

# Speech-controlled Media File Selection on Embedded Systems

Yu-Fang H. Wang, Stefan W. Hamerich, Marcus E. Hennecke, Volker M. Schubert

Temic SDS GmbH  
Speech Dialog Systems  
Ulm – Germany

{helena.wang|stefan.hamerich|marcus.hennecke|volker.schubert}@temic-sds.com

## Abstract

We present a speech-controllable MP3 player for embedded systems. In addition to basic commands such as "next" or "repeat" one main feature of the system is the selection of titles, artists, albums, genres, or composers by speech. We will describe the implemented dialog and discuss challenges for a real-world application. The findings and considerations of the paper easily extend to general audio media.

## 1 Introduction

Temic SDS is a leading manufacturer of speech control systems for the automotive market. The first product using Temic's technology was the Linguatronic system available in Mercedes cars since 1996, for further details refer to (Heisterkamp, 2001). This first system allowed the hands-free usage of the inbuilt car phone and was already completely speaker independent. The vocabulary consisted of about 30 words and allowed continuous recognition of e.g. digit strings. Since then these systems have become more and more complex. Voice control of the audio system is standard and recent systems even feature navigation destination input with voice.

Nevertheless, direct selection of audio tracks by name is not yet offered in a mobile product. Current audio systems with speech input only allow general commands such as "next title" or "previous CD". With compressed formats such as MP3 gaining strong footholds in the mobile sector, the number of titles that can be stored on such devices increases rapidly and this 'classical' speech interface quickly becomes unwieldy. In contrast, direct selection of audio media naming title, artist, or album offers an attractive and intuitive way to navigate a mobile music collection.

Since 2002 Temic SDS has been part of Harman/Becker Automotive Systems and Harman International. This provides the context for our work with voice-controlled audio systems.

In this paper we will describe a speech dialog system for selecting audio media. The paper will focus on the dialog but will also briefly discuss the challenges of such a system.

## 2 Embedded Speech Dialog Systems

Car-based speech dialog systems are available since 1996. These systems in cars are embedded solutions, provided either as a separate hardware box or integrated into infotainment systems running under a variety of real-time operating systems. Typically, such embedded systems are characterized by limited resources. To be able to run speech dialog systems on environments offering a total memory of as little as 128 KB, special algorithms and tools are needed, see e.g. (Hamerich and Hanrieder, 2004).

Currently, applications based on Temic technology are available with several car manufacturers, among them Audi, BMW, Lancia, Mercedes-Benz, Porsche, Rolls-Royce, and also in the after market. All of these systems are still closed applications whose primary task is to control on-board devices (telephone, tuner, CD player, navigation system, etc.). That is, there is no need to access data dynamically. Even for navigation destination input the domain and the vocabulary is known in advance.

## 3 Related Work

Speech enabled media selection in embedded systems is not available yet. Several concept studies have been made, see e.g. (Pieraccini et al., 2003), but no product is available at the moment.

On the other hand, projects are ongoing to allow speech-based media selection on server-based systems, e.g. refer to (Baumann and Klüter, 2002; Schulz et al., 2004). Here the challenges for speech recognizers are nearly the same as in embedded devices, without having the limitation of computing power and memory.

## 4 System Description

Since portable media players get more and more common, users wish to control these devices by speech as well. Such systems can be available nearly everywhere, carried by the user or integrated in an automotive audio system. In all cases, a huge collection of audio files can be found on such devices. This could for example be a mobile MP3 player, a USB memory stick, a memory card, or the hard disk inside a car audio system or head unit. Several such media can be available in the system. Generally, it is required that certain meta information such as name of the title, the artist, and the composer are available. In case of MP3 files these could be the ID3 tags that such files typically contain. If files do not come with meta information (e.g. titles from a CD) such information can be looked up in a database such as the Gracenote CDDB.

Our goal is to be able to select subsets from the entire music collection by voice. For example it should be possible to play songs from a certain artist or all titles of a particular album or music of a certain genre. Towards this end the system keeps a database of all available media on all connected devices with corresponding meta information. The information in the database is used to dynamically configure the speech recognizer.

The main features of the system are the intuitive dialog which attempts to minimize the steps necessary to accomplish a selection, the dynamic enrollment<sup>1</sup> of MP3 tags into the recognizer grammar, the possibility to attach external USB memory devices (including an iPod), and a special "more like this" feature which allows the user to select songs similar to the one currently playing.

The system runs on a PC simulation as well as on a 200 MHz SH4 embedded platform.

## 5 Dialog Design

MP3 players have become very popular and are available both in software (e.g., Windows Media Player, iTunes) as well as in hardware (e.g., iPod). The iPod in particular features an innovative user interface which allows the user to quickly navigate the music collection.

