sound (Klang)) at least the corresponding term “sound” should be speakable (Figure 4.15). The system may then inform the user that this application can only be used manually. It can thus be avoided that the user says “sound” whereupon the system produces recognition errors leading to wrong system behaviour the user does not understand.

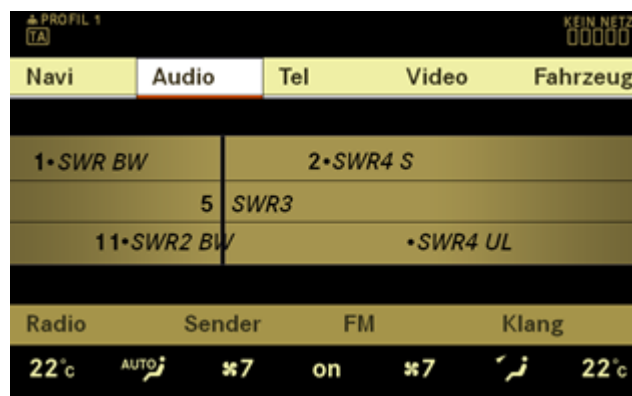


Figure 4.15: Prototype display state in the application audio

+ Synchronise system states.

While driving it might be spontaneously necessary or adequate to change from manual to spoken interaction or vice versa. It might also be the case that manual interaction is no longer possible after having exceeded a certain speed limit. To make this change easy and comfortable for the user it is important to keep system states synchronous, i.e. there must be *one* visible system state to enable changing modalities at any time. Consequently all input modalities need to have the same state of knowledge.

Imagine the user is in the application telephone intending to dial a number. The number may be entered manually or by speech. Having finished the phone call the dialled number remains on the display. Now imagine the user wants to store this number but re-accessing it is only possible manually; by speech he would have to repeat the number again even though it is already on the

display. This would be irritating and time-consuming. A smooth flow between the two modalities must guarantee that the displayed number may be accessed both manually and by speech. Alternatively, the number could be deleted from the display once the user has hung up.

+ Adjust dialogue flow of speech and manual interface.

Manual and speech specification need to be adjusted to each other such that the same dialogue flow is used. In this way the user acquires a process once, either by manual or spoken interaction, and is then able to transfer it to the other modality in an analogous way.

Take the process of storing a destination where the user is confronted with two different procedures on the speech and haptic side. Using spoken input the user is able to store a destination directly after having entered it. With manual input however an additional step is required prior to storing a destination, namely to start route guidance to the corresponding destination. To the user this kind of system behaviour is not comprehensible. An identical specification of both modalities would enable storing a destination after having input the corresponding address.

+ Establish a common vocabulary base for spoken and manual interaction.

It requires additional cognitive load if vocabulary contained in submenu lines of a display varies extremely from what is offered on a teleprompter. One reason might be that short expressions are better suitable for a display whereas for speech as input modality recognition rate is better the longer the expressions. Unless this problem is given, different terminology should be avoided.

Take an address book for example. Teleprompters visually transfer to the user a selection of speakable commands. In the address book possible utterances could for example be “store name” or “open <John Q. Public>”. On the submenu line of the address book however, these terms do not exist and thus may not be selected manually. The terms are named differently: “new entry” and “search” respectively that in turn may not be selected by speech. The aim should be pursued to unify word choice such that a teleprompter becomes irrelevant since the principle holds “what you see is what you can say”. On the one hand this eases multimodal specification and on the other hand the user is not distracted by a teleprompter constantly fading in and out.

Chapter 5

Accessing large databases using in-car speech dialogue systems

The world is complex, and so too must be the activities that we perform. But that doesn't mean that we must live in continual frustration. No. The whole point of human-centered design is to tame complexity, to turn what would appear to be a complicated tool into one that fits the task – a tool that is understandable, usable, enjoyable.

(Donald A. Norman, Interactions)

Over the past years electronic on-board devices in cars have permanently increased in number and complexity – involving an increasing risk of driver distraction. In order to minimise distraction speech has become an important input modality in the automotive environment. As a consequence of the increased functionality, however, also parts of in-car dialogue systems tend to have lost transparency (Mann, 2007a). The number of music titles, for example, has risen sharply. In former times one only had the possibility to listen to audio CDs, i.e. the number of available music titles in an in-car application was very small. Nowadays, we can also have various media carriers (see Chapter 5, 5.1.1) comprising compressed audio data which makes the number of selectable titles rise sharply. The devices are easily integrated (see, e.g., the Mercedes-Benz Media Interface (Automobilsport, 2008)). This in turn aggravates the user's

possibility of accessing particular data for he might neither correctly remember *all* data nor which titles were stored on which media carrier.

The “success” of modernity turns out to be bittersweet, and everywhere we look it appears that a significant contributing factor is the overabundance of choice (Schwartz, 2004, p.221).

Similar problems occur in context with in-car address books. The number of address book entries available in the car nowadays also rises sharply. This is due to the fact that in-car address books can be compiled from several different sources such as the on-board address book, organisers, mobile phones and PDAs (cf. Chapter 4, 4.4.7). As far as the application navigation is concerned navigation databases nowadays include a growing number of points of interest (POI) subsumed under up to 80 categories (Berton, 2007, p.155).

There is a clear trend towards text enrolments, a feature providing speakable text entries by automatically transcribing dynamic data. However, only few in-vehicle speech dialogue systems provide this option so far. This is due to recognition problems. The more entries a music database, a navigation database or an address book etc. has the higher the confusion rate of a recognition system gets.

Considering the growing amount of media, navigation and address book data, current methods of navigating them are no longer sufficient to meet customer demands. These methods comprise speech commands like for example ‘next entry (song | petrol station | hotel)’, ‘previous entry (song | petrol station | hotel)’, selecting the corresponding line number or manually searching long lists or storage devices. Unlike the vocabulary for the functions and submenus of an in-car speech dialogue system as well as the voice enrolment feature storing retrieval indices for address book entries, radio stations, destinations etc. in form of voice templates (cf. Chapter 4, 4.4.7), items of the above databases cannot be directly addressed by speech. This runs counter to the principle “what you see is what you can say” of the previous chapter. Besides, it confuses the user if superordinate categories such as artist, title, hotel or petrol station are speakable whereas the content of the corresponding categories is not.

This chapter presents an approach that offers a more user-friendly way of interacting with large databases, in particular when the user's inability to remember large amounts of data is taken into account. To begin with, the approach focuses on music data and is later extended to the applications navigation and address book.

5.1 State-of-the-art of in-car audio applications

Various approaches for speech-based access to audio data have already been published. One approach comprises accessing every database item within one utterance (Wang, 2005). Prior to speaking an item the user is required to enter the corresponding category, as for example in "play album Automatic for the People". Thus, recognition space can be pruned effectively. The approach requires the user to know the complete name of an item and does not provide options for category-independent input. Another approach is followed in the TALK project (TALK, 2007). By means of a complex disambiguation strategy it allows for speaking any item any time. It has not been proven successful for more than a few hundred songs.

Considering the large amount of audio data users will bring into their cars the aim of the following approach is to handle large vocabulary lists by means of category-based and category-free input of items. Additional wording variants, generated by means of generating rules, allow the user to speak only parts of items stored under various categories such as artist, album, title, etc. This will reduce cognitive load and improve user acceptance, since the user does not have to remember the complete name.

5.1.1 Constraints

Automotive audio systems nowadays are equipped with a variety of storage devices. These devices in turn may comprise different data formats as well as various file types (see Figure 5.1). Speech as an alternative input modality to manual input ought to guarantee the driver to keep his hands on the wheel and his eyes on the road.

However, as Figure 5.1 shows, this diversity leads to an enormous complexity. The user needs to have an overview of the technical structure and has to memorise the corresponding contents when selecting audio data. Such an application is not transparent to the user and demands too many cognitive resources while pursuing the driving task. It is therefore necessary to provide a

concept that reduces mental load by allowing the user to select audio data without previous knowledge of technical devices plus the data they contain.

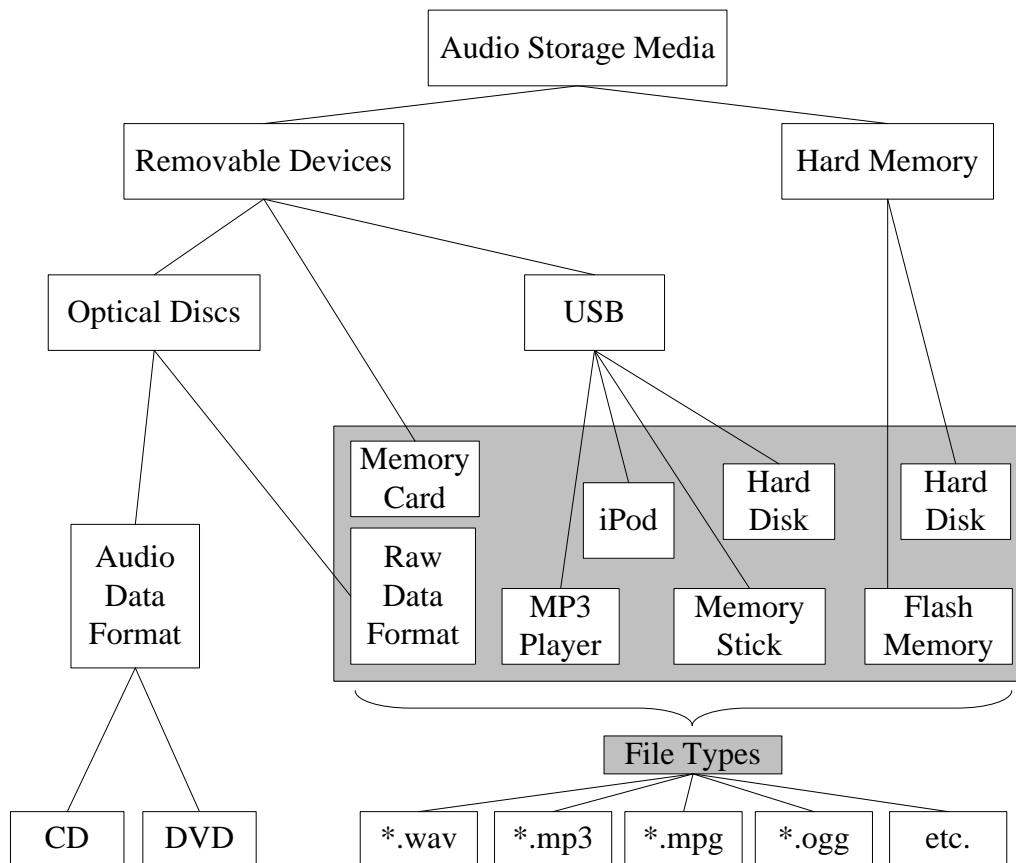


Figure 5.1: A taxonomy of audio storage devices

When an audio database is accessible by means of text-enrolments, i.e. speakable text entries, the problem arises that these entries have only one phonetic transcription. This means they can only be selected by speaking the complete name of a title, album, artist etc. In case the user slightly varies input (which might be due to lacking knowledge of the precise name), the corresponding entry cannot be selected by the system. This turns spoken interaction into a difficult task. Evidence for this assumption is taken from the user study on personal address book data described in Chapter 4, 4.4.7, analysing to what extent users remember the entries they have stored in their address book. The findings of this study showed that only in 39% of the cases was there a correct match between the speech commands uttered by the user and what had actually been stored in the address book (see Figure 4.10). The majority of 61% remained undetectable by speech due to lacking recollection.

It is obvious that similar problems will occur when it comes to selecting audio data, i.e. users often do not remember the exact name of titles, albums or other categories. Consequently the user might be tempted to switch from spoken interaction to manually navigating through hierarchies in order to accomplish a task rather than concentrating on the actual driving task.

To reduce cognitive load and ensure that the driver can concentrate on the traffic situation, it is necessary to offer a more advanced audio data retrieval that requires neither previous knowledge of technical devices and their corresponding audio data, nor the data's precise wording.

5.1.2 User needs

A user study (Rosendahl, 2006; Mann, 2007a) examined customer expectations with regard to an in-car search mode for various applications. The study comprised 21 subjects of three different age-groups: <35, 35-55 and >55 years. One of the main conclusions was that the participants generally like the idea of an in-car search engine for data such as audio, phone book or navigation – provided the design is as follows:

- It must be simple and intuitive
- It should work without a tutorial
- It should be an extension to what people are used to
- It should be efficient and time-saving
- It should be possible to select a specific category prior to activating the search function
- Favourites should be available
- Result lists should be restricted

The following approach therefore allows accessing audio data on different media carriers and in various formats in a uniform way. The underlying methods enable both expert and novice users to accomplish tasks in a shorter period of time than with current systems.

5.2 Interaction concepts for searching audio data

The previous chapter pointed out the difficulties occurring with in-car audio application management. Speech as interaction mode has the purpose to preserve the driver's safety. Therefore the aim is to design dialogue such that the user may complete tasks in a simple and

intuitive way. This chapter proposes a new method for handling increased functionality as well as large amounts of audio data.

5.2.1 Category-based and category-free search

In order to bring back transparency into the multitude of technical audio devices it takes an approach that allows accessing audio data from various media carriers and different formats in a uniform way. To achieve this three different interaction concepts are suggested: category-based search, category-free search and physical search.

Category-based search requires pre-selecting a category. A set of five categories was defined (artist, album, title, genre and year) that is of interest to the user and usually available in the metadata of audio files and two additional views on audio data: folder view and title list view.

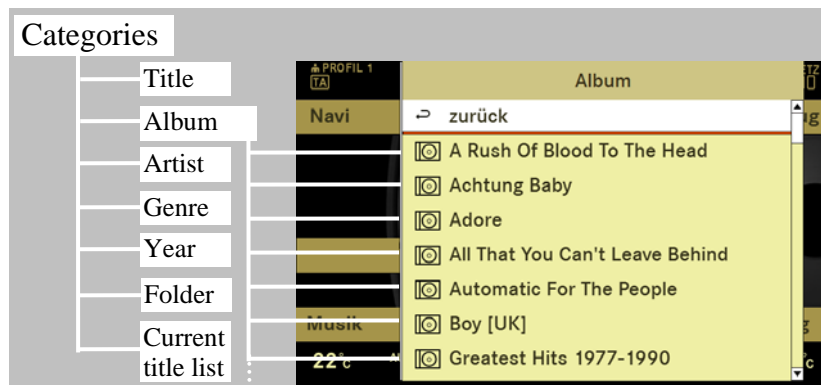


Figure 5.2: Category-based search

Each category contains data of all audio storage devices. When selecting one of the above categories, e.g. ‘album’, the user is returned a list of all albums from all connected storage media in alphabetical order (see Figure 5.2). Thus, the user does not have to go through technical devices such as MP3 players or memory cards including the embedded hierarchies to find the desired album. The result list may be scrolled through by manual or spoken input. When choosing a particular item from the list the user may do so by directly speaking the album name. This option is provided for by means of speakable text entries (i.e. text enrolments): for example, the user can say ‘Greatest Hits 1977-1990’ or simply ‘Greatest Hits’ (cf. Chapter 5, 5.2.2). The user is allowed to speak any item of the list, not just the ones currently displayed.

Global category-free search is independent of any pre-selection. Once the user got into this search mode he may enter a title, an album, an artist, a folder, a year or a genre by speaking its complete name (e.g. ‘A Rush of Blood to the Head’ or ‘Alternative Rock’) or parts thereof (e.g. ‘A Rush of Blood’). As in the category-based search the system considers the contents of all audio storage devices. Regarding these large amounts of audio data, uncertainties are likely to occur. They are resolved in two ways. In case the uncertainty of a user’s input is within one category, a result list containing the corresponding items is returned (see Figure 5.3 (left)).



Figure 5.3: Category-free search – multiple results within one category (left); resolution proposal for multiple results in different categories (right)

In case the uncertainty spans more than one category – ‘No Need to Argue’ for example could either refer to an album or a title by ‘The Cranberries’ – a supplementary step is added providing the user with a list of the corresponding categories plus the respective number of hits (see Figure 5.3 (right)).

Physical search ensures backward compatibility and provides a fall-back solution for users wishing to navigate within the contents of particular audio storage devices.

5.2.2 Fault-tolerant word-based search

Section 5.1 presented the difficulties users have in precisely recollecting large amounts of data. Additionally to the large number of in-car audio files, the structure of music file names such as artist, album and title is manifold and sometimes fairly complex.

Artist: 'Neville Marriner: Academy Of St. Martin In The Fields'
 Album: 'A Collection of Roxette Hits! - Their 20 Greatest Songs!'
 Title: 'Bach: Orchestral Suite #2 In B Minor, BWV 1067 - 3. Sarabande'

Consequently, if the user wants to select particular items by speech using the above search concepts it would not be sufficient to provide text enrolments with merely one wording variant per item. Because then, user input that might be far more likely compared to the available wording variant could not be recognised by the system (e.g. 'Laundry Service' instead of 'Laundry Service: Limited Edition: Washed and Dried'). This leads to frustration and driver distraction with the consequence that the user ends up using the manual speller.

The approach therefore allows selecting audio data by speaking only parts of complete names. To create additional useful wording variants for parts of items the available audio data are pre-processed by generating rules (Mann, 2007b). Items of all categories are decomposed according to rules such as follows (cf. Chapter 5, 5.5.2):

1. Special characters such as separators and symbols are either discarded or converted into orthography. Alternatively they may be used as separators to create additional wording variants.

Africa / Brass	Africa Brass
The Mamas & The Papas	The Mamas and The Papas

2. Abbreviations are written out orthographically.

<i>Dr. Dre</i>	Doctor Dre
Madonna <i>feat.</i> Britney Spears	Madonna featuring Britney Spears

3. Keywords such as category names including their synonyms are discarded and therefore not obligatory when entering audio data.

The Charlie Daniels <i>Band</i>	Charlie Daniels (plus rule 4)
<i>Songs</i> of Long Ago	Long Ago (plus rule 4)

4. Closed word classes such as articles, pronouns and prepositions are detected by means of morpho-syntactic analysis and can be omitted in context with particular phrases (e.g. noun phrases or verb phrases).

<i>The</i> Lemonheads	Lemonheads
<i>They</i> Might Be Giants	Might Be Giants
<i>Under</i> Pressure	Pressure

5. Secondary components (e.g. of personal names) can be discarded by means of syntactic-semantic analysis.

Ludwig <i>van</i> Beethoven	Beethoven
<i>Dave</i> Matthews Band	Matthews Band
Looking for the Perfect Beat	For The Perfect Beat Perfect Beat (plus rule 4)

Each variant is then phonetically transcribed to be accessible via voice input. Shakira's album 'Laundry Service: Limited Edition: Washed and Dried' for example contains a song called 'Objection (Tango)'. For selecting this song a normal way would be the description 'the tango 'objection'' as the album contains another tango. To cover this variant the single parts 'Objection' and 'Tango' have to be combined taking into account syntactic and semantic knowledge: 'tango' describes the music category, which is used in the descriptive expression 'the tango' to select the song of this category named 'objection'.

Another example is 'Hips Don't Lie (featuring Wyclef Jean)'. This song can be segmented into the following parts: [[Hips] [Don't Lie]] [[featuring] [[Wyclef] [Jean]]]. Possible recombinations could be 'Hips Don't Lie with Wyclef Jean' | 'Hips Don't Lie with Jean' | 'The song with Wyclef Jean' etc.

Compared to manually entering a category item by means of a speller this approach is less distracting, more comfortable and time-saving.

5.3 General requirements for the user interface

In addition to the interaction concepts on large audio data presented in Chapter 5.2 the user interface is based on the design guidelines presented in Chapter 4, 4.4. It follows the general principle *what you see is what you can speak*. All text information that can be selected manually on the display can also be used for voice input. The strategy is particularly helpful for novice users who are not yet familiar with using spoken interaction. In order to synchronise speech and graphics/haptics a synchronisation component (SYNC) (cf. Chapter 4) transfers data and events between the two modalities. The user may switch between speech and manual input at every step. Combined with the above principle, the system reflects a user concept that is consistent and uniform, giving the user the impression of having only one visible system state.

In contrast to command and control systems the approach allows for spoken input that is less restricted. Rather than demanding from the user to learn a multitude of speech commands a variety of expressions covering the same meaning (synonyms) is offered. In case the user has forgotten a particular expression, he may simply pick an alternative instead of looking at the display to search for the appropriate term.

With regard to initiative the speech dialogue is either system- or user-driven, depending on the user profile. For the novice user who is unfamiliar with a task the system takes initiative, leading him through the dialogue. The more familiar the user gets with a task, the more the number of relevant turns can be reduced. To accelerate interaction expert users may apply shortcuts. Expressions such as “search album”, “search category artist” or “play music” are straightforward, preventing him from numerous steps through a menu hierarchy as is inevitable when using manual interaction.

5.4 Prototype architecture

The new approach of accessing media data by speech was integrated into a prototype system. The prototype’s architecture is based on state-of-the-art speech dialogue systems (cf. Chapter 4) connecting to the media search engine of the media application (see Figure 5.4). Since audio data on external storage devices might vary significantly the system needs to be capable of handling dynamic data. As the size of audio data may be quite large, a background initialisation process has to be implemented.

The dialogue system is a multimodal interface (cf. Chapter 2) with two input and two output modalities: manual and speech input, and graphical and speech output. In order to accept spoken input, understand and process it and answer appropriately speech control comprises the following modules: a task-driven dialogue manager (TDDM), a natural language understanding unit (NLU) containing a contextual interpretation (CI), an automatic speech recogniser (ASR) and a text-to-speech component (TTS), which includes a grapheme-to-phoneme (G2P) converter. All speech control modules are subject to the configuration of a common knowledge base (Ehrlich, 2006).

