Elements of an acoustic anthropomorphisation of an user adaptive Digital Radio DAB

Günther Schatter
Bauhaus-Universität Weimar
Fakultät Medien, Weimar, Germany
guenther.schatter@medien.uni-weimar.de

Sebastian Schmolke
Bauhaus-Universität Weimar
Fakultät Medien, Weimar, Germany
sebastian.schmolke@medien.uni-weimar.de

Benjamin Zeller
Bauhaus-Universität Weimar
Fakultät Medien, Weimar, Germany
benjamin.zeller@medien.uni-weimar.de

Abstract
This paper will introduce the results of the development of a largely acoustically controllable audio system that additionally is capable of answering versatile request by speech synthesis. The speech-dialogue-controlled audio-content-manager was realised on the basis of a DAB-receiver attached to a personal compute via USB including a WLAN-interface. The development of a speech-communication that considers cultural habits in the sense of personification was especially concentrated on.

Keywords
Digital Radio, DAB, DMB, Metadata, Text To Speech, Automatic Speech Recognition, Dialogue System

INTRODUCTION
In the digital era and in the environment of multifaceted cordless systems the Digital Radio is an impeded broadband medium with medium broadband capacity that only contains little radio-specifics any longer and can rather be identified as a slimmed universal receiver without moving images. Despite hybridisations with texts and images (RadioText, Dynamic Label, Visual Radio, Slide Show, PAD etc.) the radio is still an acoustically oriented medium in the first place and is continuously used as such. The efforts to feature the visually displayable data services as the outstanding and commercially conveying characteristics have not been particularly successful during the past developments. Except for a few devices there are no receivers to be found on the market that are capable of graphically displaying data services. Most of the receivers limit the visual presentation to alphanumeric displays using point-matrix-displays with a typical range of 2 by 20 characters. Exceptions are made by a few receivers that utilize the functionalities of computer- or PDA-displays via USB for peripheral output. The available data rates of the Digital Radio are sufficient to transmit acoustic information, texts and as the case may be sequences of still imagery. By bundling channels of separate services an acceptable transmission of moving imagery can be achieved for miniature display via DMB. The strict technical partition of audio broadcasting and television is already increasingly disappearing. With the focus on the hearing sense the radio is able to play to its strengths in situations where visual contact is necessary neither for perception nor for operation. For audio devices an acoustic user interface is coherently obvious.

This paper will demonstrate how to activate the entire range of functionalities contained in an audio-content-manager and DAB services (Electronic Programme Guide, Dynamic Label, Programme Associated Data, Broadcast Website etc.) respectively by using voice-based interaction. The development was focussed on a dialogue system that conforms to the concepts of human-centred dialogues. By employing personified forms of communication conventional approaches of soullessly mechanical dialogue systems should be widely suppressed.

RELATED WORK
The proceedings of speech-processing technologies devices are increasingly enabled to understand voice-based commands and to communicate by using natural language respectively such as dictating machines, information retrieval systems, browser et cetera. For visually impaired persons there is a great many of devices to their disposal that are limited to voice-based output like talking watches, reading machines on the basis of scanners et cetera. Because a lot of information is simultaneously transmitted as text messages in a DAB receiver environment, it is obvious to apply text-to-speech and automatic speech recognition solutions for a communication by speech assistance, because displays are often not appropriate. The benefits of an automatic speech-conversation system have drivers, visual impaired persons or people who possess a receiver without a display. There will be car radio providing local hazard warnings in spoken form as well. Several solutions already exist for “audio-anytime” in car radios (TopNews) and for nowadays wide spread car navigation systems. A simple DAB-receiver is capable of announcing the names of radio stations, the time and running text (Dynamic Label) on the basis of stored speech particles which admittedly leads to spelling in relation with running texts [Pure]. Since a few years speech-dialogue systems do exist to control the information- and entertainment offers of high-class automobile equipments. Besides radio- and CD control mainly telephone-, navigation- and comfort functionalities are offered [Audi07][BMW07]. This paper will introduce ideas on how to extend these audio functionalities shall with new concepts of usage (see Figure 1).

