An User-Centered Development of an Intuitive Dialog Control for Speech-Controlled Music Selection in Cars

Stefan Schulz, Hilko Donker

Cooperative Multimedia Applications
Faculty of Computer Science
Dresden University of Technology
Dresden – Germany
{stefan.schulz|hilko.donker}@inf.tu-dresden.de

Abstract

During the last years speech dialog systems have proven valuable for controlling infotainment systems in cars. The increasing popularity of mobile MP3 players and their capability to store more and more MP3, suggest the development of a speech control function for car MP3 players. With technical advantages in hardware and speech recognition it is possible to focus on the development of an intuitive speech dialog, which ensures that the advantages of speech control can be utilized while disadvantages are minimal at the same time.
This paper describes the user-centered development of such an intuitive dialog, its special challenges and the reached quality in usability and ease of use. Also the generalizability of the results is discussed.
Index Terms: usability, dialog design, mp3, user-centrered development

1. Introduction

This work is based on [1], where the feasibility of a product-ready speech-controlled music selection for embedded systems was shown, but the dialog design only based on a questionnaire, general design guidelines and the designer's intuition and the resulting prototype was not evaluated in a user study. In contrast, this work only focuses on the usability of such a system.[1]
Especially in the highly controversial topic of music selection (see section music selection) a complete user-centered development incooperating users expectations on in-car music selection from the beginning seemed to be more promising.
In [2] an iterative development life cycle was suggested for the user-driven development of speech dialogs. Following this, we applied three stages of users participation (questionnaire, simulation, evaluation) during this project. The questionnaire helped to ascertain information about potential users as well as about intuitive associated functions which helped to refine the ideas on a possible dialog design. Based on this a simulation using the wizard-of-oz-technique could be hold to test users intuitive wording, strategies and general approach concerning music selection in a realistic environment.
With the knowledge obtained from this a suitable concept for a music selection dialog could been defined and implemented within the prototype 'Dorothy'. In the discussion of the evaluation of this prototype an investigation of the user expectations and the reached quality of the dialog is included, as well as future work on the topic is discussed.

[1]The work reported in this paper was done within a master's thesis carried out in cooperation between the Dresden University of Technology and Harman/Becker Automotive Systems Ulm.

2. Related Work

Speech enabled music selection in cars is not available yet, but there have been several concept studies on this topic. However the generalizability of the results of these concept studies are either restricted by the use of limited hardware and speech recognition [3], by not considering computing capabilities [4] or only by examining isolated querys and using special speech recognition components, which cannot be used with standard speech dialog systems in cars [5].
But this last work provided the idea to focus on an easy and lightweight selection ('google approach'), rather than too much on exact matches. The importance of smart selection was also highlighted by [6] in their work on a music distribution system. They assumed further, that there are different use-cases and users for such systems, so that several ways to access music should be considered.
The earlier mentioned basis for this paper [1] contributed not only in terms of feasibility of a product-ready speech-controlled music selection for embedded systems, it also introduced some dialog concepts (play/browse-modus), which were taken as a starting point in this development.

3. Music Selection

Music selection is viewed and used differently through various users or use-cases. This originates from various definitions of what (good) music is and how it can be structured and accessed.[2] Two main approaches could been distinguished in further discussion, at one side the hierarchical and on the other side the search-based method. A good music selection dialog should integrate both approaches.
When accessing music hierarchically, the user has a clear understanding of what he wants to select and where in a given hierarchy it could be found. This is the classic approach most MP3-players including the iPod are using today.
But in other cases, the user might feel like searching or explor-

[2]Even in the development team of this project it was impossible to agree on one consensus.

ing his library, by only using keywords to converge to music he would like to hear. Therefore it can be imagined that additional to keyword search a user could principally use query-by-humming [7], community-based-filtering [8] or even mood-based-selection via bio-sensors [9]. But for that you would need special narrative meta-data, which is not widely available and its automatic generation is not trivial. Because of that it was decided to use only the descriptive meta-data of ID3-tags, which are a quasi-standard for music meta-data and can contain nearly any amount of meta information imaginable [10]. But in most cases, only a small subset of such meta-data is used and useful for music selection. When using this ID3-tags accessing music via speech (which was possible as shown in [1]) also new questions arise, because often more than one possibility of system reaction is possible. To give an example, if the user simply requests 'artist madonna', the system reaction (later referred as 'play/browse-hierarchy') can vary from directly playing the first or a random madonna song to offer a list of madonna titles or albums. Finding a well-accepted subset of meta-data used for music selection as well as an intuitive play/browse-hierarchy were a main challenge of this work, because this selection covers the main majority of the dialog structure.