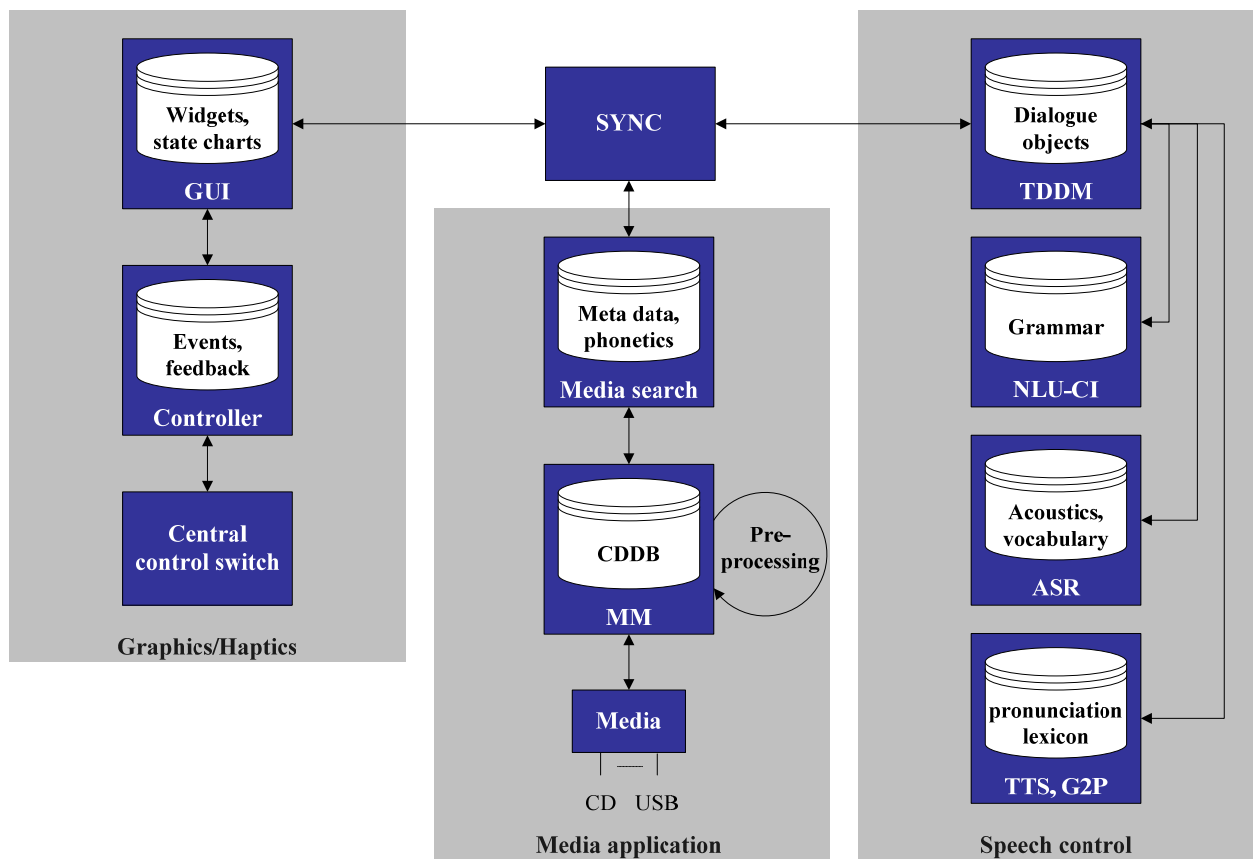


Figure 5.4: Prototype architecture view

The graphics-haptics interface consists of a module for the visual component, i.e. the graphical user interface (GUI) and a central control switch for manual interaction (cf. Chapter 4, 4.1). The controller module contains the interface to the central control switch and is responsible for the event management.

The synchronisation module (SYNC) connects and synchronizes the spoken and visual world. It is also the unique interface to the external media application.

The media application consists of the media search engine and the media manager (MM). The latter administrates the connected media. Considered are an internal hard disk and DVD drive, as well as external memory cards and MP3 players connected via USB. The MM relies on a media database, such as Cddb, which contains metadata and as many phonetics of the metadata as possible. Non-existing phonetics are generated by the language-dependent G2P engine of the speech component. The MM transfers the metadata and corresponding phonetics to the media search engine which includes the database of all metadata of the connected media. The search engine implements a database with interfaces to quickly search it by words and parts of words.

Pre-processing metadata for speech operation enables the system to also understand slightly incorrect names, nicknames, parts of names and cross-lingual pronunciations.

Slightly incorrect names are handled by filtering out insignificant particles at the beginning. ‘Beach Boys’ thus becomes a wording variant to ‘The Beach Boys’. Nicknames are a more complicated concept as it requires access to a database, such as Gracenote MediaVOCS (Gracenote, 2007). They allow for selecting the artist ‘Elvis Presley’ by saying ‘The King’.

Providing good phonetic transcriptions for all dynamic metadata of all audio files on the connected devices is one of the greatest challenges. Additionally the pre-processing should provide phonetic transcriptions for parts of names, i.e. alternative expressions of the original item. Internationality implies that music databases normally contain songs in various languages. Thus the system must be able to handle cross-lingual phenomena, which includes phoneme mappings between language of origin (of the song) and target language (of the speaker). To allow for that a two-stage algorithm is followed:

1. The phonetic representation of a song is looked up in the music database (Cddb), which contains the phonetics only in the language of origin. If the song is available, the phonetics of the metadata are used and also automatically mapped into the phonetic alphabet of the target language, so that ASR includes both pronunciation variants.

2. In case the metadata of the song in question do not exist in the database, the approach has to rely on G2P. The system contains G2P modules for all languages on the market, e.g. American English, Mexican Spanish and Canadian French for North America. Phonetic transcriptions are provided for all three languages using the corresponding G2P. The phonemes of all languages are mapped to the target language to generate pronunciation variants covering speakers not familiar with the foreign language in question.

Speech output is done by multi-language TTS (again, all languages on the market). If phonetic representations are available in the database, they are phonetically mapped to the phoneme set of the target language and then output. If a phonetic representation is not available, the name is language-identified and transcribed in a similar way as for ASR. That enables the system to speak any item as close as possible to its name in the language of origin, or if not possible due to technical restrictions, as close as possible to its name in the target language.

5.5 Verifying generating rules

The approach presented in Chapter 5 improves interaction with voice-operated in-car audio applications containing large amounts of data from various audio storage devices. It is based on intuitive interaction concepts for searching audio data (i.e. category-based search, category-free search and physical search) enabling the user to search across all media carriers available in the car in a uniform way. Rules for pre-processing metadata of all audio data allow user-friendly access to audio data by speaking only parts of category items such as artist, album, title, etc. instead of having to remember the exact wording of all items.

The following work focuses on testing to what extent an approach that allows speaking parts of audio file names performs better. In particular, it aims at analysing in how far the pre-processed metadata (wording variants) cover what users input via spoken interaction when searching for audio data (Mann, 2008b).

To verify how people select music and how much they actually remember of the names of what they want to listen to, a survey was conducted to collect user speech data.

5.5.1 The data elicitation method

To collect a test speech database of audio file names the designed survey (Mann, 2008a) combined various scenarios, ranging from completely unrestricted input of music file names to recollection of given audio files names. When progressively restricting spoken input the number of categories within one utterance was also extended to get speech data for straight-forward selection of combined category values such as 'Roxette Greatest Hits'.

The aim was to see how users actually input what they want to listen to and how distinct their knowledge is on the one hand. On the other hand, the collected data should provide a basis for testing how well the applied set of generating rules covers spoken user input and how well recognisers perform with respect to large numbers of audio titles.

The survey should elicit items from common categories such as artist, album, title, genre, year and audio books. However, these items should not be restricted to one category per input, but - to a certain extent - they should also contain combined categories such as 'Rock from the sixties'. Combining these aspects the survey came up with three different tasks for the speech recordings.

For each task, subjects were seated in front of a computer. Using a PowerPoint presentation, the proceeding of which lay in the hands of the subjects, the subjects were first presented with an introduction, using concurrent script display and playback of pre-recorded spoken instructions.

Task 1 – to get an impression on how people behave when there are no restrictions given on how they formulate a query for a certain music title, the survey was started with a task concerning free input. The task was split in two scenarios (see Figure 5.5). In scenario one no restrictions were set such as not to prime the subject in a particular direction. The aim was to find out whether the categories commonly used in metadata (i.e. artist, album, title, genre and year) are sufficient for the subjects when selecting music or not. And if not, what additional categories do they come up with? It was also of interest how they express their wishes in phrases or sentences. Scenario two provided the subjects with icons of various categories to restrict them to the categories available in state-of-the-art music collections. The subjects could select and combine the categories at will.



Figure 5.5: Auditory and visual instructions for task 1⁷

Task 2 – the second task was designed such as to find out how familiar users are with categories in the context of music. The subjects were therefore asked to input individually favoured titles according to given categories. This implied giving examples according to one category as well as combinations of two categories. The task did not cover all combinations across all categories but only those that seem plausible according to common sense. Therefore, whenever two categories had to be combined within one utterance, the subjects were requested to stick to the given sequence. Figure 5.6 illustrates single and combined category input of audio file names.

Task 3 – this task demanded reproducing given audio file metadata in pairs of two with intended cognitive overload (see Figure 5.7). Depending on the file names' length, each pair was fading out on the screen after a certain period of time. The subjects then had to repeat each file name according to what they remembered and considered plausible. By not allowing sufficient time to learn the presented items by heart, the subjects were led to filtering crucial parts of audio file

⁷ The instructions for scenario 1 ask the subject to imagine a conversation with a human driver to select music. Scenario 2 asks for a selection of music by category. Note that the icons are explained by paraphrases such as to avoid using category names.

metadata they are presented rather than trying to remember and reproduce each of them completely. The aim was to have subjects create a diversity of audio file names. The presented audio file names covered the languages English, German and, to a minor extent, French, Italian and Spanish.



Figure 5.6: Single (left) and combined category input (right) in task 2



Figure 5.7: Procedure for reproducing given audio file names in task 3

Both task 2 and 3 were designed such that the speech recordings can be used to eventually test the assumptions against a real ASR system. Therefore the liberty of subjects as to which words to use was increasingly limited from task 1 to task 3.

Great care was taken to avoid any keywords in instructions that might influence the choice of words by the subjects. Therefore, only icons were used to represent the different categories for

task 1. Apparently, the choice of the icons was good, as only few of the subjects not did recognise the intended category immediately. In the instructions for tasks 2 and 3, however, no paraphrasing was used, as the subjects in this stage of the experiment had already established their own names for categories from scenario 2 of task 1.

When subjects were ready to speak and clicked on the next item, a signal tone indicated to them that the system was ready for recording.

By presenting the items to be memorised for repetition for only a short time (3 to 9 seconds), the subjects were deliberately put under cognitive stress. It is assumed here that under cognitive stress, perceptive pre-processing in humans has a filter effect similar to the transfer of memory contents from short-term to long-term memory, viz. that pre-processing also abstracts the salient features of items from the background. In both cases similar parts of the presented items are either stored or forgotten. So the verbal description for music titles, album names etc. is according to the motto: what parts are worth remembering? Which parts go through the filter and thus can be reproduced?

Questionnaires

The subjects were asked to fill in a pre-experimental questionnaire (see Appendix B) on personal data such as demographics (e.g. dialect, gender, age), experience with speech-operated systems and experience within the audio domain (e.g. 'what does your personal music collection consist of', 'how many titles does your music collection comprise' or 'do you often listen to music while driving'). It was followed by speech recordings according to the above tasks 1-3. In case a subject had problems in giving examples, a certain number of audio file metadata was provided as fall-back solution. The session ended with the subjects filling in a questionnaire on the experiment (e.g. 'how pleasant did you find free input / input via category' or 'what kind of input do you prefer when selecting music'). In total, each session took 45-60 minutes.

Subjects

For collecting audio file names a subject base of 30 people was recruited. It comprised 16 male and 14 female subjects of 5 age groups (18-25, 26-35, 36-45 and 56-65 years). The majority of the subjects spoke standard German with a Swabian (south-western German) dialect colouring. All of them owned a music collection comprising at least one of the following audio storage

media: CD, MP3, audio DVD and vinyl records. For 70 per cent of the subjects the collection did not exceed 2000 titles. 80 per cent had MP3s. As anticipated, the number of frequent MP3 listeners decreased with increasing age of the participants (see Figure 5.8).

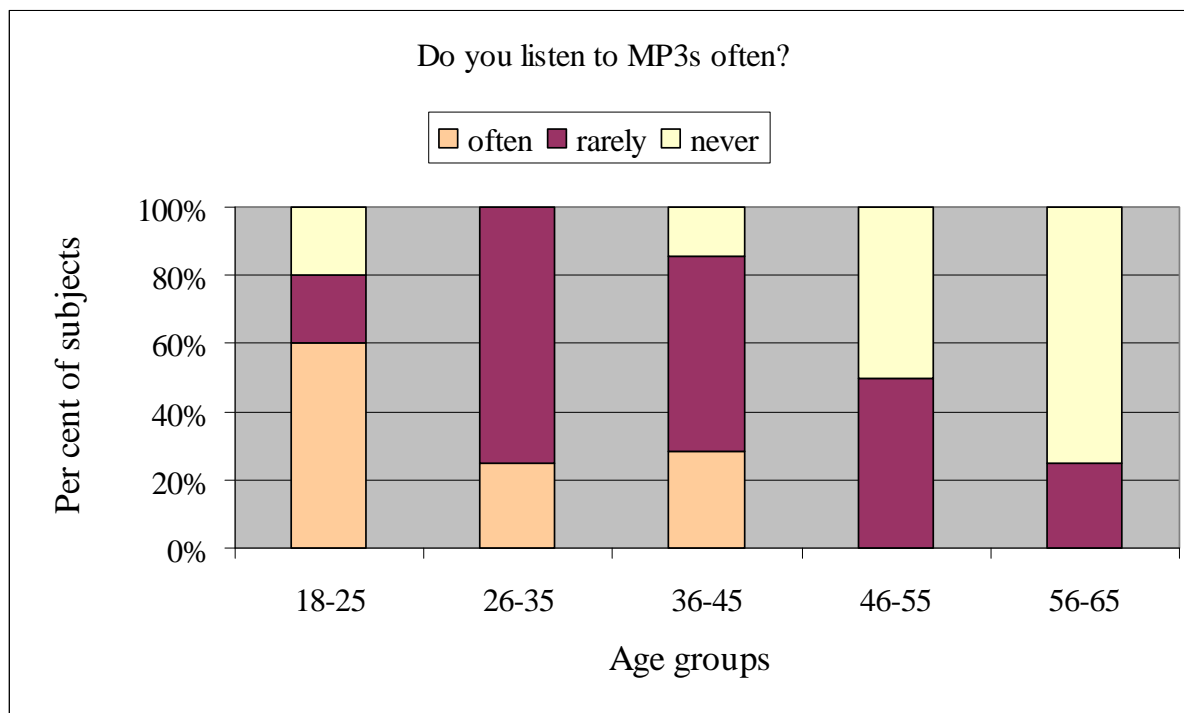


Figure 5.8: Correlation between MP3 listeners and age

5.5.2 Results

The recordings of 30 subjects resulted in approximately 15 hours of data. The utterances of each subject were then orthographically transcribed according to what has been said (e.g. 'Kuhn und Band'). The original name referring to the file in the database was also added (e.g. 'Dieter Thomas Kuhn & Band'). Table 5.1 gives examples of what has been said in the corresponding tasks.

Task	Spoken utterance	Original name
T 1	Alles von Bruce Springsteen	Bruce Springsteen
	Ich möchte gerne was schnelles Fetziges hören	Latin (Salsa)
	Ich hätte gern irgendwas aus den Siebziger Jahre	Siebziger Jahre
	Die Dire Straits bitte	Dire Straits
	CD Du bist Herr	Du bist Herr
	Was von Bon Jovi	Bon Jovi
	Ein Lied von ABBA	ABBA
	Das erste Album von Hillsong	The Power of Your Love
T 2	Mozart die kleine Nachtmusik	Wolfgang Amadeus Mozart – eine kleine Nachtmusik
	Von den Toten Hosen Kauf mich!	Die Toten Hosen - Kauf mich!
	Unplugged Leyla	Unplugged Leyla
	Tchaikovsky	Peter Ilyich Tchaikovsky
	Andrea Bocelli Time To Say Goodbye	Andrea Bocelli & Sarah Brightman Time To Say Goodbye
	Best Of Andrew Lloyd Webber	The Best Of Andrew Lloyd Webber
	Bach Toccata und Fuge	Johann Sebastian Bach Toccata und Fuge d-Moll BWV 565
T 3	Peer Gynt	Solveig's Song Peer Gynt-Suite No. 2
	Der Ketchup Song	The Ketchup Song (Aserejé)
	Boy U2	Boy [UK] U2
	Everybody Else Is Doing It	Everybody Else Is Doing It, So Why Can't We?
	Fréhel	Si tu n'étais pas là (Fréhel)
	London Symphony Orchestra	London Symphony Orchestra; Peter Maag
	Münchner Bach Orchester	Karl Richter; Munich Bach Orchestra
	Douce Guillemette	Tant vous allez, douce Guillemette

Table 5.1: Examples of spoken utterances during the experiment including their original names

90 per cent of the subjects often listen to music while driving. 40 per cent often listen to audio books. Impressions during the recording sessions and the evaluation of the data showed that the subjects experienced certain difficulties. Whereas artist and title names were remembered fairly well, the recall of album names was, as a subjective impression, much weaker than anticipated. It became even more difficult when the subjects had to combine an album name with a corresponding title.⁸ Also, people did not seem to feel at ease when it came to using genre names. The majority of the subjects are only familiar with classifying up to ten major genres, but not the diversity of hundreds of micro-genres, such as “female alternative vocal pop”.

Knowledge concerning classical music was on the average very limited, even though 22 per cent of the subjects like classical music while driving. As audio file names of this genre are generally longer compared to those of other genres, the gap between subjects’ utterances and original names was quite large.

With regard to languages half of the user utterances were English, one third was German and the remaining part was evenly distributed between Spanish, French and Italian. More than in English, which all of the subjects spoke as a second language to a certain extent, the participants had pronunciation problems for French, Italian and Spanish music titles.

Matching actually spoken utterances

To verify the hypothesis that people do not randomly remember or forget parts of music title names, the actually spoken utterances of all single-category items (1847) were first of all matched against the original names, i.e. against the target stimuli as they were presented in tasks 2 and 3.

It was found that in the category artist, 61% of the subjects’ utterances matched directly, indicating that people had spoken the 'correct' name (see Figure 5.9). This is little surprising, as

⁸ It can be assumed that the concept of a music title ‘belonging’ to one particular ‘album’ (or long-playing record) is vanishing. Music collections, re-distribution in different samplers etc. disassociate an individual title from its accompanying titles. However, the subject base is way too small to prove or disprove such assumptions.

in many cases these names consisted of only one word or were in themselves well-known, so that no extra memorization on the parts of the subjects had to take place and the 'full' name could be directly retrieved from long-term memory.

Genre names proved to be similarly short. 87% of the spoken utterances directly matched to the presented stimuli. Also, years and decades are subsumed here, the memorization of which also does not require a particular, directed effort.

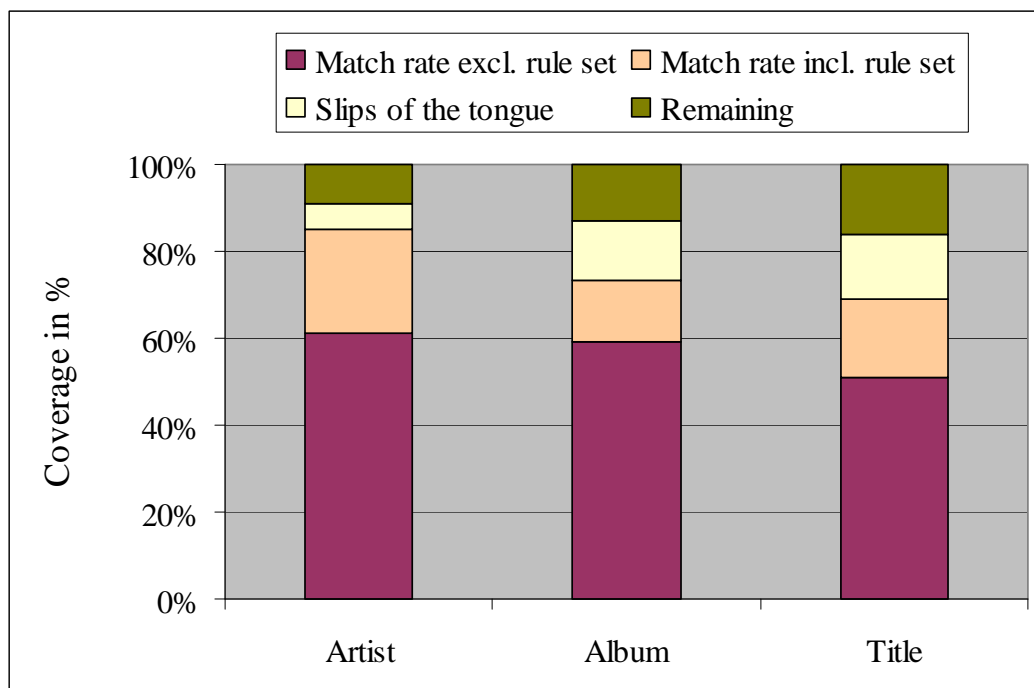


Figure 5.9: Coverage of spoken utterances

When it comes to the names of albums, the initial match rate is 60%. For the category title, however, the rate of direct matches is a little lower (51%): people did remember these fairly complex names with more difficulty.

