A MULTIMODAL DIALOGUE SYSTEM FOR INTERACTING WITH LARGE AUDIO DATABASES IN THE CAR

Sandra Mann, André Berton and Ute Ehrlich

DaimlerChrysler AG, Group Research and Advanced Engineering
P.O. Box 2360, 89013 Ulm
Sandra.Mann@daimlerchrysler.com

Abstract: The variety and complexity of audio storage devices nowadays available in automotive systems turn selecting audio data into a cumbersome task. How can users comfortably access particular audio data in a variety of media carriers containing large amounts of audio data while actually pursuing the driving task? Browsing logical and physical hierarchies can heavily increase driver distraction. This paper proposes a speech-based approach that significantly facilitates accessing database items. It consists of two interaction concepts: the concept for category-based search requiring pre-selecting a category (e.g. artist, title, genre, etc.), and the concept for category-free search allowing the user to search globally across all categories. In both concepts search space comprises all media carriers. The user may directly address audio data items by saying the corresponding name. However, evidence was taken from address book data that users do not perform well when it comes to remembering the exact name of an item. We therefore create additional wording variants to allow for fault-tolerant search. This is achieved by means of filter- and recombination rules. Thus users, who do not remember an item correctly are still able to find it when speaking only parts thereof.

1 Introduction

State-of-the-art speech dialogue systems in cars provide assistance in operating audio devices, such as radio and CD player. However, the number and variety of media carriers (e.g. hard disk, MP3 player, memory cards) containing compressed audio data has increased significantly. These media carriers allow for integrating large databases of music titles into the car. The large databases in turn aggravate the user's possibility of accessing the right item, for how should she be able to correctly remember hundreds, or even thousands of songs, let alone attributing the corresponding media carrier to the desired data.

Considering the growing amount of media data, current methods of navigating them 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 to meet customer demands. We present an approach that offers a more user-friendly way of interacting with audio data, in particular when the user's inability to remember large amounts of data is taken into account.

Various approaches for speech-based access to audio data have already been published. One approach comprises accessing every database item within one utterance [1]. 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 [2]. 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 we aim at handling large vocabulary lists

by means of category-based and category-free input of items. Additional wording variants, generated by means of filter and recombination 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.

The next section elaborates on problems coming along with accessing large amounts of audio data. Evidence is taken from users' recollection of personal address book data. The new multimodal dialogue design for efficiently interacting with complex audio applications is described in section 2. Section 3 presents the multimodal system architecture required for the new approach. The final section draws conclusions and provides an outlook on future work.

1.1 Technical and user constraints

In-car speech dialogue systems for audio applications nowadays provide the user with a considerable 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) and 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 [3]. Given the above diversity and complexity this procedure is likely to prove cumbersome, increasing cognitive load.

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 a user study on personal address book data [4] analysing to what extent users remember the entries they have stored in their address book (see Figure 1).

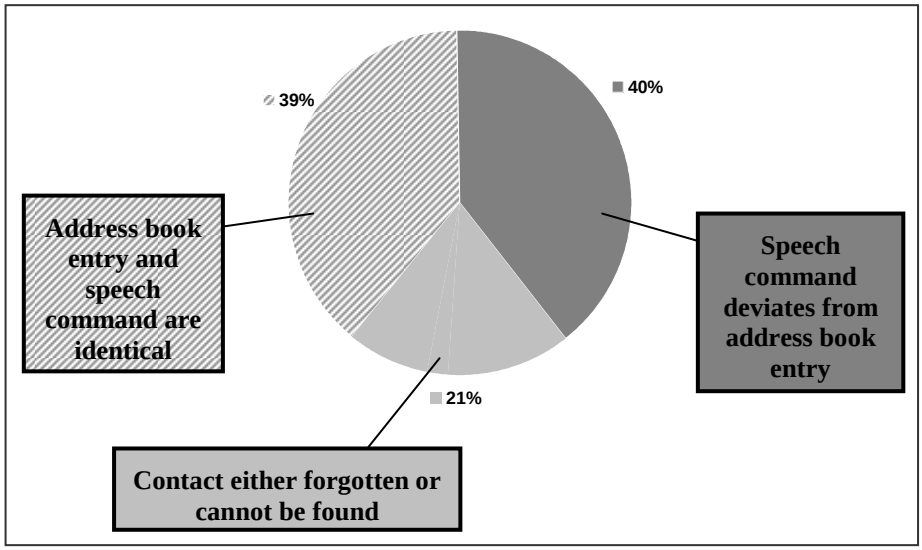

Figure 1 – User recollection rate of address book data

From a set of 21 subjects electronic address books (i.e. mobile phones and PDAs) were read out and by means of 420 scenarios on different contact groups (e.g. business, private, etc.) the users' speech commands were recorded and subsequently compared to the stored data. The findings 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. 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 certain 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.