Similar innovative interfaces for speech are not yet available, however. It is not clear what an intuitive speech interface would look like, nor what users would expect from it. Therefore, as a first step, a questionnaire was set up in order to collect user requirements and expectations. To top off the dialog design, we applied established dialog design principles. This was the basis of the dialog implementation.

<sup>1</sup>Enrollment means the addition of textual representations to the vocabulary. This addition could be done while the dialog is running.

### 5.1 Questionnaire

In order to find out which MP3 functionalities are important for a speech-controlled MP3 application, a user questionnaire was handed out to employees of Temic and Harman/Becker in several countries, the majority of them being frequent MP3 users. The questionnaire covered a number of different features which can be grouped into three categories:

1. selection of track, album, artist, genre or composer by speaking its name;
2. while track is playing: read out information about the track, such as track number, album, or artist name on demand<sup>2</sup>;
3. read out a list of tracks, albums, artists, genres, or composers to the user on demand (as alternative to visual display).

As result, subjects regarded function category 1 a required feature, while category 2 was considered nice to have. Number 3 was rejected by the majority. People said they would feel annoyed by long lists of names read out, and they would rather prefer the visual display to get the respective information. Consequently, only the features of category 1 and 2 have been integrated, while type 3 was rejected.

Moreover, especially the frequent users required handling of user-defined playlists which will be integrated in a more advanced version of the application.

There was also a general consensus favoring simple, short dialogs. Users want to enjoy the music rather than being entangled in lengthy dialogs. This is somewhat contrary to e.g., navigation destination input where people expect that separate dialog steps are required to enter an address.

The MP3 player dialog

- handles music files and audio books
- lets the user select tracks by album, track, artist, genre, or composer name
- provides conventional navigation commands while a track is played: "next", "previous", "skip", "shuffle", "stop" plus the respective album commands. Of course, it is also possible in this dialog state (not only on top-level) to speak the selection and readout commands of category 1 and category 2 described above, such that the user can hear another track or change the current album.

<sup>2</sup>This item accounts for multi-modality to provide car-drivers a second information source while driving (the respective information is also shown on display).

Of the category 2 features mentioned above, only the information 'name of current track' and 'name of current artist' are provided.

The system can also play songs of a similar genre like the current music files ("play me more like this"). There is no sophisticated mechanism to fulfill this request, instead, just a track with identical genre information will be selected. In reality, genre selection poses a non-trivial problem (see section 6.2).

## 5.2 General Design Guidelines

There exist well-known design principles for user interfaces which work pretty nice in theory. In a practical application, however, there are always reasons when deviating from the ideal. We kept to them when designing the dialog as closely as possible:

**speaking style** users can use whole sentences (speak naturally) or use command-like speech;

**consistency** similar dialog parts are modelled as analogously as possible to simplify the user's understanding of the system;

**shortcuts** there is more than one way to have a command accomplished, e.g. 'select a track'. Shortcuts will be preferred by experienced users;

**what you see is what you speak** all commands on the display are speakable;

**feedback** there is acoustic or visual feedback on a user command. While it is superfluous when the command works as expected, the user feels abandoned when the system reaction deviates from the user's expectations. However, how much feedback will be implemented, needs careful consideration and depends on the respective dialog context;

**help** the user can get help on currently possible commands.

## 5.3 Dialog Flow

As always, dialog design has been a tightrope walk between satisfying a novice's and an expert's needs - while a novice needs guidance, an expert feels bored by repetitive actions of the dialog system. Having in mind that people mainly want a system to quickly accomplish a command rather than one that is overly verbose, the system was designed to support the user when necessary but not to bother her with too many questions or numerous dialog turns. During the dialog, we designed the system to offer minimal explicit support. So, at dialog start the system prompts the user to speak only by a beep tone. Help is provided, but in an unintrusive way:

- visual help: when the system waits for user input, a help screen displaying currently speakable commands is shown (s. Figure 1). It can be turned off (and on) by voice-command.
- dialog help (timeout/nomatch): the system offers possible commands after a timeout period or if the system did not understand the user utterance;
- explicit dialog help: the user can always ask for help, which is given in a context sensitive form.

To start the MP3 player application, the user can say "MP3 player" or just "music"<sup>3</sup>. The system then prompts the user to speak, at the same time offering a help screen that displays all possible selection criteria, as shown below in Figure 1.