After establishing the initial match rate, a set of generating rules, derived from intuition and introspection, was applied to the set of 'original names' in each category (see also the approach described in Pfeil, 2008). Basically, these rules are based on omission, permutation and insertion that create variants (or paraphrases) of the 'original names' and increase the cardinality of the set against which the match takes place (see Appendix B.2). For the matching reported here, care

was taken to restrict the rules to a manageable number. The rules mainly apply to artist and title names. Their corresponding vocabulary more than doubled (see Figure 5.10). They also apply to album names often, however, in this case the number of generated album names did not exceed 81%.

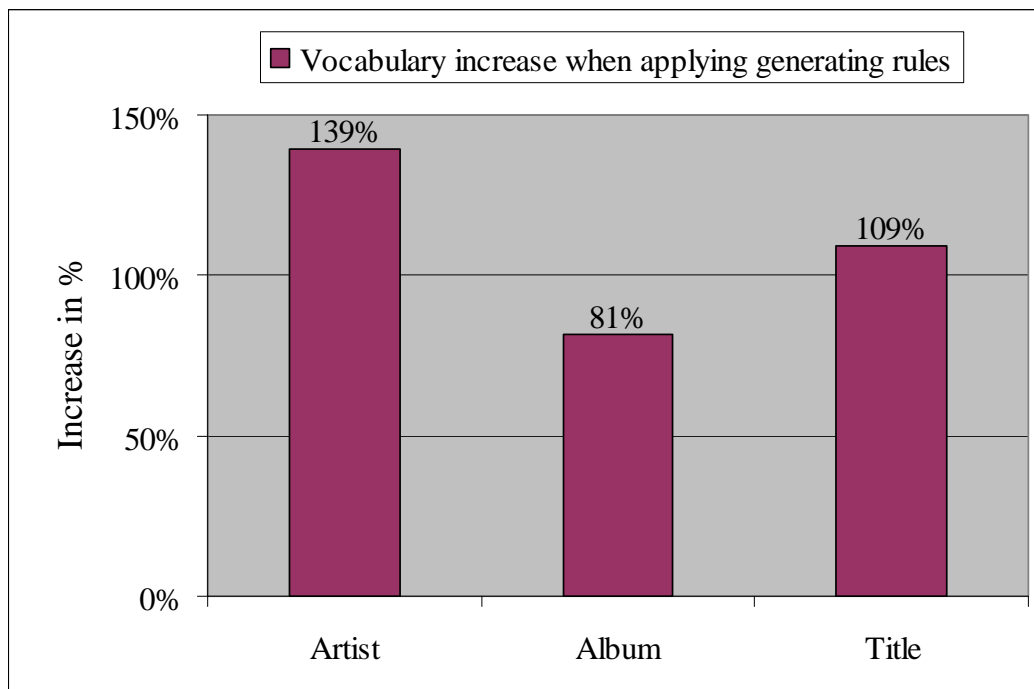


Figure 5.10: Increase of vocabulary (original names) due to generating rules

For the artist category, this set of rules proved very successful: with just 15 rules, 85% of the subjects' utterances could be covered. Again, this is attributed to the general shortness of the word combinations here, or, in other words, the relative absence of redundancy, which made it rather easy for people to memorize at least one significant part of the stimulus. If users remember only a part of an artist's name correctly, a system should be able to retrieve the desired selection list for them – at least if there is no inherent ambiguity (e.g. Elvis Presley vs. Elvis Costello).

The application of the rule set on title names increases the number of matches from an initial 51% to 69% overall, while on album names, the match rate went up from 60% to 74%. Table 5.2 shows an extract of generating rules and to what extent they successfully applied to actually spoken user utterances.

Generating rule	Example	Application in %
Artist		
Final constituent	Wolfgang Amadeus <i>Mozart</i>	14,9
Separator split	Karl Richter; <i>Munich Bach Orchestra</i>	3,7
Omit article	The <i>Rolling Stones</i>	3,0
Album		
Separator – extract anterior part	<i>Everybody Else Is Doing It, So Why Can't We?</i>	4,5
Omit article	<i>A Rush Of Blood To The Head</i>	6,0
Title		
Omit bracket constituent	<i>The Ketchup Song</i> (Aserejé)	7,1
Separator split	<i>Carmen</i> – Suite Nr. 1 (Prelude–Aragonaise)	6,0
Extract bracket content	Si tu n'étais pas là (<i>Fréhel</i>)	5,9

Table 5.2: Extract of generating rules and their application regards spoken utterances

The majority of successfully covered wording variants could be achieved by applying up to two rules in sequence (see Figure 5.11). Increasing the number of applicable rules here will increase the match rate as well, however, great care must be taken to avoid two traps here:

1. So far, the over-generation inherent in all generation approaches without a limiting factor can potentially lead to a prohibitive number of non-unique matches. While 'ordinary' ambiguities (such as in the Presley/Costello case) can always occur and do not lead to irritations on the part of a human user, ambiguities arising out of over-generation are hard to explain and can be very annoying in view of the usability of an overall system.
2. As the aim is to have an overall system using a speech recogniser, care must be taken not to increase a recogniser's lexicon to an extent where the recogniser accuracy (that is among other things also dependent on lexicon size) deteriorates because of over-generation.

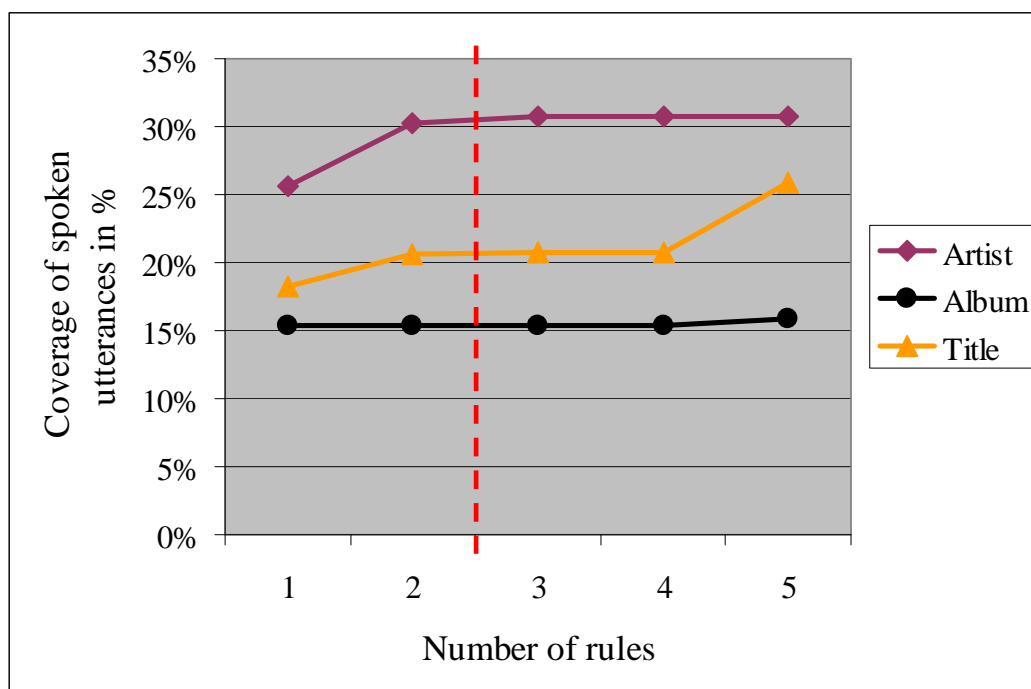


Figure 5.11: Number of rules necessary for covering utterances, i.e. wording variants

Overall, the analysis proves that generating rules significantly contribute to including the right wording variants into the vocabulary, particularly for artist names. The overall coverage of generated album and title names remains rather low due to the fact that the subjects did not know a significant amount of names, particularly foreign names that were presented and thus came up with mispronunciations and short names without knowing the context of the foreign language (cf. Figure 5.9).

Questionnaire analysis

Analysing the questionnaires showed that only three per cent of the subjects do not have difficulties when it comes to recollecting music, i.e. audio file names (see Figure 5.12). This rating is also reflected in the speech data derived from the recording sessions. Accordingly, with respect to ‘knowledge’ of music, the subject base is split into three types of listeners:

- The “music connoisseur”: owns large and diverse music collection. Most of the time precisely aware of artist, album and title name. Addressing music according to these categories is totally sufficient, expert in recollecting audio file names.

- The “average” music listener: knows an average amount of audio files according to their category names such as artist, album, title and genre. However, when selecting particular music often does not remember the audio file name precisely.
- The ”passive” music listener: likes listening to music, coming either from the radio or from own music collection. However, when confronted with particular artist and/or title names, would not know them even though they might be part of own collection. When selecting particular music most of the time does not know artist name or title. Has difficulties correlating audio file names with respective music.

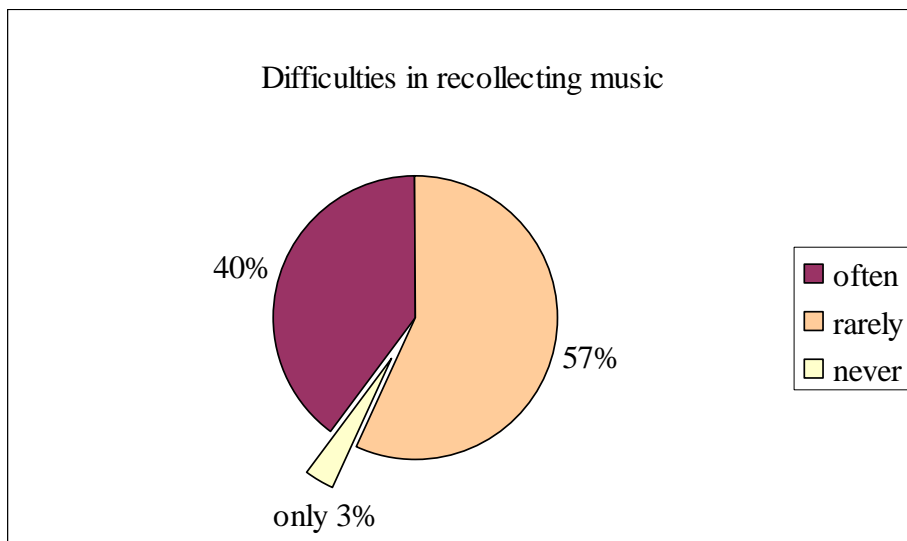


Figure 5.12: Subjects’ knowledge concerning audio file names

Subsequent to the speech recordings each participant had to subjectively evaluate the approach and their input of audio file names. When it comes to preferring one particular input type, categorical and free input are roughly equal: 53% of the participants preferred categorical input and 47% did so concerning free input.

5.6 Combined-category input

As an add-on to the above prototype it was envisaged to extend the number of voice input parameters within one utterance to allow for straight-forward combination of categories, like for example ‘Shakira Objection’ (Dausend, 2008).

Schulz (2004) presents an approach based on VoiceXML to reduce the number of dialogue steps. To achieve this up to five features may be combined within one utterance. The evaluation was restricted to a lexicon comprising 200 entries. The survey for collecting a test speech database of audio file names described in section 5.5.1 however showed that the subjects basically did not combine more than two features, i.e. categories, within one utterance. A combination of two categories occurred quite frequently, such as ‘title & artist’ or ‘genre & year’. To verify an approach allowing for combined category input, tests were conducted with a multilingual recogniser. The basis was provided by a database of 2000 music titles. This size has been selected according to the survey’s results on subjects’ personal music collections: for the majority of the subjects (70%) the collection does not exceed 2000 titles (see Figure 5.13).

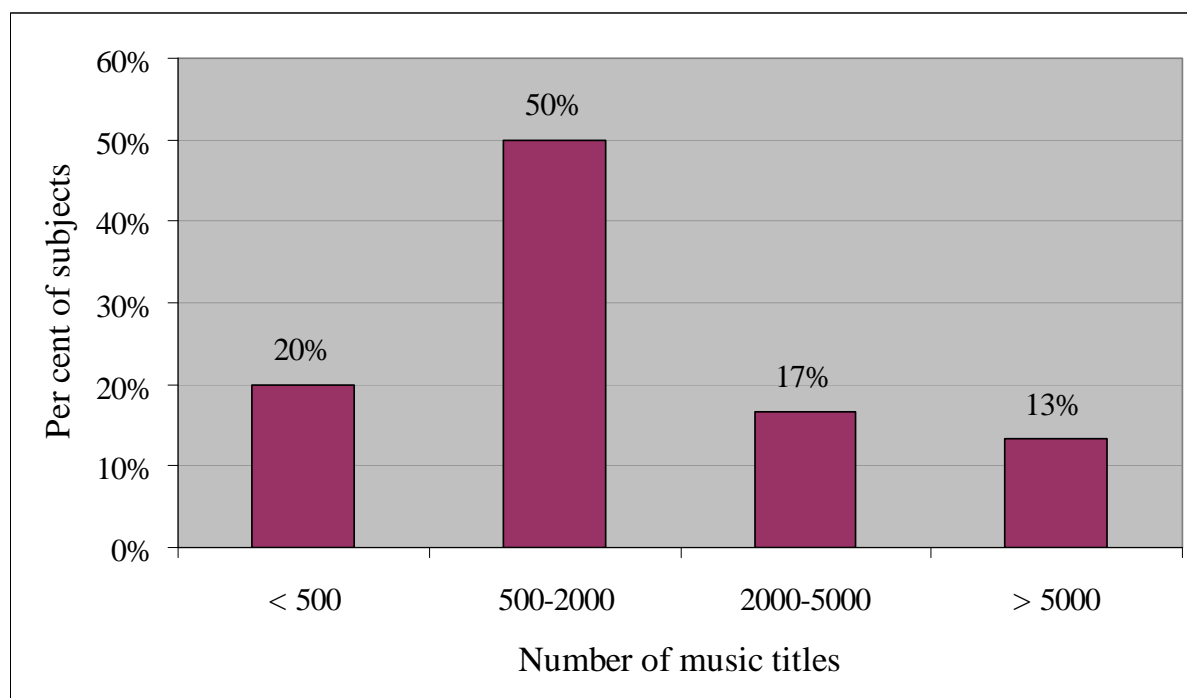


Figure 5.13: Size of personal music collection according to titles

The corresponding vocabulary for the recogniser was extracted from ID3-tags. The sample for the tests was taken from tasks 2 and 3 of the speech recordings described in section 5.5.1. It consists of 1847 recordings with one item and 912 recordings with two items. Two thirds of the data are foreign-language, mainly English.

For the recognition tests three options for inputting music were provided: obligatory input of two items with fixed sequence, i.e. title and artist (2ItemsTA); obligatory input of two items with arbitrary sequence, e.g. title and artist or vice versa (2ItemsFree); and optional input of one or two items with arbitrary sequence which obviously is the most comfortable one (1||2ItemsFree). As reference value input of one item (e.g. title) was taken.

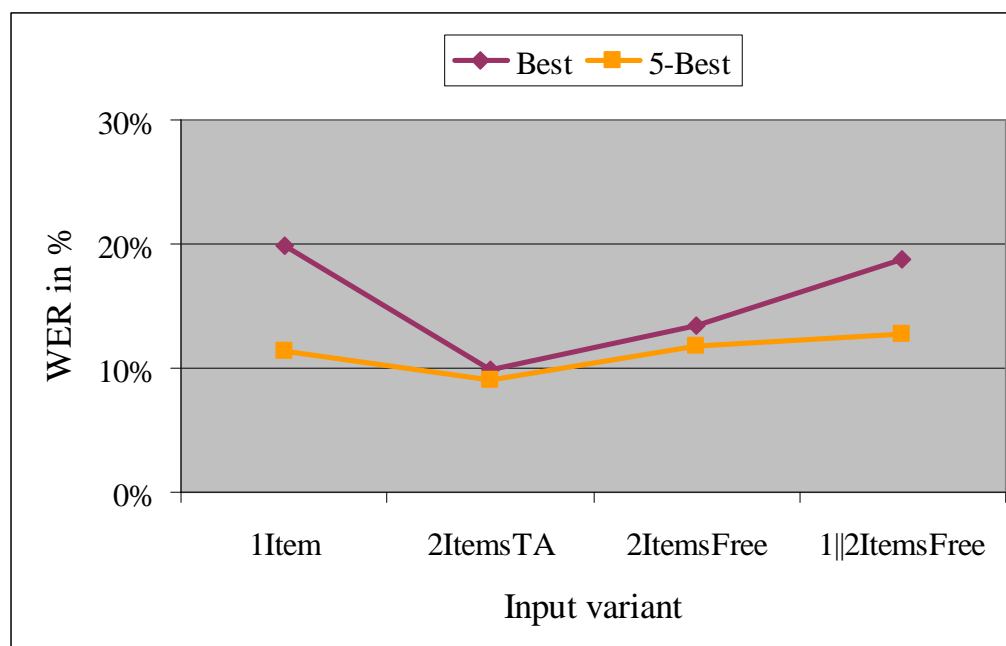


Figure 5.14: WER of different input variants with a lexicon comprising 2000 titles

Figure 5.14 summarizes the results. The word error rates (WER) are fairly high. This is due to the fact that the data are to a large extent foreign-language. The recogniser was a German recogniser extended by phonemes of various foreign languages. However, these foreign phonemes were not mapped from the language of origin into the target language, i.e. German. This means that for two thirds of the data recognition was non-native. The sample of data comprising one item (1847) contains considerably more foreign data than the sample with two items (912), leading to a higher WER. As expected, the best results are achieved by a fixed sequence of two items (2ItemsTA). However, recognition results with arbitrary sequence do not deviate strongly. The crucial result is that the WER with respect to *1||2ItemsFree* does not significantly deviate from *1Item* and *2ItemsFree*.

From a technical point of view the dialogue concepts presented in Chapter 5.2 may therefore be extended to optional combined category input. Again, the user is enabled to speak wording variants to reduce cognitive load. To select and disambiguate recognition results the result structure and confidences are taken into account. This keeps the multimodal dialogue strategy flexible and efficient (Mori, 2008). The ideal case is when the recogniser provides exactly one result per item or when the accumulated confidence of a resulting pair is very high, also with respect to other alternatives (see Figure 5.15). The corresponding result is then displayed while the user is simultaneously prompted the corresponding result with rising intonation in form of a question: “My future decided by Hillsong?”. The user may then confirm his input either explicitly or implicitly.

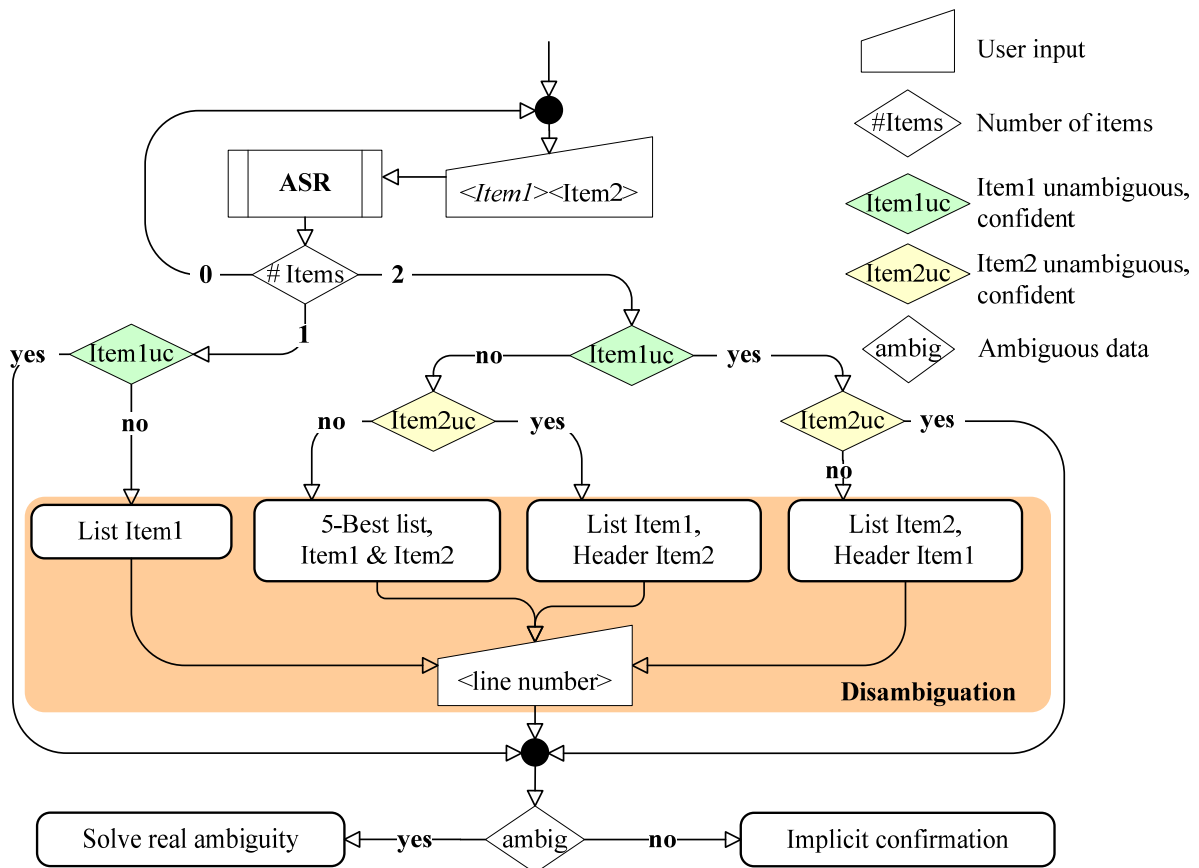


Figure 5.15: Dialogue concept for searching audio file names with combined category input

In other cases ambiguities are likely to occur. With combined category input the number of results is already reduced due to the relationship existing between two items. However, it is also possible that the recogniser gets several alternatives for a title or combinations of title and artist.