2 Dialogue design

The previous section 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. We propose a new method for handling increased functionality as well as large amounts of audio data. The next section explains our search concepts for accessing audio data. The concepts use additional wording-variants to allow for fault-tolerant search described in section 2.2. Section 2.3 presents general requirements concerning the user interface.

2.1 Interaction concepts for searching audio data

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 we suggest three different interaction concepts: category-based search, category-free search and physical search.

Category-based search requires pre-selecting a category. We defined a set of five selectable categories (artist, album, title, genre and year) that are 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.

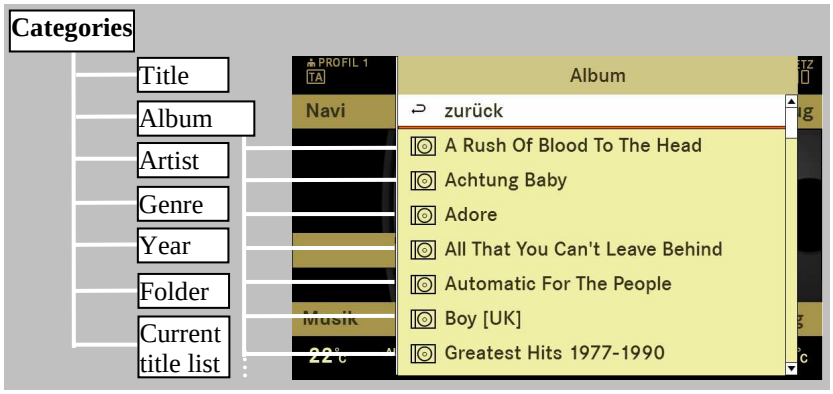

Figure 2 – Category-based search

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 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. section 2.2). We allow speaking 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 she 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 dissolved 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 3 (left)).

Figure 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’ – we add a supplementary step providing the user with a list of the corresponding categories plus the respective number of hits (see Figure 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.

2.2 Interaction concept for fault-tolerant word-based search

Section 1.1 presented the difficulties users have in precisely recollecting large amounts of data. 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. Our 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 two procedures, i.e. filter and recombination rules [5]. In the first step our method decomposes items of all categories according to the following rules:

1. **Special characters** such as separators and symbols are either discarded or converted into orthography.
   - Africa / Brass → Africa Brass
   - The Mamas **&** The Papas → The Mamas and The Papas

2. Abbreviations are written out orthographically.

   - *Dr*. Dre → Doctor Dre
   - Madonna *feat.* 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 *Band* → Charlie Daniels (plus rule 4)
- *Songs* 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).

- *The* Lemonheads → Lemonheads
- *They* Might Be Giants → Might Be Giants
- *Under* Pressure → Pressure

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

- Ludwig *van* Beethoven → Beethoven
- *Dave* Matthews Band → Matthews Band
- Looking for the Perfect Beat → For The Perfect Beat
  → Perfect Beat (plus rule 4)

In the second step linguistic rules are used to select significant components and recombine them in a way a user might say. This avoids random and useless variants. Each recombined component sequence 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 re-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.

2.3 General requirements for the user interface

In addition to the interaction concepts on large audio data presented in sections 2.1 and 2.2 the user interface 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. section 3) 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 we allow for spoken input that is less restricted. Rather than demanding from the user to learn a multitude of speech commands we offer a variety of expressions covering the same meaning (synonyms). In case the user has forgotten a particular expression, she may simply pick an alternative instead of gazing 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 her 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 so-called shortcuts. Expressions such as “search album”, “search category artist” or “play music” are straight-forward, preventing her from numerous steps through a menu hierarchy as is inevitable when using manual interaction.

For both category-based and category-free search we allow for a first combination of the categories genre and year within one utterance (e.g. play Rock ‘n’ Roll from the 70s).

GUI and SDS resemble the Mercedes-Benz design guidelines.

3 Prototype architecture

The new approach of accessing media data by speech was verified with a prototype system. The prototype’s architecture is based on state-of-the-art speech dialogue systems [6] connecting to media search engine of the media application (see Figure 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. Our dialogue system is a multimodal interface [7] with two input and two output modalities: manual and speech input, and graphical and speech output.