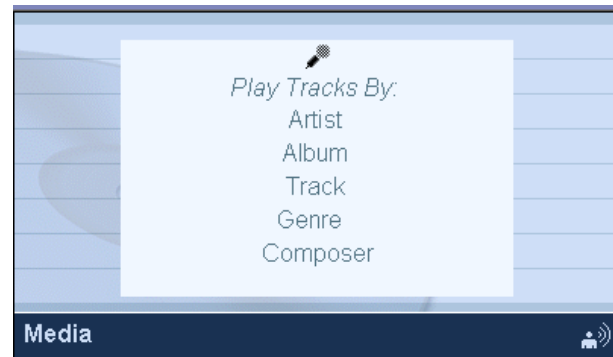


Figure 1: *Main display of the MP3 application.*

After having chosen a selection criterion, the system provides the appropriate item list. In order to optimize speech recognition, the active vocabulary is adapted dynamically to the current selection. For example, the track names of a current selection are given preference to the track names of the whole, possibly huge, MP3 collection.

The next section shows sample runs.

### 5.3.1 Sample Runs

To hear MP3 encoded music files, users can choose between the filters track, album, artist, genre, or composer which are regarded the most common use cases. After the selection, the system prompts the user to speak the respective item. The help screen, shown in Figure 1, is active per default, in this way suggesting a menu-driven dialog entry for a novice:

#### Sample A

User[1]: Music Player

<sup>3</sup>The MP3 application is part of a larger demo environment where several applications are available, but only one music application.

System[1]: MP3 Player  
 Display[1]: [s. fig. 1]  
 User[2]: select an album  
 System[2]: Which album would you like to hear?  
 Display[2]: [shows available albums for choice]  
 User[3]: Live in Paris  
 System[3]: [starts playing 1st track of album]  
 Display[3]: [shows current track information]

More experienced users can use shortcuts, ignoring the visual help:

#### Sample B

... (see dialog step [1] above)

User[2]: Play me the album 'Live in Paris'  
 System[2]: [starts playing 1st track of album]  
 Display[2]: [shows current track information]

The display showing the track information is illustrated in Figure 2.

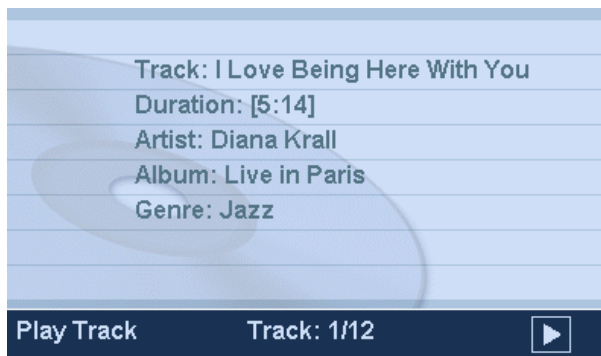


Figure 2: Display showing track information.

### 5.3.2 Refinements

The dialog flow, as described so far, does not yet account for the possibly huge size of a playlist, e.g. a list of Jazz titles, or a list of all Diana Krall titles if she is your favorite artist. Therefore, the system offers two commands modes, the usual play mode and the browse mode. The latter mode allows to search through collections by subcategories. We thereby assume that there is a natural categorization hierarchy which is *genre* → *artist* → *album* → *track*. For example, the command 'browse the genre Jazz' shows all Jazz artists, and saying 'browse the artist Diana Krall' will display the artist's albums. The command 'play all' will play the current selection. At the bottom of the hierarchy, all tracks of an album will be shown.

This additional feature will be evaluated with regard to the naturalness of the subcategories as well as to the transparency of both command modes.

## 6 Some Notes on MP3 Tags

One of the main challenges is making good use of the information contained in the MP3 tags. While there are tags for artist names, titles, album names, etc. the information is much less structured than one might expect. This leads to a number of challenges. A solution to these problems is the use of a database of meta information such as the CDDB by Gracenote. Embedded versions of this database are available and can be used to significantly improve the quality of the MP3 tags.

### 6.1 Recognition and Synthesis

One of the biggest challenges for both recognition and synthesis is finding the possible pronunciations for the titles and artists. The current system focuses on the US market but multilingual issues occur even here. There is a large Spanish speaking community in the US and even French titles are not uncommon (e.g., titles by Céline Dion).

Slang words are also very common in music titles and person names; even the human user is not always certain how to pronounce them.

All of these effects can occur within one track name, for example: "Femme like U".

Temic is working closely with Gracenote to provide a solution to these problems. Towards this end Gracenote will be providing phonetic transcriptions for the problem cases, greatly enhancing the recognition accuracy as well as speech output quality. If there is no database available, however, autotranscription will be the fallback solution.

Then there is the problem of partial matching. Many titles have longer official names than are usually spoken such as "The Shoop Shoop Song (It's In His Kiss)". In such cases it is rare that the full title is spoken.

