Voice control for in-car applications: present and future

André Berton

Daimler AG, Group Research & Development, Ulm, Germany, Email: andre.berton@daimler.com

Introduction

Electronic on-board devices in cars have permanently increased in number and complexity over the past years. That leads to an increasing risk of driver distraction. In order to minimise distraction speech has become an important input modality in the automotive environment. State-of-the-art in-car speech dialog systems provide assistance in operating audio, phone and navigation systems. This contribution first discusses past and present in-car speech dialog systems. Based on current voice control technology, it looks into future improvements of current applications and new applications which will be integrated in cars and can be voice-enabled.

History of in-car voice control

The first automotive voice control system was the Linguatronic [1], introduced by Mercedes-Benz in 1996. It came as hardware solution in a separate box containing an HMM-based ASR engine. Speech dialogs covered only very simple Command & Control tasks for phone and audio operation. Phone tasks included number dialling by saying a sequence of digits, and using voice tags (recorded speech patterns) to access phone book entries by name. Audio tasks included selecting the radio frequency, selecting the station name by voice tags, and choosing the next radio station. It also covered operating the CD changer by saying the number of the CD and title, or by just asking for the next title. Speech dialogs were started by pressing a special Push-To-Activate (PTA) button or lever. Dialogs continued until the goal of the task was achieved and the action was performed. Dialogs could be aborted by the same PTA button or lever. Speech output was based on pre-recorded prompts of professional speakers, which guaranteed a very high quality.

The next generation of speech dialog systems in cars also included voice control for navigation, which is a great challenge due to the destination entry function which chooses a address from a large database. In 1998, voice control was introduced for destination address entry by a spelling and list matching function. The driver was asked to spell the leading letters of the city name and was presented with a list of the closest matches. In a second dialog step the same procedure was repeated for street names. Addresses from the destination memory in the car could be enrolled by voice tags. In 2001, Linguatronic was further improved, such that the driver could say the names of the 800 largest cities as whole words. Finding the nearest petrol station or parking was also made available by voice control.

The current in-car voice control system of the Mercedes-Benz C class, which is the benchmark system at the moment, is a software solution on the headunit. It includes a Grapheme-to-Phoneme (G2P) converter to enable text enrolments. Phone book names and radio station names are automatically transcribed phonetically and added to the recognition vocabulary. These text enrolments are far more comfortable than the former voice tags. However, voice tags are still implemented in order allow voice control for names, where G2P fails.

The biggest technological improvement of the current benchmark system is the robust and fast speech dialog for whole word destination entry which is now available for entire Germany. The driver is asked to say the city name, then presented with the list of matching entries and asked to choose the line number of the entry in question. Selecting the street follows the same principle in a second step. Finally, a street address can be given. The benchmark system is not only based on pre-recorded speech. In can also speak dynamic content, such as traffic information. That functionality is enabled by a Text-to-Speech (TTS) module.

The future of in-car voice control

Which functions do drivers and passengers want to be able to control by voice in the future? This question is answered by customer studies. At the moment only a part of all functions of the headunit is voice-enabled. Therefore there is large room for improvements in voice control for the given applications on the headunit. Furthermore, there is a trend towards connectivity and off-board functionalities, much of these provided by the internet. Significant improvements and new functionalities are detailed in the following sections.

Improvements in on-board applications

Although the current benchmark system is a milestone in in-car voice control, there is room for improvement in almost every function and dialog. The Command & Control dialogs lack naturalness. They should be able to handle a larger variety of phrases. At the moment, the user can only enter one information item in one dialog turn. Moving towards a frame-based dialogs will allow the user to enter the entire address in one utterance, which will bring a significant improvement in naturalness and efficiency.

There is also room for improving the consistency of the user interface. In destination entry, the user can say the address by city and street name, but points-of-interest cannot be selected by their name. A consistent user interface requires such a function.

Some headunit functions are not voice-enabled yet, such as the time and date settings. Other headunit functions are only rudimentarily voice-enabled, such as the media player. More and more users are bringing their music to their cars and want to control this function by voice. New media will be integrated more and more into the headunit or connected to it by CE devices. Users want to select artists, albums and titles of their music collection on any connected device by name. [2] presents our approach to enhance accessing audio data within large databases by three different search strategies: category-based search, category-free search and physical search. We performed a user study to find out how users

refer to music content. That enables us to design an appropriate user interface and provide alternative names for music data. A particular challenge in music operation is internationality. Most music databases contain music from several languages. In order to transcribe the names correctly, the transcription engine must know, which language the name of the song belongs to. Text-based language identification has proven to be quite robust for that.