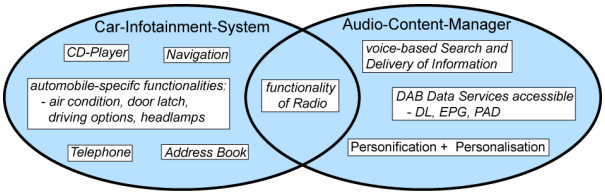

Figure 1: comparison of speech-dialogue systems

Speech-dialogue systems (SDS) are already developed since the 1970ies and are incorporated into more and more devices with the constraint that no security relevant use cases are affected. Hence the domain of AV-technology with its inherent information-and entertainment orientation is well suited for appliances of speech-dialogue systems. First interactive speech-enabled TV-EPGs were developed [Kim03][Shin06]. Very simple and inflexible methods use selection menus. Commonly used but limited command languages require the strict usage of commands. These methods are often considered to be frustrating [Witt06]. More advanced systems use error-tolerant procedures [Berg03]. Another methodology is based on Speech-In-List-Out procedures that are more robust and less burdening to the user [Divi04]. The problems of speech-based communication in relation with mobile devices in rough environments are reported on in [Sawh00]. The emotionalisation of speech-dialogues is further developed in the context of the affective computing domain [Ricc05]. From purpose-oriented monomodal speech-only systems the focus of interest shifts towards multimodal systems [Bern04] [Dyb04]. The advantages of such solutions are based in the two-way support of acoustic and visual information as references and feedback. Many of these developments tend to simulate human behaviour as naturally as possible and use concepts of personification.

PREREQUISITES

Design of the Dialogue

We define additional design goals for speech-dialogue systems that exceed the common factors for general dialogue systems [9241] and identify groups of characteristics that contain specific speech-related requirements.

Technical

Here the efforts are focussed on the computer-based synthesis of natural language. Especially the accuracy of speech recognition and the naturalness of speech output are essentially relevant. The goal is the synthesis of voices that cannot be distinguished from human voices. Additionally the acoustic appearance (energy, melody, pace, expression et cetera) of the generated voice has to be exactly controllable by parameters. Due to own experiences a natural quality of speech output often suggests a would-be intelligence of the speech system that is potentially overestimated and leads to negligence, stress tests and provocations. In the first place speech recognition still simply means the transformation of a signal into a representation adequate for the usage with computers.

Psychological

Here the predominant efforts focus on the naturalistic modality of interaction between speech-based interface and the user. The processes of conducted dialogues are designed to be comparable to interpersonal communication. Furthermore the reactions of the system to actions of the user have to conform to the expectations of human interaction. It is as well necessary to provide clear feedback, reliable help functions as well as efficient error handling. Not at least it is important to incorporate a certain error-tolerance into the system in order to automatically correct possible operating errors of the user.

Logical

The dialogues between systems and users have to be assured a comprehensible structure. The dialogue shall be comparable to human conversations. Correct reasoning of adequate reactions to actions of the user plays an important role. The system analyses available information with regard to reproducible and coherent target states.

Tactical

Here the main subject is to adjust the behaviour of the system to the behaviour of the user. The required analysis of the user's behaviour as well as the current context serves as the basis for conclusions. The aim is to recognize patterns to deduct appropriate reactions of the system that are associated with the behaviour of the user. The system has to be capable of reacting user-adaptive and situation-based. This includes the possibility to have the user define own commands for functionalities.

Emotinal

In order to personalise a system emotionally, human behaviour is simulated. The targeted anthropomorphisation generates reactions of the system that appear to be emotional. The behaviour of the system shall be associated with descriptions of a person's character (friendly, confident, entertaining). Additionally the feedbacks of the system shall not be predictable in order to enhance the impression of the system having an own personality. Synonym commands are facilitated as well.

Analysis of industrial SDS products

The analysis of several automotive speech-dialogue systems clarified that the potentials of DAB service offers is not supported by any of these systems.