In this case it is necessary to disambiguate the user's input. To do this three procedures are differentiated: selecting from titles, artists and a list of 5-best combinations.

Selecting from alternatives:

In case the recogniser determines alternative result pairs, a list of the five best results (at maximum) is displayed.

In case only one Item1 is recognised (or one with a confidence that is above average) plus several suitable Items2, a list of entries comprising the second item will be displayed. The result of Item1 is presented in the header of the list. A corresponding system prompt could for example be: "Do you want Big in Japan by Alphaville, Guano Apes or Tom Waits?". The same procedure holds for the reverse case, i.e. if multiple results have been recognised for Item1 referring to exactly one Item2.

In case the user has only entered one item, a list of Item1 is displayed without header. Correspondingly the system prompt only includes the results of Item1.

In case multiple results were recognised for both Item1 and Item2, a 5-best list of the corresponding combinations is offered. List selection could be supported by prompting such as "What would you like to listen to: Big in Japan by Alphaville? – Sounds Like a Melody by Alphaville? – Big in Japan by Tom Waits?". The ending of the list is signaled by rising intonation of the last item's final word. In case the user does not respond to the question, the first combination will be selected due to its best confidence.

Real ambiguities:

If, independent of the result structure, a 'real' ambiguity occurs it may only be resolved by taking into account additional information (Berton, 2005). The system should find significant criteria that may support the user in decision making (Kruijff-Korbayová, 2005). For example if the title Tears in Heaven by Eric Clapton occurs twice in the database, the user should be able to make his decision according to the corresponding album. However, at this point of time, selection from the result list must be completed.

Compared to conventional dialogue design with sequential input of title and artist, the number of disambiguation steps of the presented concept is at maximum equal, but in general lower.

5.7 Application-independent approach

So far the approach presented in Chapter 5 focused on audio data. As already mentioned the problems encountered with regard to accessing audio data also occur in context with navigation (i.e. points of interest) and address book data, i.e.:

- Strongly rising number of entries due to increasing number of electronic on-board devices
- Long entries
- Lacking recollection of precise entry names

Points of interest (POI) and address book data may for example contain entries as follows:

- Prof. Dr. Hans-Dieter Müller
- Dr. med. dent. Schmidt Zahnarztpraxis
- Prof. Dr. Dipl.-Ing. Ludwig van Beethoven
- Freie Tankstelle Union Heinrich-von-Stephan-Straße
- Aral Autobahntankstelle Bavaria Weil am Rhein
- Hotel Adler – Gästehaus Herzog Ulrich – Landgasthof zum Lamm
- Schneider Donaueschingen Autohaus GmbH

To extend the approach to other speech-driven applications of advanced cars, the set of recursively applied generating rules was integrated into a hierarchy comprising seven steps for pre-processing data from audio, POI and address book: default normalisation, default rules, group-specific normalisation, group-specific rules, category-based rules, synonym generation and post-normalisation (see Figure 5.16).

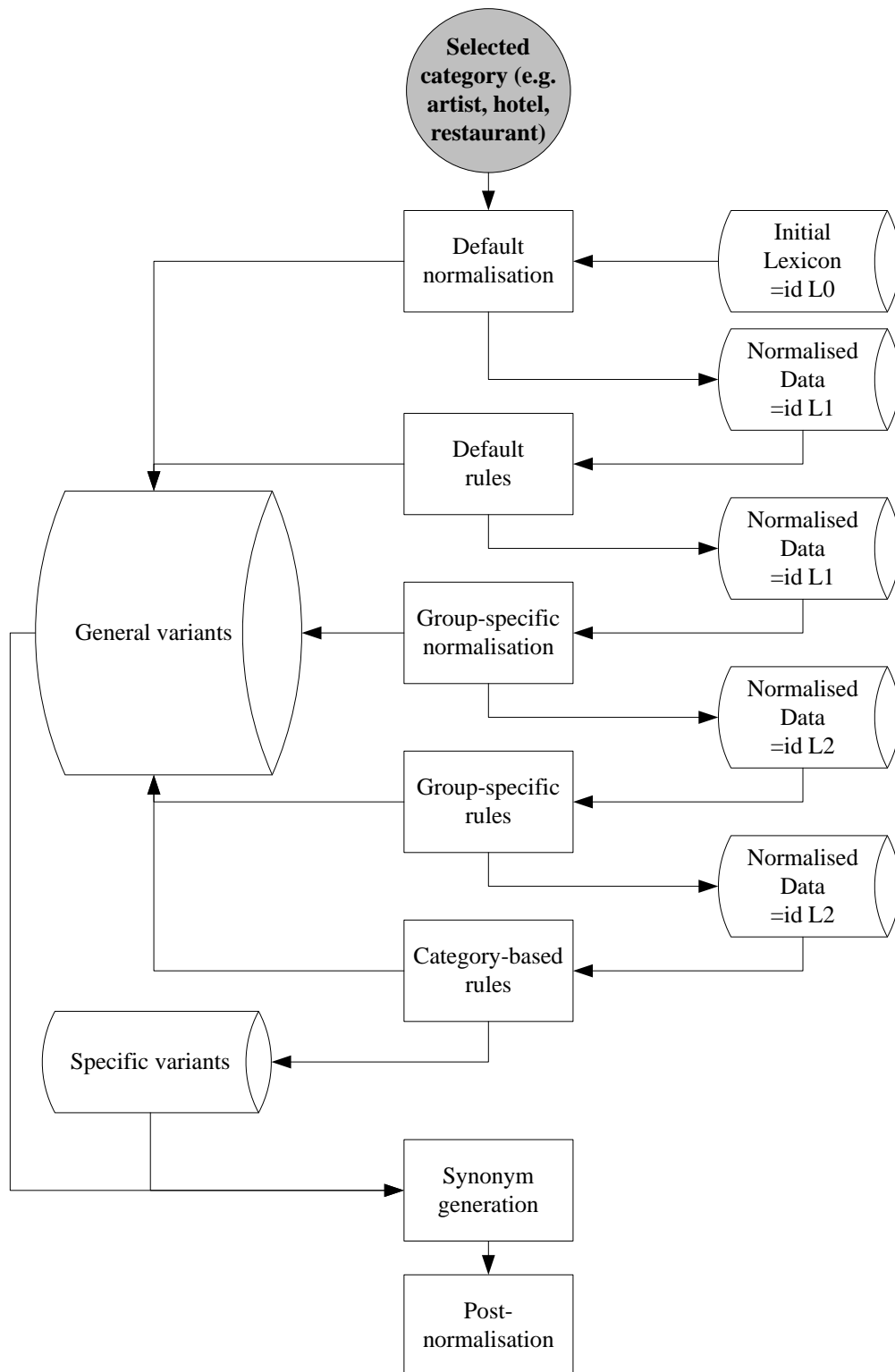


Figure 5.16: Pre-processing input data from applications audio, navigation and address book

1. Default normalisation

Default normalisation is application-independent, converting special characters into orthography and writing out abbreviations orthographically. Redundant blank spaces within an entry name are discarded.

Audio/ POI/ Address Book	
<i>Dr. Dre</i>	Doctor Dre
<i>Prof. Meier</i>	Professor Meier
Paul <i>u.</i> Lexy	Paul und Lexy
The Mamas & The Papas	The Mamas and The Papas
This Left Feels Right	This Left Feels Right

2. Default rules

The application of default rules is also application-independent. It processes data inside and outside brackets separately.

Audio/ POI/ Address Book	
Mr. Oizo feat. The Cure (live)	Mr. Oizo feat. The Cure live
	Mr. Oizo feat. The Cure
	live

3. Group-specific normalisation

Group-specific normalisation refers to each application separately. Here, application-specific abbreviations are written out orthographically and irrelevant information is discarded.

Audio	
Madonna <i>feat.</i> Britney Spears	Madonna featuring Britney Spears
<i>Jumpin'</i> Jack Flash	Jumping Jack Flash
<i>www.charthits.org</i> New Songs	New Songs
POI/ Address book	
<i>St.</i> Gallen	Sankt Gallen
Adams Apotheke im Lustnauer AG	Adams Apotheke im Lustnauer
Universität <i>fit e.V.</i>	Universität <i>fit</i>
Drebkauer <i>Str. 27</i>	Drebkauer Strasse 27
<i>Ing.</i> Meier	Ingenieur Meier

4. Group-specific rules

Group-specific rules also differ from application to application. As far as the application audio is concerned for example, initial articles and prepositions are made optional. Other so-called initial “prefixes” like personal pronouns or requests may also be omitted when selecting music. These initial articles, prepositions and other prefixes cover the languages German, English, Spanish and French. In addition symbols such as hyphen, inverted commas or colon are taken as separators for creating useful wording variants.

Audio	
<i>1989</i>	89
<i>90er Jahre</i>	90er
<i>The</i> Beatles	Beatles
<i>Let's</i> Live For Today	Live For Today
<i>Bach:</i> Air und Badinerie	Air und Badinerie

POI/ Address book

Apotheke am *ZOB*Apotheke am Zentralen
Omnibusbahnhof

Adalbert-Dengler-Weg 7

Adalbert-Dengler-Straße 7

5. Category-based rules

Category-based rules are both application- and category-dependent. For the application audio the number of categories can be seen at a glance (i.e. artist, album, title, genre and year). As already mentioned at the beginning of Chapter 5 the number of categories for points of interest may rise up to 80. For artist names category-based rules for example provide changing the sequence first name and last name. With regard to albums and titles redundant information like “Vol. 04” or “op. 257” is discarded. Within navigation data key words such as “hospital”, “restaurant”, “doctor”, “pharmacy”, etc. are detected to provide a basis for creating additional wording variants.

Audio

Artist

Eros Ramazzotti

Ramazzotti Eros

Ludwig van Beethoven

Beethoven

AlbumLooking For The Perfect Beat *1980-85*

Looking For The Perfect Beat

Chopin Etudes (*12*) For Piano

Chopin Etudes For Piano

TitlePerpetuum mobile *op.257*

Perpetuum mobile

07. - Todo Es Mentira

Todo Es Mentira

*Chopin Etudes (12) For Piano, op.10:
No.3 In E-Major, B74 “Tristesse”*

Tristesse

POI/ Address book

Petrol station

Avia Thannheimer Straße

Avia

Hospital

Helios Klinik Titisee-Neustadt

Helios Klinik

Klinik Titisee-Neustadt

Category-based rules produce two kinds of variants: general and specific variants. General variants are active on a higher level within the menu hierarchy than specific variants. For example, general variants are speakable once a particular application has been activated, e.g. by saying “search Wochner’s hotel” as in the example below. Wording variants merely consisting of a category name or a corresponding synonym (e.g. “hotel” as wording variant for “Wochner’s Hotel Sternen”) are not allowed as it would permanently conflict with selecting a category, in this case “hotel”.

Example: Wochner’s Hotel Sternen

General variants

- Wochner’s Hotel Sternen
 - Wochner’s Hotel
 - Hotel Sternen
 - Hotel Wochner’s
 - Sternen Hotel
 - ~~Hotel~~
-

Specific variants

- Wochner’s
- Sternen
- Wochner’s Sternen

Specific variants are those variants that do not immanently reflect the category they belong to. The variants “Wochner’s”, “Sternen” or “Wochner’s Sternen” for example could refer to a

shopping centre, restaurant etc. Therefore this kind of variant is speakable not until its superordinate category has been selected.

6. Synonym generation

At this level entry names and their corresponding variants are searched for category keywords. This process is application-dependent. The identified keywords are replaced by synonyms that are then stored as additional variants. In this context it is important to carefully balance the number of synonyms. Otherwise, depending on the category, the number of wording variants may rise sharply, e.g. with regard to hotels and restaurants.

Audio	
<i>London Symphony Orchestra</i>	London Sinfonie Orchester
<i>Beethoven: Piano Sonatas</i>	Beethoven Klavier Sonaten
POI/ Address Book	
<i>Professor Dr. Grosse</i>	Professor Doktor Arzt Grosse
<i>Avia Tankstelle</i>	Avia Autohof Tankautomat Tankhof Tankstop
<i>Pizzeria Giovanni</i>	Gasthof Gasthaus Gaststätte Restaurant Pizza-Service Ristorante Giovanni
<i>Parkhaus Wilhelma</i>	Tiefgarage Garage Parkgarage Parkdeck Parkhotel Wilhelma
<i>Waldhotel Sonnenblick</i>	Hotel Ferienhaus Ferienhof Pension Landhaus Gästehaus Kurhotel Privatpension Sporthotel Sonnenblick

7. Post-normalisation

At the end of the generating process of wording variants a final normalisation takes place application-independently. Generating wording variants will most likely produce artefacts, e.g.

articles or prepositions may be left over as wording variant or in a wrong context as in case of “Betriebshoftankstelle Beim”. These artefacts are either deleted or adjusted correspondingly. In addition, wording variants merely consisting of category names or synonyms are deleted, as well as variants that do not exceed two letters.

Audio/ POI/ Address Book	
<i>Am / im / zum / der / die / das / den / beim / zur / und</i>	---
Meier.	Meier
Betriebshoftankstelle Beim	Betriebshoftankstelle

Concerning the coverage of spoken utterances the hierarchical approach is identical with the initial approach presented in Chapter 5.2. However, the hierarchical approach generates less overhead of wording variants because the rules are more deterministic. Due to the hierarchical procedure also less redundancy occurs: as for the initial approach it could easily happen that different sequences of rules end up with one and the same wording variant. Finally, this approach includes synonyms for categories as well as specific variants that are only speakable at a stage where interaction within dialogue, i.e. menu hierarchy, has advanced.

In a user study the approach of directly accessing pre-processed points of interest by speech was compared with the state-of-the-art of a Mercedes-Benz series (Mann, 2008c). 30 subjects aged 18 to 65 tested both systems, the Mercedes-Benz and the prototype system. The subjects were split in two groups of 15 members according to which group one started with the Mercedes-Benz series followed by the prototype and group two did so vice versa. The tasks they were given comprised various scenarios on points of interest, e.g. navigating to a cinema, restaurant, petrol station, shopping centre, zoo, parking place etc. Care was taken not to explicitly mention the corresponding category name. The findings showed that both kinds of approaches were equally desired. Consequently a user-friendly concept for accessing large databases should comprise the following features:

1. Display of all available categories and a selection according to categories, respectively.
2. Selection according to line number.
3. List entries, i.e. categories including their corresponding content, should be speakable. Additional pre-processing of large databases to create wording variants would be useful. The number of interaction steps could thus be considerably reduced.
4. A search mode allowing for directly speaking categories as well as the categories' content (e.g. points of interest, music titles, address book entries, etc.).

Chapter 6

Conclusions

The final chapter summarizes main aims and results of this approach. It compares human-human communication with human-computer interaction and draws conclusions on what aspects from human communication may as well be applied to human-computer interaction and which should not as they would run counter to user-friendly interaction. The chapter concludes with an outlook on future work.

6.1 Summary

In the automotive environment an increasing number of assisting systems as well as luxury features has found its way into vehicles over the past years: radio, CD, MP3, DVD, on-board address book, organisers, mobile phones, PDAs, destination input, selecting points of interest etc. The variety of on-board electronic devices and its extensive functions are likely to distract drivers from pursuing the primary driving task. Speech as a means of human-machine interaction should offer the user hands-free operation while driving to concentrate on the traffic situation rather than using various displays and buttons. However, parts of in-car dialogue systems tend to have lost transparency due to the increased functionality. Although people have accepted the innovative character of speech as interaction means and are generally motivated to use it, they fairly often get frustrated when operating an in-car SDS. Interaction is considered difficult and cumbersome and the first hurdle for getting acquainted with the system is high.

It is often the conceptual design as opposed to technical mistakes that is regarded as the major obstacle for efficient and user-friendly interaction. The aim of this approach therefore was to develop concepts that facilitate human-machine interaction. The concepts range from general

design guidelines to accessing large databases within applications such as audio, navigation and address book.

General dialogue design on

- Specifying SDS
- Prompt design
- Vocabulary selection
- System support
- Recovering from errors
- Dialogue control
- Initiating speech dialogue
- Short and long cuts
- Combining spoken and manual interaction

is based on evidence taken from analysing communication among humans on the one hand and results from extensive user testing on the other. User testing of speech applications currently available in advanced cars is a crucial method for detecting typical problems users encounter when employing state-of-the-art speech technology. The subjects taken for the testing covered two types of users: the novice user, being hardly experienced with in-car speech applications and the expert user, being familiar with current in-car speech technology. Analysis of communication among humans as well as the Wizard-of-Oz experiment (Hüning et al., 2003) aimed at finding out how people interact naturally and without being limited by technical restrictions.

Concerning the findings the following components for spoken dialogue systems are considered prerequisite to enable a user-friendly interaction between human and machine:

A component for *natural language understanding* and a *statistical language model* (as opposed to parser and grammar) that allow for processing less restricted input. The collected user speech data from the Wizard-of-Oz experiment show that the user's wording is far from natural speech when interacting with a machine, but still more extended than what command and control systems may provide (Chapter 4, 4.4.3). One third of all user utterances comprise more than three words. During the experiment the users generally had a positive attitude towards a dialogue

system offering natural language input. As far as mixed initiative is concerned the users applied it in 43% of the cases (Chapter 4, 4.4.12), showing a clear tendency towards getting away from directed dialogue. Mixed initiative implies that two or more phrases may be combined within one utterance. Accordingly, they need to be included in the statistical language model. The statistical language model also needs to provide rules according to which short and long cuts are speakable (Chapter 4, 4.4.11). Users, in particular experienced users, are thus able to accelerate the task completion rate by reducing the number of interaction steps.

As far as the lexicon is concerned it is necessary to extend the vocabulary by synonyms. Evidence taken from experimental data shows that users like using their own wording. Despite given keywords and wording from the graphical user interface, tasks and system prompts, 38% of the user input still contain synonyms (Chapter 4, 4.4.3). Improving the vocabulary could thus avoid unnecessary errors and reduce the number of interaction steps. Care needs to be taken though that the number of synonyms remains compatible with recognition performance.

Taking a look at spontaneous speech and the collected user speech data the natural language unit also needs to be able to handle the following types of disfluencies (Cole, 1996, Chapter 6.4; Kronenberg, 2001, p.12; Chapter 4, 4.4.3) occurring on either the user or the system side: ungrammatical sentences, interruptions, substitutions, deletions and hesitations.

A *contextual interpretation (CI)* is necessary to interpret the NLU result within dialogue context. Being in the application navigation for example, the user utterance ‘Bochum’ may then be interpreted as city Bochum. Correspondingly, when in the application audio, ‘Bochum’ may be interpreted as song or album by Herbert Grönemeyer.

Particularly with regard to mixed initiative and less restricted user input, contextual knowledge on the dialogue flow is crucial when it comes to distinguishing utterances according to their dialogue acts (cf. Chapter 3, 3.2.2), i.e. whether their intention is to repeat, confirm, reject or correct. The next example illustrates the difficulties a system needs to solve.

Example 21:

User: I would like to call John Q. Public. His number is 0 4 3 2...5 6 7 8 4 4 9.
System: You want to call 0 4 3 2?
User: Yes 0 4 3 2 and then 5 6 7 8 4 4 9.
System: 0 4 3 2 5 6 7 8 4 4 5?
User: 4 4 9.
System: 0 4 3 2 5 6 7 8 4 4 9?
User: That's correct, dial please.

Typical dialogue acts required for speech dialogue systems can be restricted to three of Searle's illocutionary forces (Searle, 1979, p.20; also see Levinson, 1983, p.240; Searle, 1969, p.66): assertives (e.g. claim, confirm, correct), directives (e.g. question, command, recommend, warn and repeat) and commissives (e.g. offer, book).

To store dialogue flow an additional component called *dialogue history* is required. It provides an organised and consistent way of storing system data, analysis results and hypotheses of various components. This knowledge will be available to all components. In addition, dialogue history is also a basis for

- Sub dialogues that may be resumed after having been interrupted by road messages, database queries or requests for pin numbers as in example (22).

Example 22:

User: I want to store a phone number.
System: Please speak the pin code first.
User: 5 3 7 9.
System: That's correct. Which phone number would you like to store?

- Reference resolution of deictic and non-deictic expressions (cf. Chapter 3, 3.2.4), e.g. as in:

I am hungry <i>now</i> .	(<i>time deixis: current position</i>)
Can I visit a zoo <i>there</i> ?	(<i>place deixis: final destination</i>)
I am looking for a shopping centre <i>nearby</i> .	(<i>place deixis: current position</i>)
What was the last message again? Please read <i>it</i> out.	(<i>anaphor: pronoun it</i>)

- Interpreting elliptical constructions (cf. Chapter 4, 4.4.3) occurring in context with list selection, coordinating constructions or repetition, as the following examples show.

Example 23:

System: Did you mean Neustadt Aisch, Neustadt an der Weinstraße or Neustadt im Taunus?

User: The first one. (List selection)

System: Neustadt Aisch.

Example 24:

System: Which album? Have a Nice Day, Keep the Faith, Cross Road – The Best Of, Bounce?

User: What was the second one called again? (Repetition)

System: Keep the Faith.