4. Questionnaire

For obtaining information about potential user groups, their MP3 usage and expectations toward a perfect music selection, a questionnaire was hold at the beginning of the development. It had the form of a 3-page paper questionnaire, and its 227 participants were mainly young male german students. Beside some results on the users MP3 usage (music in car is literally 'always on', library size typical around 10 GB), information about expectations toward a perfect music selection could be gained. For isolating important features of such a music-player, people were asked in a open question which function they would like to control by voice the most and later to rate a list of possible features. In the open question three main functions were requested: playback controls ('Play', 'Pause'; 'Next', 'Skip'), direct selection title/album and volume/equalizer control.[3] In the list of possible features, these functions were again rated good, but play modes (shuffle, repeat) and load playlist were also considered as important functions. From the different importance of sorting criteria for the whole library (more descriptive tags) and for playlists (more narrative) it was assumed, that there are in fact several use cases for accessing music, which should be considered by dialog design.

5. Wizard-Of-Oz-Test

The goal of the simulation was to gain additional information about the wording, an intuitive structure and the general approach to speech controlled music selection as they are expected from a users perspective. For achieving this it was favorable to use the wizard-of-oz method (WOZ), and conduct the experiment in a very open structure. The used wizard-of-oz-tool [11] enabled such a form of experiment. For a more realistic impression of users interaction with a real system, a 7 inch display (as normally used in cars), a video camera for subsequent analysis of the wording and a complete driving simulation [12] was included in the test setting,

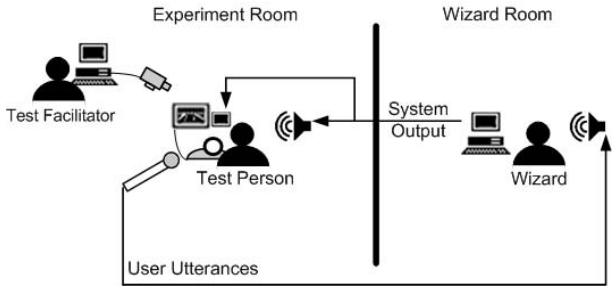

Figure 1: *WOZ test setting.*

which is shown in Figure 1.
In the process of designing tasks for this test some issues were isolated, which cannot be modeled into convincing questions. Especially the questions about when to play and when to further select after the selection of one tag (play/browse-hierarchy) needed a different approach to collect data. For this, a multimodal presentation of the possible system reactions was implemented using MS Powerpoint, which was presented to the users to help them answering to a questionnaire, which reaction they prefer most.
The test finally was hold with 20 participants, all students with barely no experience with speech dialog systems. The results show the correctness of this approach, both wizard-of-oz-test and Powerpoint-survey produced valuable results. While the WOZ-test provided information about wording, driving distraction and the integration of playlist-features, the Powerpoint-survey enabled the definition of intuitive play/browse-hierarchy.
The results for the play-browse hierarchy showed that album remains the most important ordering criteria for music. People wanted to play immediately album or title, but in the case of a selected title the associated album should be in the playlist. For artist-selection, selecting a album from this artist was the preferred choice. For genre, considered as narrative criteria before, there was no clear tendency, if to play or to select something after use of the genre information. In a little in-house survey browsing artists was preferred, qualifying it as a more descriptive criteria.[4]
When asking the users for best behavior for year-selection, about 30% rejected completely the use of it ('would not use that'), so the year-tag was not considered anymore. For ambiguous input, system-driven decisions were not favored, while a system prompt asking the user for clarification was preferred.
In the WOZ-test itself, the separation between a play and a browse command was not accepted by the users, they expected one consistent behavior when for example selecting a artist. So the idea of separate play and browse modes (which originated from [1]) was dropped. Also the explicit play-commando after each selection (which was introduced to enable playlist-management functions, but which users were not really interested in) was rejected, the direct play feature was most requested in this tests.
Also the display content was rated highly distractive, users requested more direct voice feedback. Apart from this, the overall rating of the system was good, but often together with the comment, that the users liked 'the perfect speech recognition'.

[3] As the first two function were expected before, the last one was really surprising. As it was unclear how to implement such a volume control function, which can add value to normal haptic control, this function was not considered this time.

