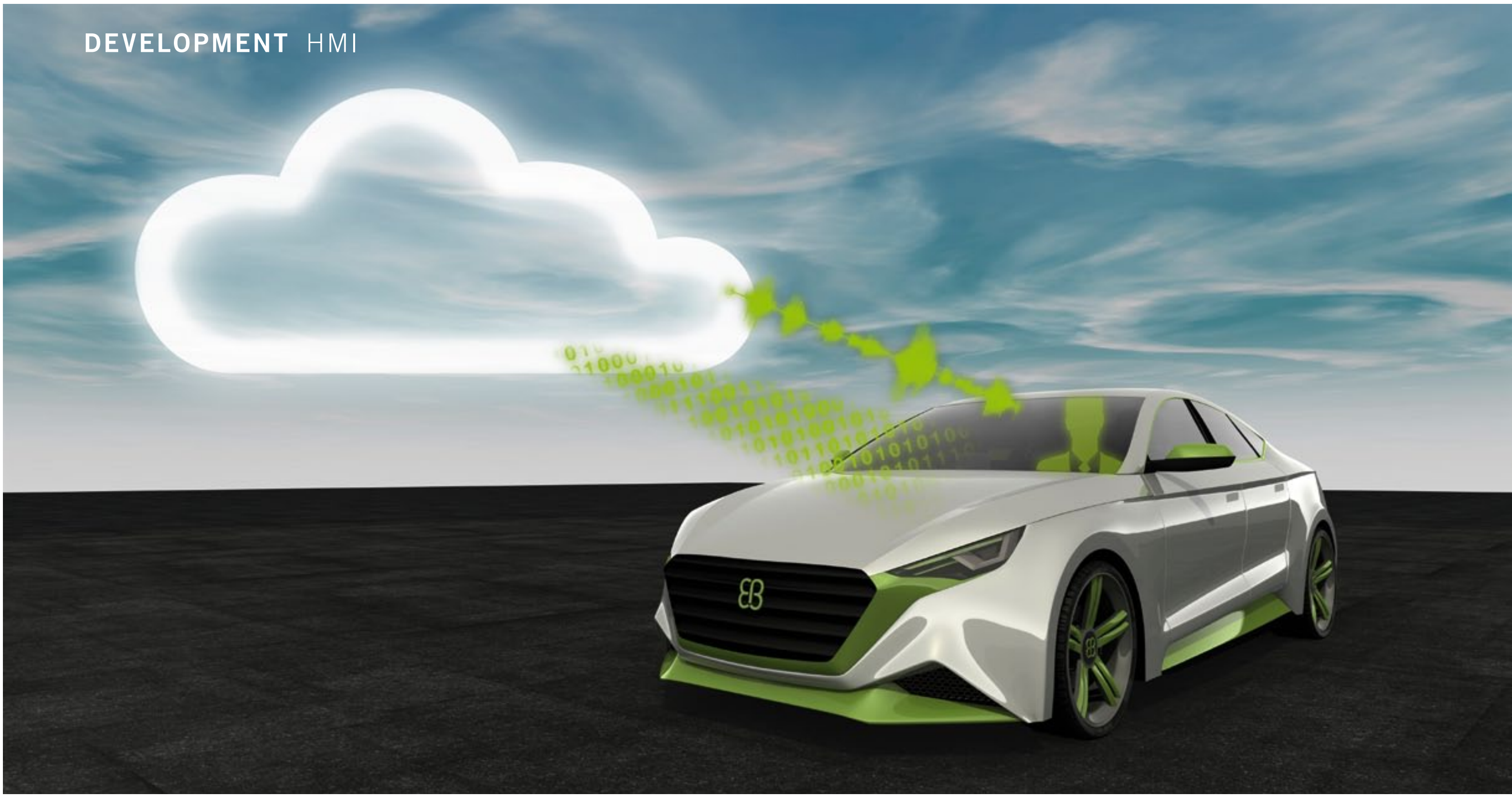


AUTHORS

DOMINIQUE MASSONIÉ
is Product Manager HMI at
Elektrobit in Erlangen (Germany).

TIMO SOWA
is Software Developer HMI at
Elektrobit in Erlangen (Germany).

FUNDAMENTAL ASPECTS OF SPEECH DIALOG SYSTEMS

Speech dialog systems are an important component of in-vehicle information and communication system operation which contribute to convenience and safety. They ensure that the driver can keep his eyes on the road and is less distracted from driving by the operation of infotainment systems. The more freedom the driver has in the formulation of voice commands, the less time he requires to familiarise himself with the operation of the system. Ideally, the driver shouldn't need any prior knowledge in order to be able to execute the required functions by voice command, which is why "Natural Language Understanding" or NLU is a key objective.

However, it is important to understand that NLU isn't a technology but a design principle, and a range of speech recognition methods can be used for NLU.

The point of origin for a dialog system is phoneme-based speech recognition. Words are broken down into their smallest acoustic components – phonemes – which are similar to syllables. The recognition process involves assigning the input signal to the most probable sequence of phonemes with the help of phoneme dictionaries, ❶. Over the last five years, considerable progress has been made in this field and results are becoming far more robust. The problem of multilingual entries, which has caused difficulties in the past (e.g. when foreign language song titles, names or destinations are included in an otherwise German language context), has been considerably alleviated by the use of multilingual dictionaries.