User: Add album to playlist and then play. (Coordinating construction)

- Detecting interaction with repetitive user rejection of one and the same recognition result, for example after having entered a particular city name. Instead of ending up in the manual speller the user may be offered different alternatives. As far as lexical disambiguation (as in the Neustadt case) is concerned the user may be prevented from having to go through the same procedure time and again. Analysis results from previous user input may be retrieved and offered to the user with high priority. This improves dialogue strategy and makes dialogues more natural to the user.

Just as in human communication it needs to be ensured that context is coherent when interacting with multimodal systems. This coherence refers to utterances and their constituents (e.g. nouns, verbs, noun phrases, verb phrases, prepositional phrases) within one modality in that they have to relate to each other on a grammatical and semantic level but also across modalities (Bußmann, 1990, p.389; Harris, 1952, Chapter 6; Bellert, 1970; Chapter 3, 3.2). Care must be taken that all modalities are synchronised with each other. The user may thus change from manual to spoken interaction and vice versa without being irritated because the two modalities reflect different system states.

When two or more participants are engaged in a human dialogue the term alternation does not imply that the boundaries between turns are always clear cut. The same holds for dialogues

between human and machine. Turns are taken when starting to speak, when pauses occur, slightly before a party has finished and also by interrupting dialogue participants in the middle of an utterance (Chapter 4, 4.4.8). As the experiments' findings indicate there is a clear tendency towards less restricted input and a dialogue flow that is more natural. Accordingly it is necessary to get away from sequential recognition that restricts the user to wait until a system prompt has ended and the system is ready for accepting speech input. Barge-in recognition enables to cover important kinds of initiatives a participant may take when he is about to take the turn. Apart from that barge-in accelerates the task completion rate by avoiding clarifying dialogues as the system has missed important information the user has given.

Providing feedback is a process that constantly accompanies dialogue among humans. It establishes mutual understanding (grounding) on both the speaker and the hearer side of the dialogue. The hearer either gives feedback while listening to the speaker (e.g. by saying yes or nodding the head) or as soon as he turned from hearer to speaker position (McTear, 2004, p.54; Chapter 3, 3.1). Turning to interaction with machines, experimental findings show that feedback from the hearer (either user or system) by means of intermediate comments such as yes, correct, right, etc. is considered inadequate and disturbing unless the comments implied selecting or confirming a particular item on the user side. When in the speaking position, however, the system needs to establish a grounding process that is two-fold:

1. The system needs to give feedback concerning the current system state whenever uncertainties on the user side are likely to occur. As speech is the crucial modality parallel to the driving task it is not sufficient to visually reflect the corresponding system state. The user simultaneously needs to be informed by speech. This might for example be necessary when the user jumps from one application to another. As recognition errors are likely to occur a system prompt should confirm the user input by initially including the corresponding application, e.g. by saying "navigation – which function?". Similar feedback should also be given on the task level in case similar tasks occur across various applications (e.g. as in the process of storing city names, address book entries and radio stations). By simply adding a keyword, unspecific prompts such as "please speak the name" may be easily disambiguated prior to causing confusion on the user side.

-
2. When it comes to selecting list elements by text enrolment or line number the system needs to give explicit feedback of what has been understood and consequently selected. Again, this confirmation has to take place visually and by speech. Take a user selecting a city name, an album or an address book entry according to line number (saying “number 2 please”) or by directly speaking the entry name (either completely or partially). A system prompt such as “city name | album | address book entry has been selected” would be insufficient as it lacks explicit confirmation. The user cannot be sure whether his item and the actually selected item are identical unless he takes a look at the display. This uncertainty may be avoided by simply integrating the corresponding line number and/or text enrolment into the system prompt.

Extending unspecific prompting by including keywords, line numbers or text enrolments makes the content transparent to the user. He may then interpret the messages easily. Analysing the grounding process in context of Grice’s conversational principles (Grice, 1989, p.26; Chapter 3, 3.2.1) shows that it follows the maxims of quantity and manner. Sentences such as “Please speak the city name” or “Number 2 – Zoo Wilhelma. Would you like to start navigation?” provide just the right amount of information the user requires, they are clear and without ambiguity. Requests such as “Please speak the name” or “Which function?” on the other hand would be too little information as it may not be clarified which application the respective utterance refers to, unless the user looks at what is visually displayed.

Basically, all guidelines by Grice, i.e. the maxims of quantity, quality, relation and manner and the underlying cooperative principle are guidelines that need to be omnipresent throughout the design of speech dialogue systems. On the system side however the maxim of quality may not be followed consequently of course. Due to the special in-car situation and speech recognition technology rejection errors, substitution errors and insertion errors (Schmandt, 1994, p.160; Chapter 4, 4.2) may occur any time, presenting the user with false results or wrong system behaviour. It is advisable to reprompt the user for confirmation if confidence is low or if the recognition result is unlikely when taking into account dialogue context, but this will not solve the actual problem.

As far as multimodal specification is concerned an important step to follow the maxim of quality is to provide dialogue components that are broadly reusable and to use a representation

formalism that unifies spoken and manual interaction. Both aspects increase consistency within and across modalities, making the actual system less prone to errors.

Some maxims in turn need to be violated deliberately when interfering with maxims that are given higher priority in certain context. Prompt design in general should be as informative and as short as possible, following the maxim of quantity in the first place. As speech is temporal and presented to the user sequentially (Schmandt, 1994, p.102; Balentine, 2001, p.11; Gibbon, 1997, p.82) too much information is difficult to memorize, in particular when primarily pursuing the driving task. Accordingly, being presented a long list of menu items, it is likely that by the end of a prompt the user has forgotten what was initially said. Unless a system provides barge-in, the number of menu items should therefore be no more than three.

In case information or instructions contained in a system prompt are misleading to the user, the maxim of quantity should be neglected for the sake of the maxim of manner. Prompt length could be extended by adding an example to clarify the prompt message. During the experiments the request “please spell the street” often turned out to be misunderstood. Rather than spelling the street name letter by letter the user entered the corresponding name in one go. An extended prompt such as “please spell the street – for Stuttgarter Straße for example say S-T-U-T-T” could make the user susceptible for the actual prompt message.

When it comes to jumping from one topic to another, communication among humans shows that this kind of behaviour is quite common. Obviously there is a violation of the maxim of relation to the extent that people do not consequently stick to a topic. Similar behaviour could be observed during the Wizard-of-Oz experiment in context with experienced users (cf. Chapter 4, 4.4.11). They applied two types of task changes: short cuts, implying a task change within an application and long cuts, implying a task change from one application to another. Evidence from both human communication and man-machine interaction clearly indicates that the maxim of relation should not be maintained in this context. Moreover, it would run counter to reducing the number of interaction steps and thus the time required to accomplish a task.

Turning to large databases in cars, accessing audio, navigation and address book data has turned into a cumbersome task. For each application nowadays a variety of electronic on-board devices is available. Audio applications for example may consist of a variety of

- Storage devices: e.g. optical disk, memory card, hard disk, flash memory, USB (MP3 player, iPod, memory stick, hard disk)
- Data formats: audio and raw
- File types: e.g. *.mp3, *.mpg, *.ogg, *.wav

In order to successfully select particular audio data, the user must have a general technical understanding of the system as well as be able to remember which medium provides what contents. Such an application is not transparent to the user. It violates the maxim of manner in that the variety of audio file names on a variety of technical devices is neither perspicuous nor orderly, demanding too many cognitive resources while pursuing the driving task. Besides current methods of navigating the growing amount of audio data by means of speech commands (like for example ‘next medium’, ‘previous medium’, ‘next song’, ‘previous song’), by selecting the corresponding line number or by manually searching storage devices are no longer sufficient. The approach on accessing large databases using in-car speech dialogue systems in Chapter 5 therefore aimed at offering an audio application management that neither requires previous knowledge on electronic devices nor the corresponding audio data they contain.

The approach is based on intuitive interaction concepts for searching audio data (cf. Chapter 5, 5.2):

- Category-based search, based on pre-selecting a category
- Category-free search, based on directly entering audio file names or parts thereof
- Physical search as fall-back solution

Category-based and category-free search consider the content of all audio storage devices available in the car. In addition to the currently available means of navigating audio data, both search modes allow for verbally selecting audio file names by means of speakable text entries (text enrolments). To prevent the user from having to manually scroll through long lists as he may not recollect the precise name of albums, titles, etc., the method is extended by generating

rules. Generating rules pre-process items of audio data in context with special characters, abbreviations, keywords, closed word classes and secondary components, thus creating additional wording variants to the 'original names' available in a music database. Evidence for the necessity of generating rules is taken from studies on selecting music and personal address book data (Mann, 2008a; Mann 2007b): when accessing large databases users often tend to violate several maxims at once as their input of file names is likely to be incomplete. This means they make contributions that are only partially true (maxim of quality) and do not provide sufficient information (maxim of quantity). The less precise their input gets, the more likely it is that ambiguities may occur (maxim of manner). Providing additional wording variants by generating rules circumvents these violations and, pursuing Grice's superordinate cooperative principle, enables the user to successfully access data that would otherwise remain undetectable by speech.

To verify the efficiency of generating rules, speech data of audio file names were collected across common categories (artist, album, title, genre, year and audio books). The survey comprised various scenarios ranging from free input to recollection of given audio file names (Chapter 5, 5.5). To find out how subjects select music if no restrictions are given, task 1 began with free input. In task 2 individually favoured audio file names had to be entered according to given categories. The task aimed at reflecting the subjects' knowledge of music file names. The requested input required both single category input as well as combinations of two categories. Task 3 asked the subjects to reproduce given pairs of audio file metadata. Intended cognitive overload led the subjects to filter crucial components as soon as each pair was fading out on the display.

Analysing the data shows that the results are promising: for the category artist the rule set increased the number of matches from initially 61% to 85%. Concerning title names the match rate went up from 51% to 69%, whereas coverage of album names with generating rules could only be increased by 14% points, from 60% to 74%. The fact that the overall coverage of album and title names is low compared to artist names can be related to mispronunciations often produced when being unfamiliar with foreign languages, in particular French, Italian and Spanish.

As far as the increase of the vocabulary is concerned, the rule set mainly applies to artist and title names. While their vocabulary more than doubles, additionally generated album names are 81%.

Results from recognition tests show that the number of voice input parameters may be extended from single category input to optional category input of one or two items with arbitrary sequence. This option goes hand in hand with the results from task 1 in which the subjects' unrestricted input does not exceed more than two items per utterance.

To conclude, this study developed new concepts that simplify the interaction with in-car speech applications providing large databases. Due to nowadays' variety of electronic devices these databases are either available in cars or can be easily integrated. Current methods of navigating data however once were focused on very small numbers of music titles and address book entries and are therefore no longer adequate. To establish general requirements for the user interface, human communication was analysed to extract principles crucial for human-machine interaction. Also, various user studies were carried out to find out how users behave in human-computer interaction and where usability problems occur. The user's lacking recollection of audio, navigation and address book data was particularly taken into account and integrated into the concepts. The concepts bring back transparency into the multitude of technical devices and large amounts of data, while simultaneously ensuring consistency within and across modalities.

6.2 Future work

The presented concepts in this approach are a first step towards human-computer interaction that enables less restricted input. In future research, various investigations with regard to generating rules are necessary. To verify the efficiency of the developed rule set more extensively, additional datasets of personal music collections are required. They could give important information about how the vocabulary of original names available in a database increases when generating rules are applied and also, whether these sizes are acceptable. Additionally, a large number of speech data correlating with the available music collections needs to be recorded. Their analysis shows how well the set of generating rules covers what users input by speech when searching for audio data and where modifications are necessary.

Furthermore, a usability study with a prototype system comprising the hierarchical approach on generating rules for POIs, audio and address book data would be important, also in context with recognition technology that is particularly focussed on dealing with large lists of embedded systems (Schreiner, to appear).

Turning from wording variants to phonetic variants, it should be considered to integrate text-based language identification for metadata of songs, albums and artist names not covered in the database. That avoids a large number of useless pronunciation variants, which is particularly useful for the European market with many different languages.

In case the user has difficulties in adequately pronouncing foreign language music titles in French, Italian or Spanish, an additional category such as ‘different country’ might be helpful. While it may be true that “Music knows no borders“ (EC, 2009), a country (or rather language region) is often associated with certain styles or sounds. The user should be able to simply select it by saying ‘French Music’ or even ‘French Chansons’ without having to make pronunciation efforts that are most likely to be in vain.

For the “passive” music listener having extreme difficulties in correlating audio file names with respective music it might be useful to offer an additional search according to mood (Musicoverly, 2009).

Zusammenfassung

Bedienkonzepte für Sprachdialogsysteme im Fahrzeug und deren Integration in ein multimodales Mensch-Maschine-Interface

Das Thema dieser Arbeit ist die Entwicklung von Bedienkonzepten für multimodale Dialogsysteme im Fahrzeug. Grundlage dafür bilden zum einen die Untersuchung zwischenmenschlicher Kommunikation über gesprochene Sprache, zum anderen die Durchführung von Benutzerstudien über die sprachliche Interaktion mit bestehenden Fahrzeuganwendungen sowie im Rahmen eines Wizard-of-Oz.

Motivation

Der Begriff Vielfalt prägt heutzutage jeden Bereich unseres täglichen Lebens. Was die Innenausstattung von Fahrzeugen betrifft, so hat dort – dieser Entwicklung folgend – in den vergangenen Jahren eine Vielzahl von Luxusfeatures Einzug gehalten. Neben dem Radio verfügt ein Fahrzeug beispielsweise über CD, MP3, DVD, fahrzeugeigene Adressbücher, Organizer, Mobiltelefone und PDAs, Zieleingabe sowie eine Auswahl von Sonderzielen. Die Vielfalt von elektronischen Geräten im Fahrzeug und ihre umfangreichen Funktionen können einen Fahrer leicht von der eigentlichen primären Hauptaufgabe des Fahrens ablenken. Sprache als Mittel einer Mensch-Maschine-Interaktion sollte dem Benutzer während der Fahrt eine sogenannte Handsfree-Handhabung ermöglichen, um sich auf das Verkehrsgeschehen konzentrieren zu können, ohne währenddessen noch verschiedenste Funktionstasten betätigen zu müssen. Nichtsdestotrotz haben Teile von Dialogsystemen aufgrund der vermehrten Funktionalität an Transparenz verloren. Obwohl viele Menschen den innovativen Charakter von Sprache als Mittel der Interaktion akzeptiert haben und im Allgemeinen auch motiviert sind, davon Gebrauch zu machen, werden sie oftmals bei der Bedienung eines Sprachdialogsystems frustriert. Die

Interaktion wird als schwierig und mühsam betrachtet. Die erste Hürde, um mit einem System vertraut zu werden, ist dadurch sehr hoch. Oftmals wird sogar das Design im Gegensatz zu technischen Problemen als Hauptursache für eine mangelnde benutzerfreundliche Interaktion gesehen. Das Ziel der Arbeit war es, Konzepte zu entwickeln, die eine Interaktion zwischen Mensch und Maschine erleichtern. Die Konzepte erstrecken sich von allgemeinen Designrichtlinien hin zum konkreten Zugriff auf große Datenbanken innerhalb der Applikationen Audio, Navigation und Adressbuch.

Von Mensch-Mensch-Kommunikation zu Mensch-Maschine-Interaktion

Um die Herausforderungen verstehen zu können, die an das Dialogdesign von Sprachdialogsystemen gestellt werden, wurde in dieser Arbeit zunächst der zwischenmenschliche Dialog untersucht (Kapitel 3):

- Was ist ein Dialog?
- Welches Wissen ist notwendig für eine erfolgreiche Kooperation?
- Welches sind die zugrunde liegenden Prinzipien der Kommunikation?

Darüber hinaus wurden vielfältige Benutzerstudien im Hinblick auf aktuell verfügbare Sprachanwendungen im Fahrzeug durchgeführt (Kapitel 4, 4.3). Diese Studien sind entscheidend für die Herausarbeitung von Problemen, denen Benutzer typischerweise im Umgang mit gegenwärtigen Sprachdialogsystemen begegnen. Die Probanden deckten dabei beide Arten von Benutzergruppen ab: einerseits den Novizen, der kaum Erfahrung mit Sprachanwendungen im Fahrzeug hat, und andererseits den Experten, der im Umgang mit der aktuellen Sprachtechnologie vertraut ist.

Die Ergebnisse zeigen, dass die Komponenten NLU und statistisches Sprachmodell, kontextuelle Interpretation und Dialoghistorie bei Sprachdialogsystemen Voraussetzung für eine benutzerfreundliche Interaktion zwischen Mensch und Maschine sind.

Eine Unit für die Verarbeitung natürlicher Sprache (NLU) und ein statistisches Sprachmodell werden benötigt, um dem Benutzer eine freiere Spracheingabe zu ermöglichen. Die gesammelten Sprachdaten des Wizard-of-Oz-Experiments (Hüning et al., 2003) belegen, dass die

Ausdrucksweise des Benutzers bei der Interaktion mit einem Computer von natürlicher Sprache weit entfernt ist, zugleich jedoch auch deutlich umfassender als es eine Eingabe bei Command-and-Control-Systemen erlaubt (Kapitel 4, 4.4.3). Ein Drittel aller Benutzeräußerungen umfasst jeweils mehr als drei Worte. Während des Experiments hatten die Probanden zudem generell eine positive Einstellung gegenüber einem Dialogsystem mit freier Spracheingabe. Was gemischte Initiative betrifft, so wurde sie von den Probanden in 43 Prozent der Fälle angewandt (Kapitel 4, 4.4.12). Dies bedeutet, dass zwei oder mehrere Phrasen innerhalb einer Äußerung kombiniert wurden. Daraus ergibt sich eine deutliche Tendenz weg vom geführten Dialog. Entsprechend müssten diese Phrasen in ein statistisches Sprachmodell integriert werden. Das statistische Sprachmodell muss ebenfalls über Regeln verfügen, die definieren, welche short und long cuts jederzeit gesprochen werden dürfen. Der Benutzer, insbesondere der Experte, ist dadurch in der Lage, die Bearbeitungsdauer einer Task zu verkürzen, indem die Anzahl der Interaktionsschritte reduziert wird (Kapitel 4, 4.4.11).

Was das Lexikon betrifft, so ist es notwendig, bestehendes Vokabular um Synonyme zu erweitern. Äußerungen aus den Studien belegen, dass Benutzer gerne von ihrer eigenen Wortwahl Gebrauch machen. Trotz der indirekten Vorgabe von Schlüsselwörtern und Formulierungen durch graphisches Benutzerinterface und Systemprompts beinhalten Benutzeräußerungen zu 38 Prozent Synonyme (Kapitel 4, 4.4.3). Durch eine Verbesserung des Vokabulars könnten unnötige Interaktionsfehler vermieden und damit die Anzahl von Interaktionsschritten verringert werden. Bei der Anzahl der Synonyme gilt es darauf zu achten, dass diese mit der Erkennungsleistung kompatibel bleibt.

Im Hinblick auf gesprochene Sprache zeigt die Sammlung der Benutzersprachdaten, dass eine NLU in der Lage sein sollte, mit verschiedenen Arten von Disfluenzen umgehen zu können. Dazu gehören das Auftreten von ungrammatischen Sätzen, Unterbrechungen, Substitutionen, Auslassungen und Zögern (Cole, 1996, Kapitel 6.4; Kronenberg, 2001, S.12; Kapitel 4, 4.4.3).

Eine kontextuelle Interpretation (CI) interpretiert die Ergebnisse der NLU im Dialogkontext. Innerhalb der Applikation Navigation kann dadurch beispielsweise die Benutzeräußerung 'Bochum' als Stadt Bochum interpretiert werden. Entsprechend wird 'Bochum' als Titel oder Album von Herbert Grönemeyer interpretiert, sobald sich der Benutzer in der Audio-Applikation befindet.

Besonders bei gemischter Initiative und freier Benutzereingabe ist kontextuelles Wissen über den Dialogablauf äußerst wichtig, wenn es darum geht, Äußerungen nach ihrem Dialogakt (vgl. Kapitel 3, 3.2.2) zu unterscheiden, das heißt herauszufinden, ob es sich dabei um eine Wiederholung, Bestätigung, Zurückweisung oder Korrektur handelt. Beispiel (1) verdeutlicht die Schwierigkeiten, die ein System bewältigen muss.

Beispiel 1:

Benutzer: Ich möchte Hans Mustermann anrufen. Seine Nummer lautet 0432 567 8449.
System: Sie möchten folgende Nummer wählen: 0432?
Benutzer: Ja, 0432 und dann 567 8449.
System: 04325678445?
Benutzer: 449.
System: 04325678449?
Benutzer: Ja, bitte wählen Sie die Nummer.

Typische Dialogakte, die bei Sprachdialogsystemen benötigt werden, können auf drei von Searles illokutionären Akten beschränkt werden (Searle, 1979, S.20; siehe auch Levinson, 1983, S.240; Searle, 1969, S.66): assertive (z.B. behaupten, bestätigen, korrigieren), direktive (z.B. fragen, befehlen, vorschlagen, warnen, wiederholen) und kommissive (z.B. anbieten, reservieren).

Um den Dialogverlauf speichern zu können, wird eine weitere Komponente, die Dialoghistorie, benötigt. Sie ermöglicht das zentrale und einheitliche Speichern von Systemdaten, Analyseergebnissen und Hypothesen verschiedener Komponenten. Dieses Wissen ist allen Komponenten zugänglich. Die Dialoghistorie ist darüber hinaus ebenfalls eine Grundlage für

- Die Wiederaufnahme von Subdialogen, die durch Verkehrsnachrichten, Datenbankabfragen oder Pinabfragen unterbrochen wurden.

Beispiel 2:

Benutzer: Ich möchte eine Telefonnummer speichern.
System: Bitte nennen Sie zunächst den Pincode.
Benutzer: 5379.
System: Pincode korrekt. Wie lautet die Telefonnummer?