The DAB data services such as Electronic Programme Guide(EPG), Dynamic Label (DL) and Programme Associated Data (PAD) can neither be accessed nor be navigated. Likewise contents of both audio- and data services cannot be systematically searched. The designed system will extent the conventional interaction with radio devices by pro-

viding interface specifications to facilitate access to all DAB services. The following functionalities will be realised using a speech-based interface:

- access and navigate DAB data services
- speech output of DAB data services
- interactive search of contents in DAB data services
- automatic search for relevant contents in data services

Speech Interface

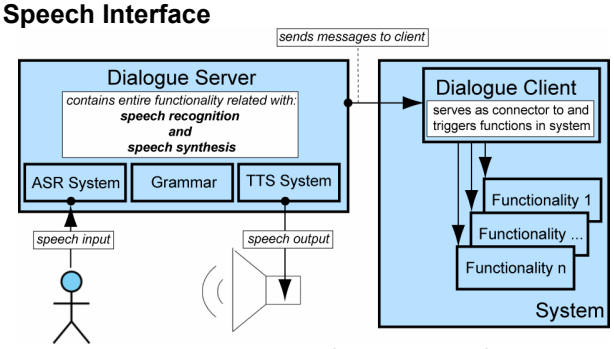

Figure 2: structure of speech interface

The modelling of the dialogue model and the speech processing was realised with the help of the DialogOS software [CLT]. The software provides a graphical interface to modularly construct dialog models based on a XML structure. An integrated speech-processing-engine is available. The structure of DialogOS is based on the Client-Server paradigm. The server incorporates all functionalities related to automatic speech recognition, speech synthesis and dialogue modelling. The actual speech-based application accesses these functionalities by connecting to the server utilising the client (see Figure 2).

We chose DialogOS because it offers a programming inteface for Java, provides a stable speech recognition as well as a very natural speech synthesis.

Text-to-Speech-System - TTS

A TTS-system generates a speech signal synthetically. The system administrates the characteristics of the analysed language internally. During the synthesis machine-readable text is structured into the contained phonemes. Subsequently a signal is generated that comprises the acoustic representation of the extracted phonemes.

Automatic Speech Recognition - ASR

An ASR-system is capable of transforming speech signals into machine- readable text. The frequencies of signal are analysed in order to extract phonemes. The phonemes are compared to the internally administered characteristics of the language. Finally it is decided which textual representation complies with the phoneme.

Hardware & Device Concept

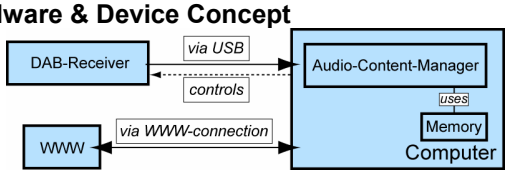

Figure 3: Audio-Content-Manager

The audio-content-manager was developed on the basis of a personal computer with WWW-access and a DAB-receiver connected via USB [Terr]. The device concept is primarily introduced in [Scha06] (see Figure 3).

Anthropomorphisation

The illusion of animacy and autonomy of devices in relation with the development of speech interfaces constitute an equally fascinating and irritating result. Although the target of anthropomorphisation of technical systems is highly controversial we are aiming to facilitate the concept of human-centred dialogues to suppress conventional ideas of soulless machine-like dialogue systems. The development of devices is not supposed to result in human behaviour but shall consider human attributes with the purpose to reduce reservations, rejections and incomprehensibility of devices by incorporating human characteristics.

When a dialogue is designed it is - based on our opinion - not sufficient to merely consider the content-related aspects of communication mentioned above. We propose an extension by additional factors to consider the relationship aspect of interaction with the aim to improve the open-mindedness, the stability of the dialogue, the familiarity and potentially even the joy of usage according the exposure to the system [Watz00]. The fundamental ideas about the relationship aspect are due to the assumption to promote the orientation of device's characteristics towards a stronger partnership-like usage.

DESIGN OF THE SYSTEM

Configuration of the Dialogue

The user is able to access the entire functional range of a digital audio-content-manager by the use of both the speech-interface (Speech User Interface SUI) and the graphical interface (GUI) (see Figure 4).

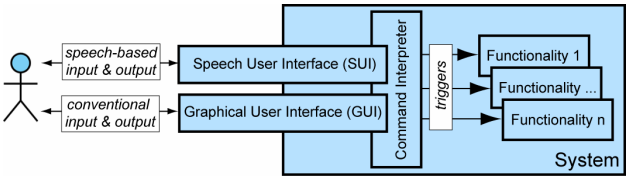

Figure 4: multimodal interaction with system

Firstly memorable terms for standardised control elements for radio devices were determined (see Figure 5). Subsequently the additional DAB service offers were considered. Particular consideration was put on the extraction of information encoded in DAB data services.