Sometimes the title contains additional information that is not really part of the spoken name such as "All Cried Out (Unplugged)", or "Ka-Ching! (Red Disc)". This information can be used to distinguish different versions of the same title (sometimes even on the same album).

Here again it is helpful to work with a partner who provides high quality meta information to ensure that names and titles are transcribed in a consistent way. This in turn makes it easier to parse the data and search through them.

### 6.2 Classification of Genre

The current system allows to play songs similar to the one currently playing. This is called the "play me more like this" feature. Although interesting approaches for automatic classification from the young field of music retrieval exist (s. (Neumayer et al., 2005; Habich et al., 2005)), they do not yet provide reliable results. Therefore, we use, so far, simple string matching which is based entirely on the genre tag contents of the songs.

Given a good genre classification, this can be very useful. Nevertheless, classification by string comparison is problematic for a number of reasons:

- Many users set up the genre descriptions of their MP3 files by themselves, using the fact that the respective ID3 tag 'genre' allows arbitrary contents.
- An attempt to establish a standardization of genres existed for ID3v1, but was given up since it was found to be inconsistent and obsolete (ID3-Homepage, 2005). Furthermore, 79 genres are neither easy to remember nor fine-grained enough to support the "more like this" feature well.
- One and the same album or track can belong to different genres.

The reason for these problems lies in the fact that the genre names are completely unrelated even for genres that are obviously similar, e.g. 'Rock' and 'Hard Rock': the taxonomic nature of genres cannot be exploited since genre names are just strings.

Using a database of titles with a consistent and fine grained genre classification helps to solve this problem.

## 7 Conclusion

In this paper a MP3 player for embedded systems was presented, featuring the selection of titles and other selection criteria by speech. We described the embedded environment and the dialog design of the application. Furthermore, we illustrated the challenges of speech-controllable ID3 tags.

The current system is based mainly on user questionnaires, general design guidelines and the designer's intuition. It is therefore a first prototype that serves as starting point for further design evaluations, undergoing constant design-implement-test iterations to finally have a stable and user-approved version after a number of cycles.

## 8 Future Work

For the future, several extensions are planned. With regard to the possible commands that the system can understand, it will be extended by simple handling of playlists ('play my favorite playlist'), following the results of the user questionnaire.

Additionally, multilingual recognition and prompting are to be improved which meets real-world requirements.

Furthermore, we want to go to the limits of the speech recognizer by allowing quasi unrestricted utterances such as "Play Live in Paris". Different from the example B, here is no keyword indicating the filter criteria 'album', which clearly complexifies the recognition task. Moreover, methods for dissolving ambiguities will also be investigated. For example, album and artist names are often identical such as in utterances like "Play Diana Krall".

Finally, alternative approaches for genre classification will be evaluated.

## References

- Stephan Baumann and Andreas Klüter. 2002. Super-convenience for Non-musicians: Querying MP3 and the Semantic Web. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France.
- Dirk Habich, Wolfgang Lehner, Alexander Hinneburg, Philip Kitzmantel, and Matthias Kimpl. 2005. Eyes4Ears - More Than A Classical Music Retrieval System. In *5th MUSICNETWORK OPEN WORKSHOP - Integration of Music in Multimedia Applications*, www.interactivemusicnetwork.org, Vienna, Austria.
- Stefan W. Hamerich and Gerhard Hanrieder. 2004. Modelling Generic Dialog Applications for Embedded Systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 237-240, Jeju, Korea.
- Paul Heisterkamp. 2001. Linguatronic - Product-Level Speech System for Mercedes-Benz Cars. In *Proceedings of the Human Language Technology Conference (HLT)*, San Diego, USA.
- ID3-Homepage. 2005. <http://www.id3.org/>.
- Robert Neumayer, Thomas Lidy, and Andreas Rauber. 2005. Content-based Organization of Digital Audio Collections. In *5th MUSICNETWORK OPEN WORKSHOP - Integration of Music in Multimedia Applications*, www.interactivemusicnetwork.org, Vienna, Austria.
- Roberto Pieraccini, Krishna Dayanidhi, Jonathan Bloom, Jean-Gui Dahan, Michael Phillips, Bryan R. Goodman, and K. Venkatesh Prasad. 2003. A Multimodal Conversational Interface for a Concept Vehicle. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2233-2236, Geneva, Switzerland.
- Christian H. Schulz, Dmitri Rubinstein, Dimitris Diamantakos, Michael Kaißer, Jan Schehl, Massimo Romanelli, Thomas Kleinbauer, Andreas Klüter, Dietrich Klakow, Tilman Becker, and Jan Alexandersson. 2004. A Spoken Language Front-end for a Multilingual Music Data Base. In *Proceedings of the Berliner XML-Tage*, Berlin, Germany.