- Referenzauflösung von deiktischen und nicht-deiktischen Ausdrücken (siehe Kapitel 3, 3.2.4), wie z.B.:

Ich habe *jetzt* Hunger.

Zeitdeixis: aktuelle Position

Gibt es *dort* einen Zoo zu besichtigen?

Lokale Deixis: endgültiger Bestimmungsort

Ich suche ein Einkaufszentrum *in der Nähe*.

Lokale Deixis: aktuelle Position

Wie lautet die letzte Nachricht? Bitte lesen Sie *sie* vor.

Anapher: Pronomen sie

- Die Interpretation elliptischer Satzkonstruktionen (siehe Kapitel 4, 4.4.3), die im Zusammenhang mit Listenauswahl, koordinierenden Konstruktionen oder Wiederholungen auftreten.

Beispiel 3:

System: Meinten Sie Neustadt Aisch, Neustadt an der Weinstraße oder Neustadt im Taunus?

Benutzer: Das erste bitte.

(Listenauswahl)

System: Neustadt Aisch.

Beispiel 4:

System: Welches Album? Have a Nice Day, Keep The Faith, Cross Road – The Best Of, Bounce?

Benutzer: Wie heißt das zweite Album gleich noch mal?

(Wiederholung)

System: Keep The Faith.

Benutzer: Bitte zur Playlist hinzufügen und anschließend abspielen.

(Koordination)

- Die Detektion wiederkehrender Benutzerrückweisungen von ein und demselben Erkennungsergebnis, beispielsweise nach Eingabe eines Städtenamens. Auf diese Weise können dem Benutzer sukzessive unterschiedliche Alternativen angeboten werden. Was die lexikalische Disambiguierung (wie im Neustadt-Beispiel) betrifft, so könnte damit vermieden werden, dass der Benutzer wieder und wieder dieselbe Disambiguierungsprozedur durchlaufen muss. Erkennungsergebnisse aus vorigen Benutzereingaben könnten abgerufen und dem Benutzer mit erhöhter Priorität angeboten werden. Dadurch wird die Dialogstrategie verbessert und die Dialoge zwischen Mensch und Maschine vermitteln einen natürlicheren Eindruck.

Genau wie bei zwischenmenschlicher Kommunikation muss bei der Interaktion mit multimodalen Systemen gewährleistet sein, dass der Kontext kohärent ist. Diese Kohärenz bezieht sich auf Äußerungen und deren Konstituenten (z.B. Nomen, Verben, Nominalphrasen, Verbalphrasen, Präpositionalphrasen) innerhalb einer Modalität dahingehend, dass sie auf einer grammatikalischen und semantischen Ebene Bezug aufeinander nehmen, als auch über mehrere Modalitäten hinweg (Bußmann, 1990, S.389; Harris, 1952, Kapitel 6; Bellert, 1970; Kapitel 3, 3.2). Es muss sorgfältigst darauf geachtet werden, dass alle Modalitäten synchron sind. Der Benutzer hat dann die Möglichkeit, jederzeit von manueller zu gesprochener Interaktion zu wechseln und umgekehrt, ohne dass dabei unterschiedliche Systemzustände von unterschiedlichen Modalitäten widergespiegelt werden.

Sobald sich zwei oder mehr Personen in einem zwischenmenschlichen Dialog befinden, bedeutet das Alternieren nicht, dass die Grenzen zwischen einzelnen Turns immer eindeutig sind. Dasselbe gilt für Dialoge zwischen Mensch und Maschine. Es wird abgewechselt wenn eine Partei gerade anfängt zu sprechen, Pausen auftreten, kurz bevor eine Partei zu Ende gesprochen hat und auch indem eine sprechende Partei inmitten ihrer Äußerung unterbrochen wird (vgl. Kapitel 4, 4.4.8). Die Ergebnisse aus den Experimenten zeigen eine deutliche Neigung des Benutzers zu freierer Eingabe und einem Dialogfluss, der ein natürliches Interaktionsverhalten ermöglicht. Entsprechend ist sequentielle Erkennung keine benutzerfreundliche Option. Der Benutzer muss dabei jedes Mal abwarten, bis ein Systemprompt beendet und das System bereit für eine Spracheingabe ist. Barge-in-Erkennung hingegen ermöglicht eine Abdeckung mehrerer Alternativen, die ein Sprecher bei einem Turnwechsel ergreift. Unabhängig davon erhöht Barge-in-Erkennung die Erfolgsrate bei Taskbearbeitungen und verringert deren Bearbeitungszeit, da Klärungsdialoge von Seiten des Systems aufgrund verpasster wichtiger Informationen vermieden werden.

Feedback geben und empfangen ist ein Vorgang, der den Dialog zwischen Menschen kontinuierlich begleitet. Dadurch wird auf Sprecher- und Hörerseite eine Grundlage für gegenseitiges Verstehen (Grounding) geschaffen. Der Hörer gibt entweder Feedback während er dem Sprecher zuhört (z.B. durch Äußerungen wie ‚ja‘ oder Kopfnicken) oder unmittelbar während seines Übergangs von der Hörer- in die Sprecherrolle (McTear, 2004, S.54; Kapitel 3, 3.1). Im Hinblick auf eine Interaktion mit Maschinen zeigen die durchgeführten

Untersuchungen, dass Feedback (sowohl von Seiten des Systems als auch von Seiten des Benutzers) während des Zuhörens als unangemessen betrachtet wird. Zwischenkommentare wie ‚ja‘, ‚genau‘ oder ‚richtig‘ werden als unpassend und störend empfunden, es sei denn die Äußerungen implizieren die Auswahl oder Bestätigung eines Listenelements durch den Benutzer. Sobald das System sich jedoch in der sprechenden Position befindet, ist es ratsam, auf zweifache Art Feedback zu geben:

1. Das System muss über den aktuellen Systemzustand Rückkopplung geben, insbesondere wenn das Auftreten von Unsicherheiten auf der Benutzerseite wahrscheinlich ist. Da Sprache die entscheidende Modalität während der Fahraufgabe ist, reicht es nicht aus, den entsprechenden Systemzustand visuell wiederzugeben. Der Benutzer sollte ebenfalls sprachlich darüber informiert werden. Dies kann beispielsweise dann erforderlich sein, wenn der Benutzer innerhalb einer Applikation zur nächsten springt. Da Fehlererkennungen häufig auftreten können, sollte ein Systemprompt an dieser Stelle die Benutzereingabe bestätigen, indem beispielsweise die entsprechende Applikation integriert wird: ‚Navigation – welche Funktion?‘. Ähnliches Feedback sollte auch auf der Taskebene gegeben werden, sobald unterschiedliche Applikationen Tasks enthalten, die in ihrer Vorgehensweise ähnlich sind (z.B. das Speichern von Städtenamen, Radiosendern oder Adressbucheinträgen). Durch Hinzufügen eines Schlüsselwortes ist es ein Leichtes, unspezifische Systemprompts wie ‚Bitte sprechen Sie den Namen‘ zu disambiguieren – bevor es zu Verunsicherungen beim Benutzer kommt.
2. Bei der Auswahl von Listenelementen per Textenrolment oder entsprechender Zeilennummer ist es notwendig, dass das System dem Benutzer explizites Feedback darüber erteilt, was verstanden und damit selektiert wurde. Auch diese Bestätigung sollte sowohl visuell als auch sprachlich erfolgen. Nehmen wir als Beispiel die Auswahl eines Städtenamens, Albums oder Adressbucheintrags über die entsprechende Zeilennummer (z.B. „Nummer zwei bitte“) oder ein direktes Sprechen des Eintrags. Ein anschließender Systemprompt „Städtename | Album | Adressbucheintrag übernommen“ wäre nicht ausreichend, da er den vom Benutzer selektierten Eintrag nicht explizit bestätigt. In diesem Fall weiß der Benutzer ohne auf das Display zu sehen nicht, ob sein Eintrag mit dem tatsächlich ausgewählten identisch ist. Diese Unsicherheit kann dadurch behoben werden, indem die entsprechende Zeilennummer und/oder das Textenrolment in den Systemprompt integriert werden.

Das Erweitern unspezifischer Prompts mittels Schlüsselwörtern, Zeilennummern oder Textenrolments macht die Interaktion mit einem System für den Benutzer transparenter. Systemäußerungen können dadurch einfach interpretiert werden. Im Hinblick auf die Grice'sche Konversationsmaxime (Grice, 1989, S.26; Kapitel 3, 3.2.1) erfüllt diese Art des Grounding die Maxime der Quantität und der Art und Weise. Sätze wie „Bitte sprechen Sie den Städtenamen“ oder „Nummer zwei – Zoo Wilhelma. Möchten Sie die Zielführung starten?“ liefern dem Benutzer genau die richtige Menge an Informationen, die er benötigt. Sie sind eindeutig und schließen jegliche Ambiguität aus. Aufforderungen wie z.B. „Bitte sprechen Sie den Namen“ oder „Welche Funktion?“ hingegen vermitteln dem Benutzer zu wenig Information, da daraus nicht eindeutig hervorgeht, auf welche Applikation sich die Äußerung bezieht, es sei denn der Benutzer verifiziert die Aufforderung durch einen Blick auf die Displayanzeige.

Für das Design von Sprachdialogsystemen sollten generell alle Richtlinien von Grice Maßstab sein: die Maxime der Quantität, Qualität, Relation und Art und Weise und das allem zugrunde liegende Kooperationsprinzip. Auf Seiten des Systems kann die Maxime der Qualität selbstverständlich nicht dauerhaft eingehalten werden. Aufgrund der besonderen Fahrzeugumgebung und der Spracherkennungstechnologie können jederzeit Rückweisungs-, Substitutions- und Einfügefehler auftreten, die dem Benutzer falsche Ergebnisse liefern oder zu ungewünschtem Systemverhalten führen (Schmandt, 1994, S.160; Kapitel 4, 4.2). Ein erneutes Auffordern, die Eingabe aufgrund niedriger Konfidenz zu bestätigen oder weil das Erkennungsergebnis im Dialogkontext äußerst unwahrscheinlich ist, ist zwar empfehlenswert, wird jedoch das eigentliche Problem nicht lösen können.

Was die Spezifikation von multimodalen Systemen betrifft, so ist für die Erfüllung der Qualitätsmaxime entscheidend, Dialogbausteine zu entwickeln, die wieder verwendbar sind, und bei deren Spezifikation ein Werkzeug einzusetzen, welches sprachliche und manuelle Interaktion vereint. Dies verbessert die Konsistenz innerhalb einer Modalität und über mehrere Modalitäten hinweg erheblich und macht das eigentliche System weniger fehleranfällig.

Andere Maximen wiederum gilt es absichtlich zu verletzen, da abhängig vom Kontext unterschiedliche Maxime eine höhere Priorität gegenüber anderen Maximen haben. Promptdesign folgt in erster Linie der Quantitätsmaxime, d.h. es sollte so informativ und gleichzeitig so kurz wie möglich sein. Da Sprache temporär ist und dem Benutzer sequentiell

präsentiert wird (Schmandt, 1994, S.102; Balentine, 2001, S.11; Gibbon, 1997, S.82), ist es für den Benutzer schwierig, zu viel Information im Gedächtnis zu behalten, insbesondere dann, wenn er sich primär auf die Fahraufgabe konzentriert. Demzufolge ist es wahrscheinlich, dass der Benutzer am Ende eines Prompts, der eine lange Liste von Menüpunkten beinhaltet, bereits wieder vergessen hat, was zu Beginn gesagt wurde. Die Anzahl von Menüpunkten sollte daher eine Anzahl von drei nicht überschreiten, es sei denn, ein System verfügt über Barge-in-Erkennung.

Für den Fall, dass Informationen oder Anweisungen innerhalb von Systemprompts für den Benutzer irreführend sind, sollte die Quantitätsmaxime zu Gunsten der Maxime der Art und Weise vernachlässigt werden. Der Prompt könnte durch Hinzufügen eines Beispiels verlängert werden, um dem Benutzer den entsprechenden Inhalt klar und deutlich zu vermitteln. Während der Untersuchungen wurde beispielsweise die Aufforderung „Bitte buchstabieren Sie die Straße“ missverstanden. Anstatt den Straßennamen Buchstabe für Buchstabe einzugeben, nannte der Benutzer den vollständigen Namen am Stück. Ein erweiterter Prompt wie z.B. „Bitte buchstabieren Sie die Straße – sagen Sie beispielsweise anstelle von Stuttgarter Straße S-T-U-T-T“ könnte den Benutzer empfänglicher für den eigentlichen Promptinhalt machen.

Wenn es darum geht, von einem Thema zum nächsten zu wechseln, so zeigt zwischenmenschliche Kommunikation, dass dieses Verhalten durchaus typisch ist. Offensichtlich verstößt sie gegen die Relationsmaxime, dahingehend dass Gesprächspartner oftmals nicht konsequent beim Thema bleiben. Ähnliches Verhalten konnte beim Wizard-of-Oz-Experiment im Zusammenhang mit erfahrenen Benutzern beobachtet werden (siehe Kapitel 4, 4.4.11). Sie benutzten zweierlei Taskwechsel: short cuts (Taskwechsel innerhalb einer Applikation) und long cuts (applikationsübergreifende Taskwechsel). Die Erfahrung mit Kommunikation unter Menschen sowie zwischen Mensch und Maschine zeigt, dass die Relationsmaxime in diesem Kontext nicht aufrechterhalten werden sollte. Sie würde außerdem dem Ziel zuwiderlaufen, die Anzahl an Interaktionsschritten möglichst gering zu halten und damit die benötigte Zeit für eine Task zu verringern.

Zugriff auf große Datenbanken mittels Sprachdialogsystemen im Fahrzeug

Während der vergangenen Jahre hat sich der Zugriff auf Audio-, Navigations- und Adressbuchdaten im Fahrzeug in eine mühselige Aufgabe verwandelt. Für jede Applikation gibt es heutzutage eine Vielzahl von elektronischen Geräten. Audioapplikationen können beispielsweise aus folgender Vielfalt bestehen:

- Speichermedien: z.B. CD, Speicherkarte, Harddisk, Flash-Speicher, USB (MP3-Player, iPod)
- Datenformate: Audio, Rohformat
- Dateitypen: z.B. *.mp3, *.mpg, *.ogg, *.wav

Um Audiodaten erfolgreich auswählen zu können, benötigt der Benutzer technisches Verständnis für das System und muss in der Lage sein, sich daran zu erinnern, welches Speichermedium über welche Inhalte verfügt. Applikationen dieser Art sind für den Benutzer nicht transparent. Sie verletzen die Maxime der Art und Weise dahingehend, dass eine Vielzahl von Audiodateinamen auf einer Vielzahl von technischen Geräten weder geordnet noch verständlich ist und damit zu viele kognitive Ressourcen während des Fahrens in Anspruch nimmt. Darüber hinaus sind derzeit gängige Methoden zur Navigation einer zunehmenden Menge von Audiodaten nicht mehr ausreichend. Diese sehen eine Auswahl durch Sprachkommandos vor wie z.B. 'nächstes Medium', 'voriges Medium', 'nächster Titel', 'voriger Titel', eine Auswahl über die entsprechende Zeilennummer oder ein manuelles Durchsuchen der vorhandenen Speichermedien. Der in Kapitel 5 beschriebene Ansatz bezüglich des sprachlichen Zugriffs auf große Datenbanken im Fahrzeug hatte daher zum Ziel, dem Benutzer ein Management seiner Audioapplikation zu bieten, das weder ein Vorwissen über elektronische Geräte erfordert noch über die entsprechenden Audiodaten, die sie enthalten.

Der Ansatz zur Suche von Audiodaten basiert auf drei intuitiven Bedienkonzepten (siehe Kapitel 5, 5.2):

- Eine kategoriebasierte Suche, die die Vorauswahl einer bestimmten Kategorie vorsieht
- Eine kategoriefreie Suche, nach der Audiodateinamen (oder einzelne Bestandteile davon) direkt eingegeben werden können
- Die physikalische Suche als Rückfalllösung

Die kategoriebasierte und kategoriefreie Suche berücksichtigt den Inhalt sämtlicher Speichermedien, die im Fahrzeug angeschlossen sind. Zusätzlich zu den gängigen Möglichkeiten, Audiodaten zu navigieren, erlauben beide Suchmodi, Audiodateinamen sprachlich über sprechbare Texteinträge (Textenrolments) zu selektieren. Um ein manuelles Scrollen durch lange Listen zu vermeiden, weil sich der Benutzer nicht an die genauen Namen von Alben, Titeln etc. erinnert, wurde der Ansatz um Generierungsregeln erweitert. Generierungsregeln bereiten Audiodateinamen beispielsweise im Hinblick auf Sonderzeichen, Abkürzungen, Schlüsselwörter, geschlossene Wortklassen und sekundäre Komponenten auf. Auf diese Weise werden zu den ursprünglichen Einträgen einer Musikdatenbank zusätzliche Wortvarianten generiert. Evidenz für die Notwendigkeit von Generierungsregeln lieferten Untersuchungen über die Auswahl von Musik und persönlichen Adressbuchdaten (Mann, 2008a; Mann, 2007b): Beim Zugriff auf große Datenbanken neigen Benutzer häufig dazu, mehrere Grice'sche Maxime auf einmal zu verletzen, da die Eingabe ihrer Dateinamen mit hoher Wahrscheinlichkeit unvollständig ist. Dies bedeutet, Benutzer machen Angaben, die nur bedingt wahr sind (Maxime der Qualität) und liefern zudem Informationen, die unzureichend sind (Maxime der Quantität). Je unpräziser ihre Eingaben werden, desto häufiger ist die Wahrscheinlichkeit, dass Mehrdeutigkeiten auftreten (Maxime der Art und Weise). Die Erzeugung von zusätzlichen Wortvarianten mit Hilfe von Generierungsregeln umgeht diese Verstöße und, Grices übergeordnetem Kooperationsprinzip folgend, ermöglicht dem Benutzer einen erfolgreichen Zugriff auf Daten, die andernfalls per Sprache unauffindbar wären.

Um die Effizienz der Generierungsregeln zu verifizieren, wurden Sprachdaten von Audiodateinamen aus gängigen Kategorien (Künstler, Album, Titel, Genre, Jahr und Audiobücher) gesammelt. Die Studie umfasste verschiedene Szenarien, die sich von freier Eingabe bis hin zur Rückerinnerung von vorgegebenen Audiodateinamen erstreckten (Kapitel 5, 5.5). Task 1 begann mit freier Eingabe, um herauszufinden, wie Benutzer ihre Musik ohne jegliche Einschränkungen auswählen. Bei Task 2 wurden individuell bevorzugte Audiodateinamen nach vorgegebenen Kategorien eingegeben. Diese Task sollte Aufschluss darüber geben, wie detailliert das Wissen der Testpersonen hinsichtlich Musikdateinamen ist. Die Vorgaben in dieser Task sahen sowohl eine Eingabe von einzelnen Kategorien sowie eine Kombination von zwei Kategorien vor. In Task 3 wurden die Testpersonen gebeten, vorgegebene Paare von Audiodateinamen wiederzugeben. Die dabei beabsichtigte kognitive

Überlastung führte die Testpersonen dazu, entscheidende Komponenten herauszufiltern, sobald ein Paar ausgeblendet wurde.

Die Analyse der Daten zeigt, dass die Anzahl der Treffer innerhalb der Kategorie Künstler durch das Regelset von ursprünglichen 61 Prozent auf 85 Prozent erhöht werden konnte. Bei den Titelnamen stieg die Trefferrate von 51 Prozent auf 69 Prozent. Die Abdeckung von Albumnamen hingegen konnte mit Hilfe der Generierungsregeln lediglich um 14 Prozentpunkte erhöht werden, von 60 auf 74 Prozent. Die Tatsache, dass die Gesamtabdeckung von Alben- und Titelnamen im Vergleich zu Künstlernamen geringer ist, lässt sich auf Versprecher zurückführen, die häufig dann produziert wurden, wenn eine Testperson im Umgang mit einer Fremdsprache wie insbesondere Französisch, Italienisch und Spanisch nicht vertraut war.

Ergebnisse aus Erkennertests zeigen, dass die Anzahl von Spracheingabeparametern bei der Musiksuche von einer Kategorie auf eine optionale Eingabe von ein oder zwei Kategorien in beliebiger Reihenfolge erweitert werden kann. Diese Option geht Hand in Hand mit den Ergebnissen aus Task 1, gemäß denen die uneingeschränkte Eingabe der Versuchspersonen eine Kombination von zwei Kategorien pro Äußerung nicht überschritt.

Zusammenfassend kann man festhalten, dass die in diesem Ansatz entwickelten Konzepte den Zugriff auf große Datenbanken im Fahrzeug mittels sprachlicher Eingabe erheblich erleichtern. Aufgrund der heutigen Vielfalt an elektronischen Geräten sind diese Datenbanken bereits im Fahrzeug verfügbar oder es ist ein Leichtes, sie ins Fahrzeug zu integrieren. Die Konzepte stellen einen ersten Schritt in Richtung Mensch-Maschine-Interaktion dar, die dem Benutzer eine freiere Eingabe ermöglicht. Derzeitig verfügbare Methoden, diese Daten zu navigieren, sind nicht länger adäquat, da sie auf eine einst überschaubare Menge von Musiktiteln oder Adressbucheinträgen ausgerichtet waren. Für die Erstellung allgemeiner Richtlinien eines Benutzerinterface im Fahrzeug wurde einerseits die zwischenmenschliche Kommunikation untersucht, um daraus Prinzipien für eine Interaktion zwischen Mensch und Maschine abzuleiten. Andererseits wurden zahlreiche Studien durchgeführt, um herauszufinden, wie sich Benutzer im Kontext einer Mensch-Maschine-Interaktion verhalten und wo Probleme bezüglich der Bedienbarkeit von Sprachdialogsystemen im Fahrzeug auftreten. Die mangelnde Erinnerungsgüte von Audio-, Sonderziel- und Adressbuchdaten auf Seiten des Benutzers fand besondere Berücksichtigung in den Konzepten. Die entwickelten Konzepte bringen Transparenz

in eine Vielzahl von technischen Geräten und großen Datenmengen, während sie gleichzeitig Konsistenz sowohl innerhalb einer Modalität als auch modalitätsübergreifend gewährleisten.