[4] If using genre in more general terms, like "relaxing music" or "with guitars", results may vary, but with only id3 metadata available, such functions could not offered here.

```
usr:  singer madonna.
sys shows album list from Madonna
sys:  artist Madonna.
sys:  album 1.
sys plays sound sample from album 1.
sys:  album 2.
sys plays sound sample from album 2.
usr <PTT>
sys *BEEP*
usr:  play.
sys plays album 2.
```

Figure 2: Scan mode example

6. Prototype "Dorothy" And Evaluation

The results from the questionnaire and the WOZ-test made it possible to formulate system development guidelines for the further development.

The system should:

- be operated as far as possible without a display,
- use an intuitive play/browse hierarchy (WOZ-results),
- play music as early as possible in the interaction,
- use real speech recognition for evaluation.

These guidelines as well as many other small facts about wording and strategies, served as the requirements for the development of the prototype. The prototype was implemented using the Harman/Becker dialog development toolkit [13]. These tools are product-level tools and together with the used StarRec speech recognizer are used for product systems by Harman/Becker.
For archiving more display independence the use of lists in the interaction was questioned. As presenting lists using speech only is generally problematic, a new way to make the selection experiencable was searched. A scan mode was defined, where all selectable music is played one after another to the user. By that, the concept of lists, which is know from normal GUI-interaction, could still be used. But in the same time more information is available than a display could offer, without even look at a display (see an example in Figure 2). For testing reasons, a read-out mode (where all list entries were read) and a read mode (without presenting anything about the list items) were implemented and tested against each other.
For the play/browser-hierarchy discussed earlier the results for the WOZ were applied and the combination is shown in Figure 3. In this illustration the blue boxes describe auditive system outputs, while the white ones show the display content. In the cases of a selected title, genre, playlist or a random play (something), the system would turn directly into play mode. In contrast, when genre, artists or the concept only (with utterances like 'show albums' or 'all artists') are selected, the user is asked to first select from a list. From this illustration can be derived, that every selection can be finished after a maximum of four utterances, but more likely after one or two. This helps to play music early in the interaction, as requested in the discussed guidelines.
Beside list mode and tag selection, the prototype featured functions of play modes (shuffle, repeat), play controls (play, pause, next, etc.) and some global commands for information and help. A screen shot of the GUI of the prototype is shown in Figure 4.

Figure 4: *Screen shot of the GUI of prototype 'Dorothy'.*

The evaluation of this prototype was designed to provide information about the reached quality of the dialog design in general. To archive that, it was necessary to test the assumptions on the dialog structure (play/browse-hierarchy), wording and list modes, which were made after the first user tests. For a broader generalizability of this tests, it was decided that the tests will not be run with students again, but rather with 24 co-workers from Harman/Becker (mainly between 26 and 40 years, 58% female), who are not working on speech dialog systems. Also, as mentioned before, a real speech recognition component was used to ensure that the impression of the dialog is not dominated by the 'spectacular speech recognition' of a WOZ-experiment.

The results showed that the play/browse-hierarchy was accepted for music selection via all kinds of tags. Only in 4% of the cases the subjects expected a different system reaction. The display dependence was effectivley reduced, the percentage of users, which felt distracted by the display dropped from 2/3 (WOZ) to 1/3. The list modes, which are probably responsible for this reduction to a great extent, were rated controversial. While the read-out mode was favored by a majority of users, both the scan-mode and the read-mode had only a few vehement fans. Finally, the general system satisfaction was high, the objective ratings turned out to be still good, even though real speech recognition was used.

7. Conclusion

The discussed results of the final evaluation underlined the correctness of the taken approach of user-driven development of the speech dialog. At the end all wanted functions were included in an intuitive dialog structure, which the users accepted. To this end only standard-components and code were used. Starting with a very open structure, a directly applicable dialog for selecting music could be defined.

This intuitive way of selecting music is believed to be valid for other use-cases of music selection, where the user should not be forced to use a display, e.g. in smart-home or smart-clothes scenarios.
For automotive applications other than music selection, the reduction of the display dependence should also be an important objective. As in the music selection use-case the use of fewer lists and more feedback is believed to help there.