The choice of voice-commands was based on the following requirements:

Conciseness
The user should be able to intuitively associate the command to a function. The commands should consist of obvious terms that are directly related to the function.

Briefness
The number of syllables of a voice-command was kept as small as possible. A target size of one to two syllables was considered as ideal.

Unambiguousness
Also the clear acoustic distinction of commands was important. The use of homonymes was strongly avoided.

Memorability
All commands had to be easily memorable in order to enable the user to reliably utilise all commands. The concepts of mnemonics served as a point of reference.

The combination of these four requirements should ensure a reliable navigation of the user within the command structure with a low error rate. The possibility to define alternative voice-commands for functions that could be used synonymously should contribute to this goal as well.

On this basis a hierarchic dialogue model the structure of a tree was developed that constitutes the navigational structure. Each node within the tree structure contains both a function call to control the radio and a voice-command to activate the call. After a voice-input the system is able to map the obtained command unambiguously to a function. The structure consists of three levels. The first level contains representations of the DAB service offers. The second level specifies required functions to control the services offers of the first level. The third level optionally contains additional control functions (see Figure 5).

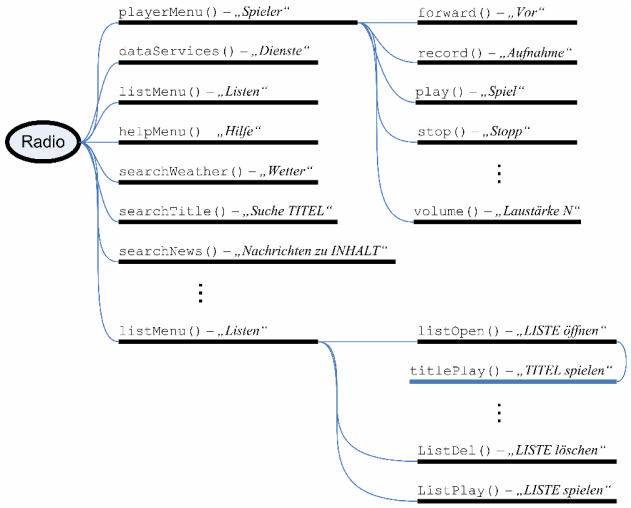

Figure 5: extract of dialogue model (strongly simplified)

To improve the orientation within the structure the user is provided with a context-sensitive acoustic feedback every time a voice-command is directed to the system. During critical situations the systems informs the user what was understood and prompts the user to confirm the command. In addition to that shortcut-commands can be used to access functionalities defined in another area within the navigation structure but without requiring the user to move within the hierarchy. A permanently available help-function offers the user to query the currently available command-options. Additionally features such as abortion of the dialogue, transparent error-handling as well as harmless restart are provided.

Voice-based Search and Response of Information

Besides the radio programmes all available textual information contained in the DAB service offers are made available via speech-output. This information includes ensemble- and service-information as well as all data services like Dynamic Label, PAD and EPG. A systematic search for contents in DAB data services a search query is acoustically verbalised. Provided that stations broadcast relevant newsworthy information textually encoded in DAB data services simultaneously to audio signals the comparatively complex semantic analysis can be neglected and a less complex search in PAD- or BWS-data records may be carried out. By doing so, themes can be subscribed to as well conforming to the idea of On-Air-Podcast-systems [Rotz06]. The voice-command contains a subject matter and content that the user is interested in (see Table 1).