Aufbau der Arbeit

Das **Kapitel 2** vermittelt die wesentlichen Grundlagen zum aktuellen Stand von multimodalen Dialogsystemen. Neben einer Einführung der für diese Arbeit relevanten Begriffe wird insbesondere der spezielle Status von Sprachdialogsystemen im Fahrzeug beleuchtet: inwieweit unterscheiden sich Sprachdialogsysteme im Fahrzeug von anderen Sprachdialogsystemen und welche Probleme ergeben sich aufgrund der Fahrzeugumgebung.

Da gesprochene Sprache eine wichtige Rolle bei der Mensch-Maschine-Interaktion spielt, wird in **Kapitel 3** zunächst gesprochene Sprache als natürliches Mittel der zwischenmenschlichen Kommunikation untersucht. Eine Analyse von Kommunikationsprinzipien und Diskurs bildet die Grundlage für die Entwicklung von Strategien für kooperative Bedienkonzepte bei Sprachdialogsystemen, welche in die folgenden Kapitel einfließt. Nicht außer Acht gelassen werden darf dabei die Tatsache, dass zwischenmenschliche Dialoge oftmals von Prinzipien abweichen. Diese Verstöße gilt es ebenfalls mittels geeigneter Strategien in einen verbalen Austausch zwischen Mensch und Maschine zu integrieren.

Eine weitere Grundlage für die Entwicklung von kooperativen Bedienkonzepten und deren Integration in ein multimodales Interface ist die Evaluierung von aktuellen Fahrzeugapplikationen. Sie zeigt potentielle Schwierigkeiten auf, die im Umgang mit einem Sprachdialogsystem im Fahrzeug entstehen können. **Kapitel 4** beschreibt zunächst die Architektur, Funktionalität und Grenzen von multimodalen Sprachdialogsystemen. Des Weiteren werden verschiedene Untersuchungsmethoden eingeführt und erläutert, die im Rahmen dieser Arbeit durchgeführt wurden. Empfehlungen für das Design von multimodalen Sprachinterfaces im Fahrzeug bilden den Kernpunkt des vierten Kapitels. Die Empfehlungen resultieren aus einer Vernetzung von Aspekten der zwischenmenschlichen Kommunikation und der Benutzerbedürfnisse, die sich während der Benutzerstudien heraus kristallisiert haben. Der Fokus ist dabei auf allgemeines Dialogdesign gerichtet.

In einem weiteren Schritt werden in **Kapitel 5** Bedienkonzepte für kooperative Sprachdialogsysteme entwickelt. Ziel ist es, einen benutzerfreundlichen Zugriff auf große Datenbanken, wie sie beispielsweise bei Audio-, Navigations- und Adressbuchanwendungen auftreten, zu ermöglichen. Aufgrund der steigenden Anzahl und Komplexität von elektronischen Geräten im Fahrzeug sind derzeitige Zugriffs- und Navigationsmethoden bei Audio-, Ziel- und Adressbuchdaten nicht länger ausreichend. Entsprechende Anwendungen dafür sind für den Benutzer nicht mehr transparent. Die vorgestellten Konzepte kombinieren verschiedene Suchstrategien mit Empfehlungen des vorangegangenen Kapitels und ermöglichen eine benutzerfreundliche Mensch-Maschine-Interaktion sowohl für Novizen als auch für Experten.

Eine Kombination von manueller Ein-/Ausgabe und sprachlicher Ein-/Ausgabe, wie sie innerhalb von Fahrzeugen mit HMI-Systemen auftritt, stellt zusätzliche Anforderungen an eine erfolgreiche Kommunikation zwischen Mensch und Maschine. Wichtig ist daher, dass jede Art von Benutzereingabe stets zwischen beiden Modalitäten ausgetauscht wird. Indem beide Modalitäten synchron sind, kann das manuelle Interface eine hilfreiche Ergänzung für das sprachliche Interface darstellen und umgekehrt. Der Dialog zwischen Mensch und Maschine wird dadurch unterstützt und ermöglicht dem Benutzer ferner, innerhalb einer Task einen Modalitätswechsel zu vollziehen, ohne deshalb mit der Task von vorne beginnen zu müssen. Die Mensch-Maschine-Interaktion wird dadurch effizienter und trägt nicht unwesentlich zu Benutzerakzeptanz bei.

Im **Kapitel 6** werden die Ideen und Ergebnisse der vorangegangenen Kapitel zusammengefasst. Daraus resultierende Ansätze für künftige Untersuchungen werden in einem Ausblick erläutert, mit dem die vorliegende Arbeit abschließt.

References

- André, E., Rehm, M., Minker, W. and Bühler, D. (2004). Endowing spoken language dialogue systems with emotional intelligence. In: E. André, L. Dybkjaer, W. Minker and P. Heisterkamp (eds.), *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 2004, Proceedings*. Springer-Verlag, Berlin/ Heidelberg, 178-187.
- Audi (2005). *Bedienungsanleitung Audi A8 Infotainment/MMI – deutsch 5.05*.
- Austin, J.L. (1962). *How to Do Things with Words*. Oxford University Press, Oxford.
- Automobilsport (2008). Web page. Available at: <http://www.automobilsport.com/news-mercedes-benz-in-car-iphone-connection-easier-vehicle-architecture-germany-apple--39509.html>
- Balentine, B. and Morgan, D.P. (2001). *How to Build a Speech Recognition Application. A Style Guide for Telephony Dialogues*. 2nd edition, EIG press, California.
- Beckett, S. (1952). *En attendant Godot*. Minit, Paris.
- Bellert, I. (1970). On a condition of the coherence of texts. *Semiotica* 2, 335-363.
- Bernsen, N.O., Dybkjaer, H. and Dybkjaer, L. (1998). *Designing Interactive Speech Systems – From First Ideas to User Testing*. Springer, London.
- Berton, A., Mann, S. and Regel-Brietzmann, P. (2007). How to access large navigation databases in cars by speech. In: K. Fellbaum (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, Band 46, 18.Konferenz, Dresden: TUD-press, Cottbus, 155-162.
- Berton, A., Schreiner, O. and Hagen, A. (2005). How to speed up voice-activated destination entry. In: *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, 16.Konferenz, Dresden: TUD-press, Prag.
- Berton, A. (2004). *Konfidenzmaße und deren Anwendungen in der automatischen Sprachverarbeitung*. PhD thesis, W.e.b. Universitätsverlag & Buchhandel Eckhard Richter & Co. OHG, Dresden.
- BMW (2006). *Bedienungsanleitung BMW 740i*. Bestell-Nr. 01400012268, deutsch.
- Bolt, R.A. (1980). “Put-that-there”: Voice and gesture at the graphics interface. In: *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, 262-270.
- Braun, D., Sivils, J., Shapiro, A. and Versteegh, J. (2008). Unified Modeling Language (UML) Tutorial. Available at: http://atlas.kennesaw.edu/~dbraun/csis4650/A&D/UML_tutorial/index.htm

- Bühler, K. (1934). *Sprachtheorie*. Fisher, Jena. Neudruck 1965, Stuttgart.
- Bunt, H. (1979). Conversational principles in question-answer dialogues. In: D. Krallmann and G. Stickel (eds.), *Zur Theorie der Frage*. Narr, Tübingen.
- Burch, D. (2002). *The mobile phone report – A report on the effects of using a ‘hand-held’ and ‘hands-free’ mobile phone on road safety*. Direct Line Motor Insurance. Available at: http://www.dft.gov.uk/think_media/241042/241120/02-mobilephonereport-directline
- Burnett, D. (ed.) (2000). *SpeechObjects Specification*. W3C, Nuance Communications. Available at: <http://www.w3.org/TR/speechobjects/>
- Burns, P.C. and Lansdown, T.C. (2002). *E-distraction: The challenges for safe and usable internet services in vehicles*. NHTSA Internet Forum on Driver Distraction. Available at: <http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/PDF/29.PDF>
- Bußmann, H. (1983). *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, 1. Auflage, Stuttgart.
- Bußmann, H. (1990). *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, 2. Auflage, Stuttgart.
- Chomsky, N. (2006). *Language and Mind*. Cambridge University Press, Cambridge.
- Cohen, M., Giangola, J. and Balogh, J. (2004). *Voice User Interface Design*. Addison-Wesley, Boston.
- Cohen, P., McGee, D., Clow, J. (2000). The efficiency of multimodal interaction for a map-based task. In: *Proceedings of the sixth conference on Applied Natural Language Processing*, Morgan Kaufmann, Seattle, Washington, 331-338.
- Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., Varile, G.B. and Zampolli, A. (eds.) (1996). *Survey of the State of the Art in Human Language Technology*. Center for Spoken Language Understanding (CSLU), World Wide Web, Oregon Graduate Institute. Available at: <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>
- Corley, M. and Stewart, O.W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass* 2/4, 589-602.
- Crain, S. and Lillo-Martin, D. (1999). *An Introduction to Linguistic Theory and Language Acquisition*. Blackwell Publishers Ltd., Oxford.
- Daimler AG (2008). *Betriebsanleitung für die S-Klasse – Online Version*. Available at: http://www.mercedesbenz.de/content/germany/mpc/mpc_germany_website/de/home_mpc/passenger_cars/home/services/interactive_manuals.html Latest Version 2009: http://www.mercedesbenz.de/content/germany/mpc/mpc_germany_website/de/home_mpc/passengercars/home/servicesandaccessories/services_online/interactive_manual.html

-
- Dausend, M., Berton, A., Kaltenmeier, A. and Mann, S. (2008). Was möchten Sie hören? – Zielsicheres Suchen in großen Datenmengen mit integrierten multimodalen Systemen. In: A. Lacroix (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, Band 50, 19. Konferenz, Dresden: TUD-press, Frankfurt am Main, 77-85.
- EC, European Commission (2009). Music knows no borders. Available at: http://ec.europa.eu/news/culture/070301_1_en.htm
- EC, European Commission (2006). Commission Recommendation of 22 December 2006 on safe and efficient in-vehicle information and communication systems: update of the European Statement of Principles on human machine interface. *Official Journal of the European Union*. Available at: http://eur-lex.europa.eu/LexUriServ/site/en/oj/2007/l_032/l_03220070206en02000241.pdf
- Ehrlich, U. and Jersak, T. (2006). Definition und Konfiguration der Wissensbasen und Schnittstellen von Sprachdialogapplikationen mit XML. In: *Proceedings of XML Tage*, Berlin, 51-61.
- Ehrlich, U. (1990). *Bedeutungsanalyse in einem sprachverstehenden System unter Berücksichtigung pragmatischer Faktoren*. PhD thesis, Max Niemeyer Verlag GmbH & Co. KG, Tübingen.
- Enigk, H. et al. (2004). *Internal Study: Akzeptanz von Sprachbediensystemen im PKW – Längsschnittstudie*. DaimlerChrysler AG, Berlin.
- Enigk, H. and Meyer zu Kniendorf, C. (2004). *Internal Study: Akzeptanz von Sprachbediensystemen im PKW – Anforderungsanalyse zur Struktur und Nutzung von Adressbuchdaten*. DaimlerChrysler AG, Berlin.
- Fox, B.A. (1993). Discourse structure and anaphora: written and conversational English. *Cambridge studies in linguistics*, 48. Cambridge University Press, Cambridge.
- Fromkin, V. and Rodman, R. (1993). *An Introduction to Language*. Harcourt Brace College Publishers, Orlando.
- Gellatly, A.W. (1997). *The Use of Speech Recognition Technology in Automotive Applications*. PhD thesis, Blacksburg, Virginia.
- Gibbon, D., Mertins, I. and Moore, R.K. (eds.) (2000). *Handbook of Multimodal and Spoken Dialogue Systems*. Kluwer Academic Publishers, Norwell, Massachusetts.
- Gibbon, D., Moore, R. and Winski, R. (eds.) (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter & Co., Berlin.
- Goodwin, C. (1981). Conversational organization: Interaction between speakers and hearers. In: E.A. Hammel (ed.), *Language, Thought and Culture: Advances in the Study of Cognition*. Academic Press, New York.

- Goronzy, S. and Beringer, N. (2005). Integrated development and on-the-fly simulation of multimodal dialogs. In: *Proceedings of the Interspeech*, Sixth Annual Conference of the International Speech Communication Association, Lisbon, Portugal, 2477-2480.
- Gracenote MediaVOCS (2007). Web page. Available at:
http://www.gracenote.com/gn_products/mediaVOCS.html
- Greenbaum, S. and Quirk, R. (1990). *A Student's Grammar of the English Language*. Longman Group UK Limited, Essex.
- Grice, H.P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts, London, England.
- Haegeman, L. (1998). *Introduction to Government & Binding Theory*. 2nd edition, Blackwell Publishers.
- Hänsler, E. and Schmidt, G. (eds.) (2006). *Topics in Acoustic Echo and Noise Control. Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing*. Springer-Verlag, Berlin Heidelberg.
- Harel, D. (2007). Statecharts in the making: a personal account. In: *Proceedings of the third ACM SIGPLAN conference on History of programming languages*, San Diego, California.
- Harris, R.A. (2005). *Voice Interaction Design – Crafting the New Conversational Speech Systems*. Elsevier Inc., San Francisco, California.
- Harris, Z.S. (1952). Discourse analysis. *Language* 28:1, 1-30.
- Heidingsfelder, M., Kintz, E., Petry, R., Hensley, P., Sedran, T., Reers, J., Berret, M., Endo, I. and Tanji, K. (2001). *Telematics: How to hit a moving target. A roadmap to success in the Telematics arena*. Roland Berger Strategy Consultants, Detroit, Stuttgart, Tokyo.
- Heinrich, A.T. (2007). *Zustandsdarstellung für die Spezifikation der Sprachbedienung im KFZ*. Diploma thesis, Daimler AG, Ulm.
- Heisterkamp, P. (2003). "Do not attempt to light with match!": Some thoughts on progress and research goals in Spoken Dialog Systems. In: *Proceedings of Eurospeech 03*, Geneva, Switzerland.
- Heisterkamp, P. (2001). Linguatronic: Product-level speech system for Mercedes-Benz cars. In: *Proceedings of Human Language Technology (HLT)*, San Diego, California.
- Herbst, T., Stoll, R. and Westermayr, R. (1991). *Terminologie der Sprachbeschreibung*. Max Hueber Verlag, Ismaning, Germany.

-
- Heute, U. (2006). Noise reduction. In: E. Hänsler and G. Schmidt (eds.), *Topics in Acoustic Echo and Noise Control. Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing*. Springer-Verlag, Berlin Heidelberg, 325-384.
- Huckvale, M. (1996). *Learning from the experience of building automatic speech recognition systems*. Speech Hearing and Language, Phonetics and Linguistics, University College London, London. Available at: <http://www.phon.ucl.ac.uk/home/shl9/markh/huckvale.htm>
- Hüning, H. et al. (2003). *DaimlerChrysler AG Internal Study: Results of a Wizard of Oz Experiment*. DaimlerChrysler AG, Ulm.
- Hulstijn, J. (2000). Modelling usability – Development methods for dialogue systems. *Natural Language Engineering, 1*, 1-16.
- Hunt, A. and McGlashan, S. (eds.) (2004). *Speech Recognition Grammar Specification Version 1.0*. W3C. Available at: <http://www.w3.org/TR/2004/REC-speech-grammar-20040316/>
- Hunt, A. (ed.) (2000). *JSpeech Grammar Format*. W3C. Available at: <http://www.w3.org/TR/jsgf/>
- IBM (2009). Web page. Available at: http://www01.ibm.com/software/pervasive/embedded_viavoice/
- IBM (2002). *Reusable Dialogue Components for VoiceXML Applications*. 4th edition, International Business Machines Corporation, USA.
- ISO (2009). International Organization of Standardisation. Available at: <http://www.iso.org/iso/home.htm>
- JAMA (2004). *Guidelines for in-vehicle display systems, version 3.0*. Japan Automobile Manufacturers Association. Available at: http://www.jama.or.jp/safe/guideline/pdf/jama_guideline_v30_en.pdf
- Jelinek, F. (1990). Self-organized language modelling for speech recognition. In: A. Waibel and K. Lee (eds.), *Readings in Speech Recognition*. Morgan Kaufmann Publishers, San Francisco, CA, 450-506.
- Jersak, T., Kronenberg, S. and Mann, S. (2006). Command-and-Control-Dialogführung im Fahrzeug mit optimiertem Barge-In. In: R. Hoffmann (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, Band 42, 17. Konferenz, Dresden: TUD-press, Freiberg, 142-149.
- Jersak, T. (2004). *Sprachdialog-Benutzerschnittstelle mit optimierter Barge-In-Funktionalität im Prototyp eines multimodalen WAP-Browsers*. Project at DaimlerChrysler AG, Research and Technology Centre Ulm, diploma thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Stuttgart.

- Kaspar, B., Schuhmacher, K. and Feldes, S. (1997). Barge-in revised. In: *Proceedings of Eurospeech 97*, Rhodes, Greece, 673-676.
- Kölzer, A. (2002). *DiaMod – ein Werkzeugsystem zur Modellierung natürlichsprachlicher Dialoge*. PhD thesis, Mensch und Buch Verlag, Berlin.
- Kronenberg, S. (2001). *Cooperation in Human-Computer Communication*. PhD thesis, Universität Bielefeld, Technische Fakultät, Bielefeld. URN (NBN): urn:nbn:de:hbz:361-3425.
- Kruijff-Korbayová, I., Becker, T., Blaylock, N., Gerstenberger, C., Kaißer, M., Poller, P., Schehl, J. and Rieser, V. (2005). Presentation strategies for flexible multimodal interaction with a music player. In: *Proceedings of DIALOR '05, Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*, Nancy.
- Kuhn, T., Fetter, P., Kaltenmeier, A. and Regel-Brietzmann, P. (1996). DP-based wordgraph pruning. In: *Proceedings ICASSP'96*, Volume 2, Atlanta, USA.
- Lee, J.D, Caven, B., Haake, S. and Brown, T.L. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors*, 43, 631-640.
- Levinson, S.C. (1983). *Pragmatics*. University Press, Cambridge.
- Lombard, E. (1911). Le signe de l'élévation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37, 101-119.
- Mann, S., Berton, A. and Heisterkamp, P. (2008a). *Daimler AG Internal Study: Speech Database Recordings for the Application Audio*. Daimler AG, Ulm.
- Mann, S., Berton, A., Dausend, M. and Heisterkamp, P. (2008b). "Beethoven's Ninth" – An experiment on naming usage for audio files. In: A. Lacroix (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, Band 50, 19.Konferenz, Dresden: TUD-press, Frankfurt am Main, 124-132.
- Mann, S., Berton, A., Dausend, M. and Eberhardt, A. (2008c). *Daimler AG Internal Study: Accessing Points of Interest by Speech. A Comparative Study of Current Mercedes-Benz Series Versus Prototype System*. Daimler AG, Ulm.
- Mann, S., Berton, A. and Ehrlich, U. (2007a). How to access audio files of large data bases using in-car speech dialogue systems. In: *Proceedings of the Interspeech*, Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 138-141.
- Mann, S., Berton, A. and Ehrlich, U. (2007b). A multimodal dialogue system for interacting with large audio databases in the car. In: K. Fellbaum (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studentexte zur Sprachkommunikation, Band 46, 18.Konferenz, Dresden: TUD-press, Cottbus, 202-209.

-
- Mann, S., Heisterkamp, P., Hüning, H., Jersak, T. and Kronenberg, S. (2006). How to systematically store and retrieve voice-enrolled list elements in spoken dialogue systems. In: R. Hoffmann (ed.), *Elektronische Sprachsignalverarbeitung (ESSV)*, Studententexte zur Sprachkommunikation, Band 42, 17. Konferenz, Dresden: TUD-press, Freiberg, 173-178.
- Mann, S. (2003). *User Concepts for Spoken Dialogue Systems in Car Environments*. Project at DaimlerChrysler AG, Research and Technology Centre Ulm, Magisterarbeit, Universität Stuttgart, Stuttgart.
- McTear, M.F. (2004). *Spoken Dialogue Technology – Toward the Conversational User Interface*. Springer, London.
- McTear, M.F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, 34, 1-80.
- Mercedes-Benz (2003). *Betriebsanleitung Linguatronic*. Bestell-Nr. 6515969400, deutsch.
- MindSpring Enterprises (1997). Phonetic Alphabets. Available at: <ftp://ftp.cs.ruu.nl/pub/NEWS.ANSWERS/radio/phonetic-alph/full/>
- Mori, R.D., Béchet, F., Hakkani-Tür, D., McTear, M., Riccardi, G. and Tur, G. (2008). Spoken language understanding. *Signal Processing Magazine, IEEE*, 25, 50-58.
- Müller, V. (2003). *Die Deixis im "Theater des Absurden"*. PhD thesis, Universität Stuttgart, Stuttgart. URN (NBN): urn:nbn:de:bsz:93-opus-21402
- Musicoverly (2009). Web page. Available at: <http://musicoverly.com>
- NHTSA, National Highway Traffic Safety Administration (2006). *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. National Technical Information Service, Springfield, Virginia.
- Nielsen, J. (2005). Heuristics for user interface design. Available at: http://www.useit.com/papers/heuristic/heuristic_list.html
- Nielsen, J. (1993). *Usability Engineering*, Morgan Kaufmann, San Francisco, CA.
- Nuance (2009). Web page. Available at: <http://www.nuance.de/naturallyspeaking/>
- Ogawa, T. (2007). Adequacy analysis of simulation-based assessment of speech recognition system. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, 1153-1156.
- Oomen, C.C.E. and Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research* 30, 163-184.
- Oviatt, S.L. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42, 74-81.