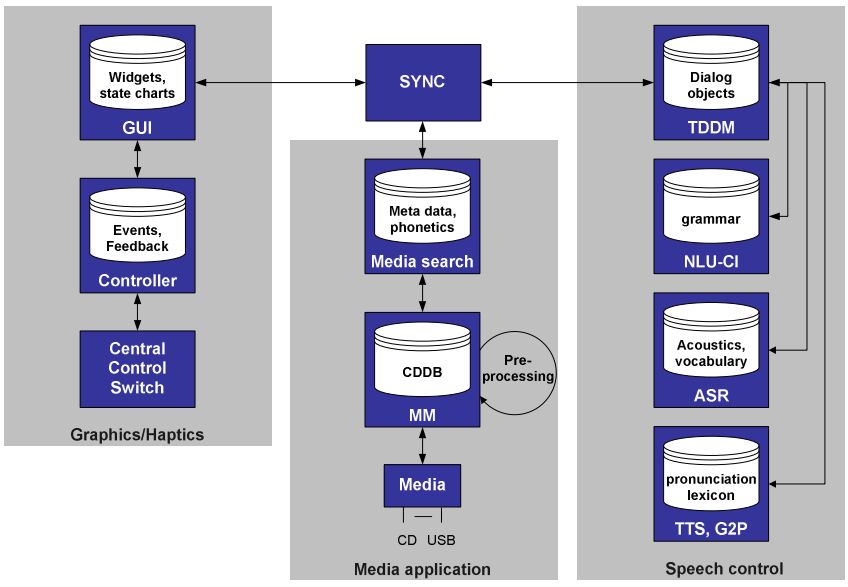

Figure 4 - Prototype architecture view

Speech control consists of a task-driven dialogue manager (TDDM), a natural language understanding unit (NLU) containing a contextual interpretation, an automatic speech recogniser (ASR) and a text-to-speech component, which includes a grapheme-to-phoneme (G2P) converter. All speech control modules are subject to the configuration of a common knowledge base [8].

The *graphics-haptics interface* follows the mode-view-controller paradigm [9]. Models (state charts) and views (widgets) are described in the 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.

The *media application* consists of the media search engine and the media manager (MM). The latter administrates the connected media. We consider 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. Not 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 [10]. 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 we follow a two-stage algorithm:

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, we have 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. We provide phonetic transcriptions for all three languages using the corresponding G2P. 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.

4 Conclusion

The approach presented in this paper 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 filtering and recombining 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.

Future work focuses on testing and evaluating to what extent this approach performs better. It would also be interesting to analyse in how far the pre-processed metadata (wording variants) cover what users input via spoken interaction when searching for audio data.

An add-on to the above prototype envisages extending the number of voice input parameters within one utterance to allow for straight-forward combination of categories, like for example ‘Shakira Objection’.

We also consider integrating text-based language identification for generating phonetic transcriptions for metadata of songs 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.

5 Acknowledgements

This work was partially funded by the German Ministry of Education and Research BMBF in the framework of the SmartWeb project.

References

[1] Wang, Y., Hamerich, S., Hennecke, M. and Schubert, V.: Speech-controlled Media File Selection on Embedded Systems, SIGdial Workshop, Lisbon, Portugal, 2005.
[2] EU project TALK: Talk and Look: Tools for Ambient Linguistic Knowledge, www.talk-project.org
[3] Mann, S., Berton, A. and Ehrlich, U.: How to Access Audio Files of Large Data Bases Using In-car Speech Dialogue Systems, Interspeech, Antwerp, Belgium, 2007.
[4] Enigk, H.; Meyer zu Kniendorf, C.: DaimlerChrysler Internal Study: Akzeptanz von Sprachbediensystemen im PKW – Anforderungsanalyse zur Struktur und Nutzung von Adressbuchdaten. DaimlerChrysler AG, Berlin, 2004.
[5] Berton, A., Ehrlich, U. and Mann, S.: Aufarbeitung von Aussprachevarianten für Text-enrolments von Sprachbediensystemen. Patent application filed, 2007.
[6] McTear, M.F.: Spoken Dialog Technology – toward the conversational user interface, Springer, London, 2004.
[7] Gibbon, D., Mertins, I. and Moore, R.K. (Eds): Handbook of Multimodal and Spoken Dialogue Systems, Kluwer Academic Publishers, Norwell, Massachusetts, USA, 2000.
[8] Jersak, T. and Ehrlich, U.: Definition und Konfiguration der Wissensbasen und Schnittstellen von Sprachdialogapplikationen mit XML, XML Tage, Berlin, 2006.
[9] Reenskaug, T.: Thing-Model-View-Editor - an Example from a Planningsystem, Technical Note, Xerox PARC, 1979. http://heim.ifi.uio.no/~trygver/mvc/index.html
[10] Gracenote MediaVOCS: www.gracenote.com/gn_products/mediaVOCS.html