Table 1: search query to data services

user: " Nachrichten zu Doping."
subject matter: "Nachrichten"
content: "Doping"

Subsequently the DAB data services and other available resources (WWW etc.) are analysed for relevant contents. The result of the search consists of consistently structured textual paragraphs. These texts are acoustically output by the system. This functionality enables the user to systematically access the DAB data services (see Figure 1)

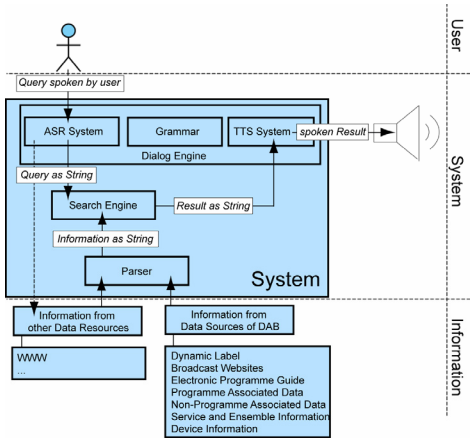

Figure 6: interactive search for contents

This functionality is subject to various potentials related to the organisations of contents of DAB data services. The verbalised search query possibly allow for conclusions to the preferences of a user. These could be used to autonomously search, store and offer contents to a user. To do so concepts have to be developed to manage the refreshing periods of contents in data services.

Anthropomorphisation

We pursue simple but effective approaches to support the creation of a partnership-like relation between users and the system. The following basics serve as a starting point.

Variable Feedback

Each action performed by the user generates an acoustic acknowledgment to confirm the voice-command. For every possible feedback an amount of several responses was set up. A given voice-command is confirmed by randomly choosing a feedback from the predefined amount. This supports the multiplicity of possible progressions of the dialogue. The system seems to be more human compared to monotonously reacting systems. Additionally the choice of the feedback according expression and content of the actual response can be adjusted to the context of the user

Definition of own commands

To avoid the inevitability of using the predefined voice-commands every function can be assigned with an arbitrary amount of individual voice-commands. This facilitates the development of personalised dialogues between user and system and thereby improves the memorability of the commands. Furthermore by using own commands a personal relation between user and radio may emerge since users interact with the system on an individual basis.

Analysis of user behaviour

In order to provide faster access to functionalities often used, recognise typical usage patterns and automatically search for contents the system continuously analyses the actions of users. The system exactly monitors how often which voice-commands are used in which situation. The collected information is used to deduct information relevant for a user in certain situations (see Figure 7).

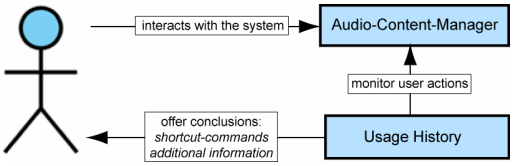

Figure 7: Scheme: Analysis of user behaviour

For example the system determines the frequent use of a specific sequence of voice-commands at a certain time to access some information. As a result the system automatically offers a shortcut-command to the user that enables the direct access of the functionality. In addition the system determined a user always searching for additional information in the DAB data services after listening to sport news. The system offers the user to retrieve this information autonomously and presenting it by speech output. By enabling direct access to functionalities often used and the automatic retrieval of relevant contents the dialogue with the system is exactly adjusted to the users needs.

IMPLEMENTATION

Speech Control

In the server of the speech system a graphical representation of the dialog structure with the form of a graph is constructed. The nodes of the graph offer functions for speech processing and the communication with the client. During an interaction with the system the graph is traversed according to the voice-commands of the user triggering certain functionalities. Depending on the connections between the nodes the user is able to navigate the dialogue model. The vocabulary of the server contains the amount of words that can be detected by the dialogue system. The server administrates a vocabulary by using grammar. The server continuously updates the grammar on a regular basis by analysing the DAB data services adding words not yet contained in the grammar. The server utilises regular expressions to contextually classify speech-input (see Table 2).

Table 2: Indexing of speech commands

speech input: "Nachrichten zu London."
extraction: /Nachrichten zu (.*)/= (extract)
„London“ extracted → parameter for search in.

Afterwards the server sends a message to the client containing the appropriate function call and optionally extracted parameters.

Voice-based Search and Response of Information

After an acoustically verbalised search query the DAB data services are analysed for relevant contents. The subject matter contained in the voice-command as well as the additionally specified content is send as a function-call with parameters to a parser. The parser decomposes the contents of the available data services into consistently structured paragraphs. All paragraphs are analysed with regard to conformances with the desired subject matter and the specified content. The paragraphs that are positively analysed are assembled in a list. The elements of this list are presented to the user by speech synthesis. (See the following examples of the radio station “Deutschlandfunk” from 06/21/2007 18-19 o’clock.)