- Oviatt, S.L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12, 93-129.
- Oviatt, S. L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9, 19-35.
- Pfeil, M., Buehler, D., Gruhn, R. and Minker, W. (2008). Evaluating text normalization for speech-based media selection. In: *Perception in Multimodal Dialogue Systems*. Springer-Verlag, Berlin Heidelberg, 52-59.
- Philips (2009). Web page. Available at: <http://www.myspeech.com/>
- Philopoulos, A. (2002). *Speech Based Interaction for In-Vehicle Information Systems: A Design Case*. Project at DaimlerChrysler AG, Research and Technology Centre Ulm, for the Degree of Master in Technological Design in User System Interaction, Technische Universiteit Eindhoven, Eindhoven.
- Pinker, S. (1994). *The Language Instinct*. Penguin Books Ltd., London.
- Puder, H. (2006). Noise reduction with Kalman-Filters for hands-free car phones based on parametric spectral speech and noise estimates. In: E. Hänsler and G. Schmidt (eds.), *Topics in Acoustic Echo and Noise Control. Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing*. Springer-Verlag, Berlin Heidelberg, 385-427.
- Radford, A. (1988). *Transformational Grammar*. Cambridge University Press, Cambridge.
- Ranney, T.A., Mazzae, E., Garrott, R. and Goodman, M.J. (2000). *NHTSA driver distraction research: Past, present and future*. NHTSA Internet Forum on Driver Distraction. Available at: <http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/Papers.htm>
- Reenskaug, T. (1979). Thing-Model-View-Editor - an example from a planning system. Technical Note, Xerox PARC. Available at: <http://heim.ifi.uio.no/~trygver/mvc/index.html>
- Reithinger, N. and Herzog, G. (2006). An exemplary interaction with SmartKom. In: W. Wahlster (ed.), *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer-Verlag, Berlin Heidelberg, 41-52.
- Reithinger, N., Bergweiler, S., Engel, R., Herzog, G., Pflieger, N., Romanelli, M. and Sonntag, D. (2005). A look under the hood: design and development of the first SmartWeb system demonstrator. In: *Proceedings of the Seventh International Conference on Multimodal Interfaces*, Trento, Italy. Available at: <http://delivery.acm.org/10.1145/1090000/1088492/p159-reithinger.pdf?key1=1088492&key2=8430840421&coll=GUIDE&dl=GUIDE&CFID=31619933&CFTOKEN=96731198>
- Roberts, I. (1997). *Comparative Syntax*. Arnold, a member of the Hodder Headline Group, London & New York.

-
- Rosendahl, I. et al. (2006). *DaimlerChrysler AG Internal Study: Analysis of Customer Needs in Context of an In-car Search Engine*. DaimlerChrysler AG, Stuttgart.
- Saab Automobile AB (2003). *Bedienungsanleitung Saab 93 Sport-Limousine*. Bestell-Nr. 425652, deutsch.
- Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organization of turntaking for conversation. *Language*, 50, 696-735.
- SAE (2007). Society of Automotive Engineers International. Available at: <http://www.sae.org/servlets/index>
- Saussure, F. de (1975). *Cours de Linguistique Générale*. Payot, Paris.
- Schleß, V. (2000). *Automatische Erkennung von gestörten Sprachsignalen*. PhD thesis, Shaker Verlag, Aachen.
- Schmandt, C. (1994). *Voice Communication with Computers*. Van Nostrand Reinhold, New York.
- Schmidt, G. and Haulick, T. (2006). Signal processing for in-car communication systems. In: E. Hänsler and G. Schmidt (eds.), *Topics in Acoustic Echo and Noise Control. Selected Methods for the Cancellation of Acoustic Echoes, the Reduction of Background Noise, and Speech Processing*. Springer-Verlag, Berlin Heidelberg, 547-598.
- Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., LeGoff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K. and Blauert, J. (1995). *A taxonomy of multimodal interaction in the human information processing system. Technical Report, Esprit Project 8579* MIAMI. Nijmegen Institute for Cognition and Information (NICI), Nijmegen. Available at: <http://www.ai.rug.nl/~lambert/publications.html>
- Schreiner, O. (to appear). *Ansätze für die automatische Spracherkennung von grossen Listen in Embedded-Systemen*. PhD thesis, Ulm University, Ulm.
- Schulte, J. (ed.) (2003). *Ludwig Wittgenstein – Philosophische Untersuchungen. Auf der Grundlage der kritisch-genetischen Edition*. Suhrkamp Verlag, Frankfurt am Main.
- Schulz, C. H., Rubinstein, D., Diamantakos, D., Kaißer, M., Schehl, J., Romanelli, M., Kleinbauer, T., Klüter, A., Klakow, D., Becker, T. and Alexandersson, J. (2004). A spoken language front-end for a multilingual music data base. In: *Proceedings of XML Tage*, Berlin, 276-290.
- Schwartz, B. (2004). *The Paradox of Choice – Why Less Is More*. HarperCollins Publishers Inc., New York.
- Schwarz, M. and Chur J. (1996). *Semantik: ein Arbeitsbuch*. Gunter Narr Verlag, Tübingen.
- Searle, J.R. (1979). *Expression and Meaning*. Cambridge University Press, Cambridge.

- Searle, J.R. (1969). *Speech acts – an Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.
- Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 4th edition, Addison-Wesley, Amsterdam.
- SIL International (2004). Glossary of linguistic terms. Available at: <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAPresupposition.htm>
- Silbernagel, D. (1979). *Taschenatlas der Physiologie*. Thieme Verlag, Stuttgart.
- Strawson, P.F. (1952). *Introduction to Logical Theory*. Methuen, London.
- TALK, EU project. Talk and Look: Tools for Ambient Linguistic Knowledge. Available at: <http://www.talk-project.org>
- Transport Canada (2003). *Strategies for reducing driver distraction from in-vehicle telematics devices: A discussion document*. Standards Research and Development Branch, Road Safety and Motor Vehicle Regulations Directorate, Canada.
- Tsimhoni, O., Smith, D. and Green, P. (2004). Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human Factors*, 46, 600-610.
- Van Tichelen, L. and Burke, D. (eds.) (2006). *Semantic Interpretation for Speech Recognition (SISR) Version 1.0*. W3C. Available at: <http://www.w3.org/TR/2006/CR-semantic-interpretation-20060111/>
- Vaseghi, S.V. (2006). *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons Ltd., West Sussex, England.
- Wahlster, W. (2007). SmartWeb – ein multimodales Dialogsystem für das semantische Web. In: B. Reuse and R. Vollmar (eds.), *40 Jahre Informatikforschung in Deutschland*. Springer-Verlag, Berlin Heidelberg. Available at: http://smartweb.dfki.de/Vortraege/SmartWeb_Ein_multimodales_Dialogsystem_fuer_das_semantische_Web.pdf
- Wahlster, W. (ed.) (2006). *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer-Verlag, Berlin Heidelberg.
- Wahlster, W. (2006). Dialogue systems go multimodal: The SmartKom experience. In: W. Wahlster (ed.), *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer-Verlag, Berlin Heidelberg, 3-27.
- Wang, J.S., Knipling, R.R., and Goodman, M.J. (1996). The role of driver inattention in crashes; new statistics from the 1995 crashworthiness data system. *40th Annual Proceedings of the Association for the Advancement of Automotive Medicine*, Vancouver. Available at: http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/777.pdf

- Wang, Y., Hamerich, S., Hennecke, M. and Schubert, V. (2005). Speech-controlled media file selection on embedded systems. *SIGdial Workshop*, Lisbon, Portugal.
- Weevers, I. (2004). *“I’d Rather Play a Speech Game than Read the Manual”*: A Game-Based Approach for Learning How to Use an In-Vehicle Speech Interface. Project at DaimlerChrysler AG, Research and Technology Centre Ulm, for the Degree of Professional Doctorate in Engineering in User System Interaction, Technische Universiteit Eindhoven, Eindhoven.
- Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM, Volume 9, Number 1*, 36-45. Available at: <http://www.fas.harvard.edu/~lib51/files/classics-eliza1966.html>
- Young, K., Regan, M. and Hammer, M. (2003). *Driver distraction: A review of the literature*. Monash University Accident Research Centre Australia, report no. 206. Available at: <http://www.monash.edu.au/muarc/reports/muarc206.pdf>

A Usability Guidelines

A.1 Jakob Nielsen: Ten Usability Heuristics

The following principles on user interface design are adopted from Nielsen (2005). He calls them "heuristics" because they are more in the nature of rules of thumb than specific usability guidelines.

Visibility of system status – The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

Match between system and the real world – The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

User control and freedom – Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency and standards – Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

Error prevention – Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

Recognition rather than recall – Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the

dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Flexibility and efficiency of use – Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetic and minimalist design – Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Help users recognise, diagnose, and recover from errors – Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Help and documentation – Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

A.2 Ben Shneiderman: Eight Golden Rules

These rules have been adopted from Shneiderman (2004, p.74).

Strive for consistency – Consistent sequences of actions should be required in similar situations; identical terminology should be used in prompts, menus, and help screens; and consistent commands should be employed throughout.

Enable frequent users to use shortcuts – As the frequency of use increases, so do the user's desires to reduce the number of interactions and to increase the pace of interaction. Abbreviations, function keys, hidden commands, and macro facilities are very helpful to an expert user.

Offer informative feedback – For every operator action, there should be some system feedback. For frequent and minor actions, the response can be modest, while for infrequent and major actions, the response should be more substantial.

Design dialog to yield closure – Sequences of actions should be organized into groups with a beginning, middle, and end. The informative feedback at the completion of a group of actions gives the operators the satisfaction of accomplishment, a sense of relief, the signal to drop contingency plans and options from their minds, and an indication that the way is clear to prepare for the next group of actions.

Offer simple error handling – As much as possible, design the system so the user cannot make a serious error. If an error is made, the system should be able to detect the error and offer simple, comprehensible mechanisms for handling the error.

Permit easy reversal of actions – This feature relieves anxiety, since the user knows that errors can be undone; it thus encourages exploration of unfamiliar options. The units of reversibility may be a single action, a data entry, or a complete group of actions.

Support internal locus of control – Experienced operators strongly desire the sense that they are in charge of the system and that the system responds to their actions. Design the system to make users the initiators of actions rather than the responders.

Reduce short-term memory load – The limitation of human information processing in short-term memory requires that displays be kept simple, multiple page displays be consolidated, window-motion frequency be reduced, and sufficient training time be allotted for codes, mnemonics, and sequences of actions.

A.3 Sharon Oviatt: Ten myths of multimodal interaction

In context with multimodal interaction Oviatt (1999) has come up with 10 interaction myths.

If you build a multimodal system, users will interact multimodally.

Speech and pointing is the dominant multimodal integration pattern.

Multimodal input involves simultaneous signals.

Speech is the primary input mode in any multimodal system that includes it.

Multimodal language does not differ linguistically from unimodal language.

Multimodal integration involves redundancy of content between modes.

Individual error-prone recognition technologies combine multimodally to produce even greater unreliability.

All users' multimodal commands are integrated in a uniform way.

Different input modes are capable of transmitting comparable content.

Enhanced efficiency is the main advantage of multimodal systems.

B Speech recordings of audio file names

B.1 Questionnaires

B.1.1 General information***Persönliche Angaben******Allgemeine Informationen***

1. Name: _____

2. Emailadresse: _____

3. Beruf: _____

4. Welchen Dialekt sprechen Sie? _____

5. Geschlecht männlich weiblich
 6. Alter

18 – 25	26 – 35	36 – 45	46 – 55	56 – 65
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fahrerfahrung

7. Haben Sie einen Führerschein? Ja Nein

Wenn ja, wie viele Kilometer fahren Sie etwa pro Jahr? _____

Sprachbediente Systeme

8. Haben Sie jemals eine über Sprache bediente Anwendung benutzt?

Ja Nein

Wenn ja, was für eine Anwendung war das? _____

Erfahrung im Audibereich

9. Woraus besteht Ihre Musiksammlung?

CD

MP3

DVD-Audio

Schallplatten

Wie viele Musikstücke umfasst diese Sammlung?

<500

500-2000

2000-5000

>5000

10. Hören Sie häufig MP3s?

oft

selten

nie

11. Sind Sie ein Musikkenner (hinsichtlich Künstlernamen, Alben, Titel, Genre)?

Ja

Nein

12. Welche Audiogeräte haben Sie in dem von Ihnen genutzten Fahrzeug?

Radio

DAB

CD-Spieler

DVD-Audio

Speichermedium MP3

Sonstige

Sind Geräte mit Sprachbedienung darunter?

Ja

Nein

Wenn ja, welche Bedienung bevorzugen Sie:

haptisch

sprachlich

13. Hören Sie während der Fahrt häufig Musik?

Ja

Nein

Wenn ja, was bevorzugen Sie dabei:

Radio

eigene Musiksammlung

Wenn ja, hören Sie dabei gerne klassische Musik?

Ja

Nein

14. Passiert es Ihnen oft, dass Sie etwas Bestimmtes/ein bestimmtes Stück hören möchten und nicht wissen, wie es heißt und von wem es gesungen wird?

oft

selten

nie

15. Hören Sie während der Fahrt häufig/gerne Hörbücher?

Ja

Nein

B.1.2 Questions on the experiment***Fragen zum Versuch***

1. Wie angenehm fanden Sie es, wenn Sie etwas frei sagen konnten?

angenehm

neutral

unangenehm

Wie hoch denken Sie, war dabei Ihre Trefferrate?

hoch

mittel

niedrig

2. Wie angenehm fanden Sie es, Ihre eigene Musik in Form von Künstler, Album, Titel etc. zu sprechen?

angenehm

neutral

unangenehm

Wie hoch denken Sie, war dabei Ihre Trefferrate?

hoch

mittel

niedrig

3. Welche Art der Eingabe bevorzugen Sie bei einer Musiksuche?

Eingabe über eine Kategorie

Freie Eingabe

4. Welche alternativen Suchkriterien würden Sie sich wünschen?

5. Könnten Sie sich auch vorstellen, Musikstücke auch auszusuchen einfach nach

der Stimmung

dem Tempo

6. Würden Sie gerne neue Musikstücke angeboten bekommen, die so ähnlich sind wie das aktuell gespielte?

Ja, ich möchte gerne unbekannte Stücke kennen lernen

Gelegentlich möchte ich etwas Abwechslung

Nein, ich möchte selbst bestimmen, was gespielt wird

B.2 Extract of generating rules

Rule description	Example	Regular expression	Sub-match
Category artist			
Nachname bei	Frédéric	$(^{\backslash}\{p\{IsWord\}\{4,\}})$	\$2
Künstlerbezeichnung in der Form <Vorname Nachname> extrahieren	Chopin	$(\{p\{IsWord\}+\$)$	
Nachname bei 2 Vornamen extrahieren	Peter Ilyich Tchaikovsky	$^{\backslash}\{p\{Print\}+\?)\x20(\{p\{Print\}+\?)\x20(\{p\{Print\}+\?)\$$	\$3
Angaben in Klammern weglassen	Natalie Cole (Duet with Nat King Cole)	$(^{\backslash}\{p\{Print\}+\})(\x20*\{[\{1\}\{p\{Print\}+\?}\{1\}\x20*\}(\{p\{Print\}*\$)$	\$1 \$3
The <Künstlerbezeichnung> : Artikel weglassen	The Cure	$The(\x20\{1,\})(\{p\{Print\}+\$)$	\$2
Künstler vor Bindezeichen (&/und/and) extrahieren	Tom Petty & The Heartbreakers - 1993	$^{\backslash}\{p\{Print\}+\?)\x20(?:\x26 \und and)\x20?(\{p\{Print\}+\})\b(?:\x20?[\x2c\x2d\x2f\x20]\{1\}\x20?)\d\{0,4\}$	\$1
Bestandteil vor Trennzeichen extrahieren	Karl Richter; Munich Bach Orchestra	$^{\backslash}\{p\{Print\}+\?)(?:\x20[\x2d\x3a]+\x20 (?<\x20)[\x2c\x3b]+\x20)(\{p\{Print\}+\?)\$$	\$1
Bestandteil nach Trennzeichen extrahieren	Karl Richter; Munich Bach Orchestra	$^{\backslash}\{p\{Print\}+\?)(?:\x20[\x2d\x3a]+\x20 (?<\x20)[\x2c\x3b]+\x20)(\{p\{Print\}+\?)\$$	\$2

Category album

Verwerfen von Suffixen:

- | | | | |
|----------------------------|---|--|-----|
| • Suffix <YYYY>-
<YYYY> | Looking For
The Perfect
Beat 1980-1985 | (^\p{Print}+)\x20[0-9]{4}\x20?-
\x20?[0-9]{4}\$ | \$1 |
|----------------------------|---|--|-----|

- | | | | |
|-----------------------------|--------------------------------------|--|-----|
| • Suffix römische
Zahlen | Queen
Greatest Hits
II | (^\p{Print}+?)\x20(I II III IV V){
1}\$ | \$1 |
|-----------------------------|--------------------------------------|--|-----|

Verwerfen von Präfixen:

- | | | | |
|---|---------------------------------|---|-----|
| Präfix
<Zahl>+ggf.<punkt>+
ggf.<trennzeichen> | 07.- Todo Es
Mentira | ^[0-
9]{1,3}\x2e?\x20*\x2d?\x20*(\p
{Print}+) | \$1 |
|---|---------------------------------|---|-----|

- | | | | |
|---------------------------------------|---|--|-----|
| Präfix (The) very best of
<artist> | The Very Best
Of Burt
Bacharach | ([a-zA-Z0-9]*[-\ \\]*(The)?
Very Best Of\x20(\p{Print}+)\$) | \$3 |
|---------------------------------------|---|--|-----|

- | | | | |
|--|--|--|-----|
| Extrahiert
Zusatzbezeichnung aus
klassischem Titel | Chopin Etudes
(12) For Piano,
Op. 10: No 3 In
E Major, B 74
" Tristesse " | Op(?:\x3a x2e)?\x20*(?:[0-
9]{1,2})[\x3a\x20\x2d]*\x3a?\x2
0?(?:Nr\x2e No\x2e Nr No \x2f)?\
\x20*(?:[0-
9]*)\p{Print}+\ (\p{IsWord}+)" | \$1 |
|--|--|--|-----|

- | | | | |
|-------------------------------------|---|---|---------|
| Verwirft Nummer aus
Beschreibung | Chopin Etudes
(12) For Piano | (^\p{Print}+?)(?:\x20?\x28+[\p{I
sWord}\d]+\x29+\x20?)(\p{Print
}+) | \$1 \$2 |
|-------------------------------------|---|---|---------|

- | | | | |
|------------------------------------|---|--|-----|
| Extrahiert Teil vor
Stücknummer | Beethoven:
Piano Sonatas
#14, 17 & 23 | ^([\x23]+?)\x20?\x23\x20?((?:[\
d\x26\x2c\x20] and und)+)\$ | \$1 |
|------------------------------------|---|--|-----|

- | | | | |
|----------------------------------|------------------------------------|--|-----|
| Künstlerbezeichnung
weglassen | Bach: Air und
Badinerie | ^[^\x3a]*\x20?\x3a*\x20?(\p{Pri
nt}*)\$ | \$1 |
|----------------------------------|------------------------------------|--|-----|

Category title			
Angaben in Klammern weglassen	Natalie Cole (Duet with Nat King Cole)	(^\p{Print}+)(\x20*[\(\[\{ 1 }\p{Pr int}+[\)\]\]}{ 1 }\x20*)(\p{Print}*\$)	\$1 \$3
<i>Verwerfen von Präfixen:</i>			
Präfix direkte Anrede	You Look So Fine	(usted yo ustedes nosotros nosotra s vosotros vosotras Tu Who\x27s Who You)\x20(\p{Print}+)\$	\$2
Präfix 3. Person (zweiteilig)	It's A Wonderful Life	(It's a)\x20(\p{Print}+)\$	\$2
Extrahiert Bestandteil vor Trennwort	Edgar M. Böhlke - Hectors Reise oder die Suche nach dem Glück CD 1	(\p{Print}+\x20\p{Print}+?)\x20(?:oder aus in und)\x20(\p{Print}+ ?)\x20\p{Print}+)	\$1
Extrahiert Klammerbestandteil	Si tu n'étais pas la (Fréhel)	(\p{Print}+?)\x20?\x28\x20?(\p{ Print}+?)\x20?\x29\x20?(\p{Print }*)	\$2
Extrahiert Teil vor Trennzeichen	Carmen - Suite Nr. 1 (Prelude - Aragonaise)	^\p{Print}+?)(?:\x20[\x2d\x3a]+ \x20 (?<\x20)[\x2c\x3b]+\x20)(\ p{Print}+?)\$	\$1
Extrahiert Teil nach Trennzeichen	NDR Info - Zwischen Hamburg und Haiti	^\p{Print}+?)(?:\x20[\x2d\x3a]+ \x20 (?<\x20)[\x2c\x3b]+\x20)(\ p{Print}+?)\$	\$2