New functionalities and voice control

People get used to having current information available anywhere anytime, so they expect to have it in their cars, too. The internet is the medium, which contains the great majority of the information requested. First telematics services, such as BMW Online, already introduced internet access from the car, but without speech interface. This service covers information about local mobility, news, travel, and spare time, etc. The portal is maintained by the car manufacturer, but the content is provided by several suppliers, such as Deutsche Presse-agentur (dpa), 11880, ViaMichelin, apotheken.de, Parkinfo and others. Such an internet portal can be extended by voice access using VoiceXML technology. However, there are technological challenges, particularly due to inferior recognition quality of speech transmitted over the very limited phone bandwidth in combination with the noisy car environment.

In the car scenario of the project SmartWeb, funded by the German Ministry of Education and Research, we investigated three architectures for making available internet information for voice access in the car. These architectures are considered three stages to incorporating voice-enabled access to internet information, starting with transmitting automatically generated dialog applications by radio, ending with complete off-board server-based speech processing.

Our first approach is based on generating question answering speech dialogs automatically from well-structured internet pages using an off-board server processing. The speech dialogs together with the data they access are then transferred to the car by digital broadcasting, such as Digital Media Broadcast (DMB). The dialogs are added to the existing speech dialogs on the fly by updating a second dialog manager, the Natural Language (NatLang) DM. A Meta Dialog Manager MDM coordinates the activities of both speech dialog managers and transfers new application names and details to the GUI. No uplink from the car is required. The architecture is shown in Figure 1, stage 1.

Since automatic dialog generation does not consistently lead to natural speech dialogs and to understandable prompts, we considered a second approach, where speech dialogs were manually designed to provide state-of-the-art quality. These dialogs can still be transferred to the car by digital broadcasting. However, they also enable interfacing standard web services using the WSDL interface. These interfaces annotate data fields, so that some particular data, such as phone numbers and addresses, can directly be transferred to the headunit phone and navigation applications. This tight integration (s. Figure 1, stage 2) allows the user to navigate to an address and to call a number received from the internet. Subjects did appreciate this technological combination of well-designed speech dialogs and current information.

The third approach was based on bi-directional audio and data links from the car to a server (s. Figure 1, stage 3). The server enabled a deep semantic processing and internet search using resources that will not be available in the car. However, audio processing including echo compensation needed to be processed in the car. This technological challenge has not been solved yet and recognition results from noisy in-car speech transferred over the limited GSM bandwidth proved to be currently technologically unfeasible.

Finally, we looked into personalizing the user interface. All internet applications were represented in a list. This list was sorted according to explicit user preferences and implicit scores accumulated during speech access to the services [4].

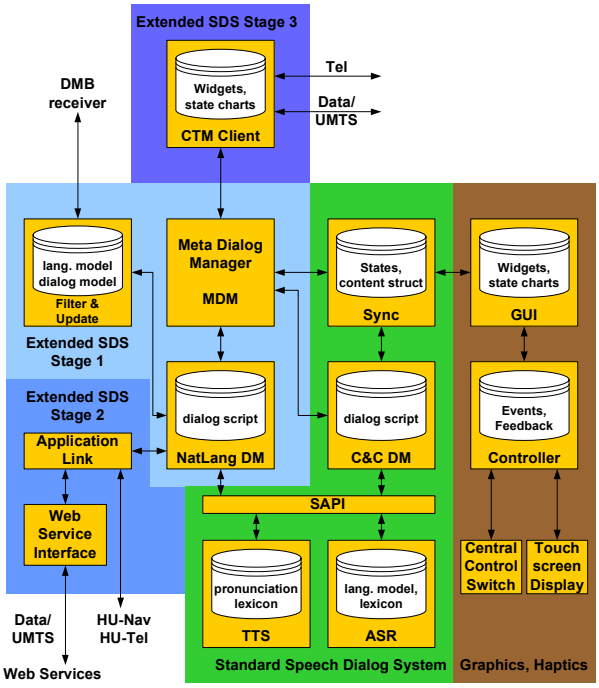

Figure 1: Dialog system architecture.

Conclusion

This paper summarizes the history and state-of-the-art of in-car voice control systems. Although current voice user interfaces receive good press reviews, we conclude that there is much room for improving naturalness of dialogs (frame-based dialogs), for new applications, such as accessing large multi-lingual data sets (music selection), and incorporating internet information dynamically in speech dialogs.

References

[1] Heisterkamp, P.: Linguatronic: *Product-Level Speech System for Mercedes-Benz Cars*, In Proc. Of Human Language Technology, HLT, San Diego, USA, 2001

[2] Mann, S., Berton, A. and Ehrlich, U.: *How to Access Audio Files of Large Data Bases Using In-car Speech Dialogue Systems*, Interspeech 2007, Antwerp, 2007

[3] Berton, A. et al.: *How to Integrate Speech-Operated Internet Information Dialogs into a Car*, Interspeech 2007, Antwerp, 2007

[4] Fischer, P., Berton, A. and Regel-Brietzmann, P.: Personalisierte Sprachinteraktion zur Priorisierung von Internetinformationen im Auto, ESSV, Cottbus, 2007