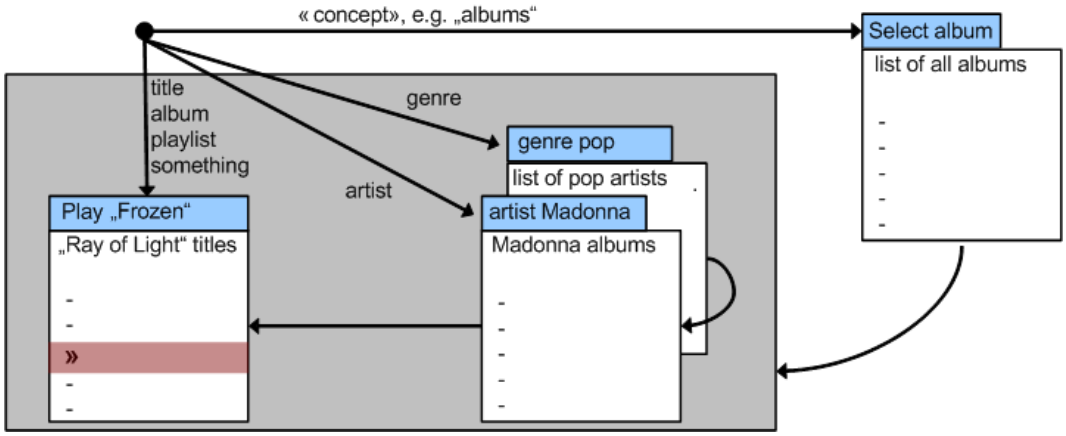

Figure 3: *Tag hierarchy used in prototype 'Dorothy'.*

8. Future Work

The dialog structure developed for this work is in German, the application to other languages and countries should be evaluated in a next step. Also, the use of narrative meta-data and their impact on the dialog structure should be subject of further investigation.

9. Acknowledgments

As this work was done in close cooperation with Harman/Becker Automotive Systems in Ulm, Germany, we would like to thank the whole 'dialog research and tools' team there, with special thanks to Stefan W. Hamerich and Patrick Langer, which were irreplaceable for the success of this project.

10. References

[1] Y.-F. H. Wang, S. W. Hamerich, M. E. Hennecke, and V. M. Schubert, "Speech-controlled Media File Selection on Embedded Systems," in *Proceedings SIGdial*, Lisbon, Portugal, 2005, pp. 217–221.

[2] N. O. Bernsen, H. Dybkjær, and L. Dybkjær, *Designing Interactive Speech Systems - From First Ideas to User Testing*, Springer, London, Great Britain, 1998.

[3] G. McGlaun, F. Althoff, H.-W. Rühl, M. Alger, and M. Lang, "A Generic Operation Concept for an Ergonomic Speech MMI under Fixed Constraints in the Automotive Environment," in *Proc. HCI*, 2001.

[4] R. Pieraccini, K. Dayanidhi, J. Bloom, J.-G. Dahan, M. Phillips, B. R. Goodman, and K. V. Prasad, "A Multimodal Conversational Interface for a Concept Vehicle," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 2233–2236.

[5] C. Forlines, B. Schmidt-Nielsen, B. Raj, K. Wittenburg, and P. Wolf, "A Comparison Between Spoken Queries and Menu-based Interfaces for In-Car Digital Music Selection," in *Proc. HCI*, 2005.

[6] F. Pachet, A. La Burthe, A. Zils, and J.-J. Aucouturier, "Popular music access: The Sony music browser," *Journal of the American Society for Information Science Technology*, vol. 55, no. 12, pp. 1037–1044, 2004.

[7] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," in *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 231–236.

[8] "Last.fm," - URL: http://www.last.fm/, last access: April 2006.

[9] Thomas Rist, "Affekt und physiologische verfassung als parameter für hochgradig personalisierte fahrerassistenzdienste," in *Proceedings Workshop Automobile Cockpits und HMI*, 2004.

[10] M. Nilsson, "ID3 Tag Version 2.4.0 – Main Structure," Informal Standard - URL: http://www.id3.org/id3v2.3.0.html, 2000.

[11] S. Petrik, Y.-F. H. Wang, and S.W. Hamerich, "Hierarchical Dialogue Structure Representation for Wizard of Oz Experiments on Speech Dialogue Systems," in *Proceedings of the 10th Conference Speech and Computers (SPECOM)*, Patras, Greece, 2005, pp. 207–210.

[12] S. Mattes, "The Lane Change Task as a Tool for Driver Distraction Evaluation," in *Proceedings of the ISOES*, Munich, Germany, 2003, pp. 57–60.

[13] S. W. Hamerich and G. Hanrieder, "Modelling generic dialog applications for embedded systems," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2004.