Table 3: News

user : "Radio"
system: *"Sie wünschen?"*
user: "Presseschau zu Mindestlohn."
system: *"In Presseschau wurden keine Meldungen gefunden, aber in Nachrichten, möchten sie diese Meldungen hören?"*
user: "Ja."
system: *"Streit um Mindestlohn dauert an - Kritik an […]"*

Table 4: Traffic

user: "Radio"
system: *"Ich höre."*
user: "Verkehr A3"
system: *"A3 Köln Richtung Würzburg zwischen Frankfurt [...]"*

Table 5: Weather

user: "Radio"
system: *"Ja, bitte?"*
user: „Wetter“
system: *" Wetter: Regen und Gewitter, Höchstwerte bis 28 [...]"*

Personification

Variable Feedback

In the graph of the dialogue model held in the server of the speech system additional nodes are inserted that contains alternative responses. These nodes are prepended another new node. This node contains the random-function (see Table 6) that chooses the succeeding node. The randomly activated node contains the response to the called function.

Table 6: Pseudocode of random function

```
i = random(1,n);
switch(i) {
        case 1: prompt = "Ja bitte?";
        case 2: prompt = "Sie wünschen?";
        case 3: prompt = "Ich höre?";
        case 4: …
        …}
```

Definition of own commands

A new sub-graph that enables the user to acoustically specify own voice commands was additionally integrated in the server of the speech system. This functionality can be accessed like any other radio-specific functionality by using a voice-command. After a user verbalised the new command the existing dialogue model is automatically extended with the new command. A new node is inserted at the appropriate position in the graph mapping the new command to the desired function.

Analysis of user behaviour

For the analysis of the user's behaviour the MPEG-7 standard for the description of usage histories is employed [15938]. The actions of the user are logged in a machine-readable format. The relevant information about the actions of the user, the point in time and the related content are described. The monitored usage patterns may be utilised for subsequent retrievals of relevant contents.

Furthermore it is monitored which voice-commands are used more frequently than others. As a result the system offers shortcut-commands to the users that can be used to directly access functionality.

CONCLUSION

The extension of a Digital Radio with the possibilities of speech-based interaction potentiates the following improvements of the functionalities of radio devices:

- completely controllable by voice-commands,
- interactive search for contents,
- speech-based output of contents,
- basic personification and personalisation.

The following problems were handled during development:

- find a compromise between extent of possible commands and reasonable complexity,
- quality of speech output,
- dynamic extension of vocabulary in order to provide natural speech-based interaction,
- utilise basics of personification considering the efforts of implementation.

The introduced system realises a hierarchic dialogue model facilitating the full speech-based control of a digital radio. Furthermore the system was extended with the capability to systematically search for contents in DAB data services that are presented to the user by speech synthesis. This functionality was perceived very positively by test users. The possibility to define individual commands to control the radio, the variety of possible feedbacks of the system as well as the analysis of usage behaviours and the automatic search of contents support the development of a partnership-like relation between radio and user.

The combination of these elements to an individual audio-content-manager represents a fundamental modernisation of conventional usage patterns with radio devices. Thereby the development of radio-usage from passive listening towards an interactive and individual dialogue between radio and user is strongly supported. The improved functionalities render the radio to be an appropriate device to satisfy much more multifarious necessities of information than before.

FUTURE CHALLENGES

The targeted convergence of conventional interfaces to a human-centred dialogue system can basically be achieved using two different approaches. On the one hand side there is the intended personalisation of the interface as well as of the offered contents and the adaptation of functionalities to individual concerns. On the other hand there is the attempt to personify the system in the sense of anthropomorphisation. Here continuous efforts are required concerning two subjects. In order to adapt the system to the behavioural patterns of humans the system requires extensive information about the general situation of the user and the context. This information could be acquired using sensors in the environment of the user. The processing of this information and the reliable deduction of situations and coherences as

well as the determination of adequate reactions constitute further requirements.

The second future challenge to personified systems relates to the additional phonetic capabilities of speech-based systems. Although current speech systems already endorse very high quality of natural accentuation, speech rhythm and speech pace the capabilities to have synthetically generated speech carry moods as well are insufficiently supported by speech systems. But an important capability of personified systems is based in the potential to adaptively react on sentiments of users. The reliable analysis of users moods and the adequate reaction constitute the crucial basics for the development of actually personified systems.

A further consequent continuation of the introduced system features the thorough usage of interest profile in the sense of a personal content manager. The personal content manager realises the retrieval and the organisation of relevant information extracted from DAB data services as well as from radio programmes on the basis of analysing user behaviours. An individual selection of radio programmes and contents from data services will be composed and offered to the user. By further usage of the radio the continuous analysis of user behaviours gradually improves the reliability concerning the relevance of extracted contents. The personal content manager supports the development of the radio to a personal media-assistant that monitors vast amounts of information and - aligned with the individual needs of users - stores relevant contents that will be offered to the user.

ACKNOWLEDGEMENTS

The authors wish to thank Daniel Bobbert (CLT) fort he patient support. Furthermore the authors thank Martin Klusmann and Dong Xin for their contributions.

REFERENCES

[9241] EN ISO 9241-110 Ergonomie der Mensch-System-Interaktion. Grundsätze der Dialoggestaltung.

[15938] ISO/IEC 15938-1, -2, -5, -10, -11: Information Technology - Multimedia content description interface. 2005.

[Audi07] Audi AG. Bedienungsanleitung MMI, 2007.

[Berg03] Berglund, A.: Augmenting the Remote Control. Studies in Complex Information Navigation for Digital TV. PhD Thesis. Linköpings Universitet, 2003.

[Bern04] Bernsen, N. O., Dykbjaer, L.: Evaluation of spoken multimodal conversation. Proc. of the 6th Int. Conf. on Multimodal interfaces, 2004.

[BMW07] BMW AG: Betriebsanleitung zum Fahrzeug. 2007.

[CLT] CLT Sprachtechnologie GmbH, DialogOS Version 1.1 Releasedate 12.07.2007, Saarbrücken, 2007

[Divi04] Divi, V. u.a.: A Speech-In List-Out Approach to Spoken User Interfaces. TR2004-023. Mitsubishi, 2004.

[Dyb04] Dybkjær L., Bernsen N. O., Minker W. (2004): Evaluation And Usability Of Multimodal Spoken Language Dialogue Systems. In: Speech Communication, Vol.43, Issues 1-2, June 2004, Elsevier. S. 33- 54.

[Kim03] Kim, H.; Hwang, E.: VoiceEPG: Speech Interface for Electronic Program Guide. In: Proc. of the IASTED Conf. on Internet and Multimedia Systems and Applications. Honolulu, Hawai, August 14-16, 2003.

[Pure] Sonus-1, www.pure.com

[Ricc05] Riccardi, G.; Hakkani-Tür, D.: Grounding Emotions in Human-Machine Conversational Systems. In: Intelligent Technologies for Interactive Entertainment. Springer, 2005.

[Rotz06] Rotzoll, C.; Schatter, G.; Toennies, H. Ch.: Crossreferencing of EPG and RSS Metadata. An Approach for a Broadcast-Podcast Hybridisation. Workshop 2006 Erlangen

[Sawh00] Sawhnwy, N.; Schmandt, C.: Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments. ACM Transactions on Computer-Human Interaction, Vol. 7, No. 3, September 2000, Pages 353–383.

[Scha06] Schatter, G.; Bräutigam, C.; Neumann, M.: Personal Digital Audio Recording via DAB. Enhanced Radio as Interface. 7th Workshop Digital Broadcasting. Fraunhofer IIS Erlangen, 2006.

[Shin06] Shinjo, H. et. al.: Intelligent User Interface based on Multimodal Dialog Control for Audio-visual systems. Hitachi Review March 2006.

[Watz00] Watzlawick, P.; Beavin, J. H.; Jackson, D. D.: Pragmatics of Human Communication A Study of Interactional Patterns, Pathologies, and Paradoxes. W. W. Norton & Company, Incorporated, 1987.

[Witt06] Wittenburg, K. et al.: The prospects for unrestricted speech input for TV content search. In: Proc. of the Working Conf. on Advanced Visual interfaces AVI '06. Venezia, Italy, May 23-26, 2006.