Up to now, voice commands have predominantly been interpreted in a grammar-based process. Strict construction rules exist to define admissible voice commands. This makes the analysis of voice commands and responses comparatively easy, though only previously defined sentence sequences can be processed.

In modern and flexible systems, the grammatical interpretation of voice commands is supported by a separate semantic analysis (also called 'topic classification'). The main issue in this connection is the meaning that the user intended with the voice command. The interpretation of longer sentences therefore poses a considerable challenge to the system's comprehension capacity, particularly since it may have to identify possible contextual ambiguities. A practical approach is to use dialog to solve these ambiguities. If necessary, the system asks questions in order to clarify context and meaning. This means that potential dialogs are

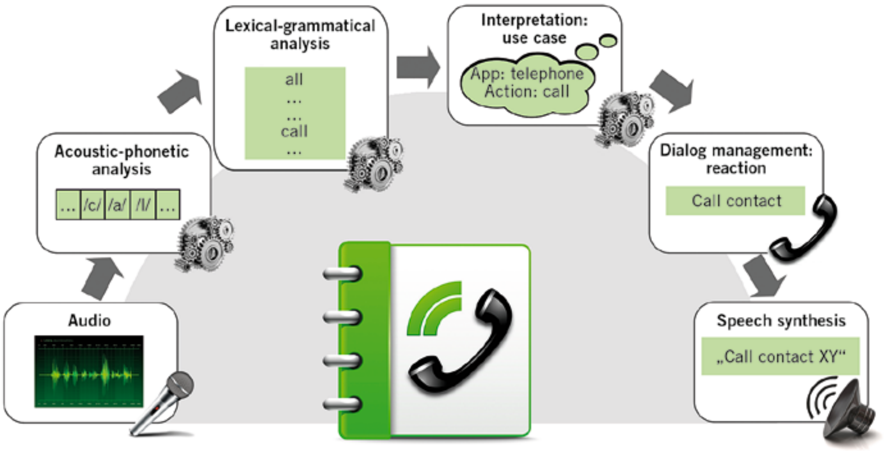
SUPPORTED BY THE CLOUD
DIALOG BETWEEN CARS AND THEIR DRIVERS

The range of functions offered by cloud-based in-vehicle infotainment systems is expanding all the time. The operation concept is still a decisive criterion in this context because the driver has to be able to concentrate on the road, the safety of other road users being a top priority. The degree to which the driver is distracted by the systems that are designed to support him and enhance the driving experience is therefore a focal topic of discussion. Voice recognition systems offers one of the need solutions. Some are cloud-based – on the one hand very innovative, on the other hand this must be discussed critically. Elektrobit describes newest technologies and discussions.

more complex but more natural, and therefore involve less stress for the driver.

Voice output in dialog systems is generally text-to-speech-based (TTS). While allowing greater freedom of voice command input, these systems usually only have one voice output option. A logical requirement of an efficient and "natural" system is 100 % compatibility of input and output, which means that in multilingual dialog situations the system selects its responses in the language that the user gave the voice command. However, this is a challenge that has not yet been overcome in speech dialog systems.

Another interesting trend in human-machine interface (HMI) modeling is multi-modality. Voice commands are generally activated by pressing a voice button on the steering wheel. In future, additional or alternative operation steps will be possible such as combination with touchscreens or camera-based gesture recognition. This will permit the development of innovative operation concepts. For example, the driver can tap a point on a map that is displayed on a touchscreen and ask "What is that?" or command the navigation system to "Take me here."



❶ Sequence diagram of a speech dialogue based on the example of an address-book

CURRENT SPEECH RECOGNITION TECHNOLOGY

As already mentioned, the older grammar-based speech recognition systems have been augmented by limited semantic analysis. The dialog involves a relatively rigid formulation of voice commands which are then analysed by the system, ❷. Modern systems break down these processing stages and a dedicated semantic analysis is applied to the grammatical analysis.

One important prerequisite for natural language understanding and more flexible speech recognition is the use of Statistical Language Models (SLM). Admissible input is defined by using a statistical model of possible words and phrases. The systems can be trained by processing large volumes of possible words and phrases so that they can calculate the probability of certain word combinations occurring, or the probability of a word occurring at a certain point in a sentence. On the basis of this knowledge, the system analyses voice commands to ascertain which word at that point in the sentence would make the most sense in a given context. To develop a 'training corpus' for every supported language involves vast volumes of data of up to several hundred megabytes per language. Comprehensive training corpora and statistical models are available from providers such as Nuance or Voicebox.

Dynamic content such as names or terms which aren't initially contained in the dictionary such as user contacts pose additional challenges. The fact that modern systems can have several thousand name entries is an issue that should be addressed by the dialog concept. The same applies to calendar entries, song titles, point of interest entries in the navigation database and many other things. The phonetic transcription which would be necessary to detect the names that are entered in the directory is usually not stored in the system. For this reason the phonetic transcription has to be generated, based on a set of rules, when these data are imported in the infotainment system. In the same process, abbreviations in contact and address lists such as Dr. are adapted to "doctor" for speech recognition purposes.

New infotainment system services such as political and sports news, fuel prices and other information mean additional content that the speech recognition system has to understand and the dialog system has to be able to process for language output.

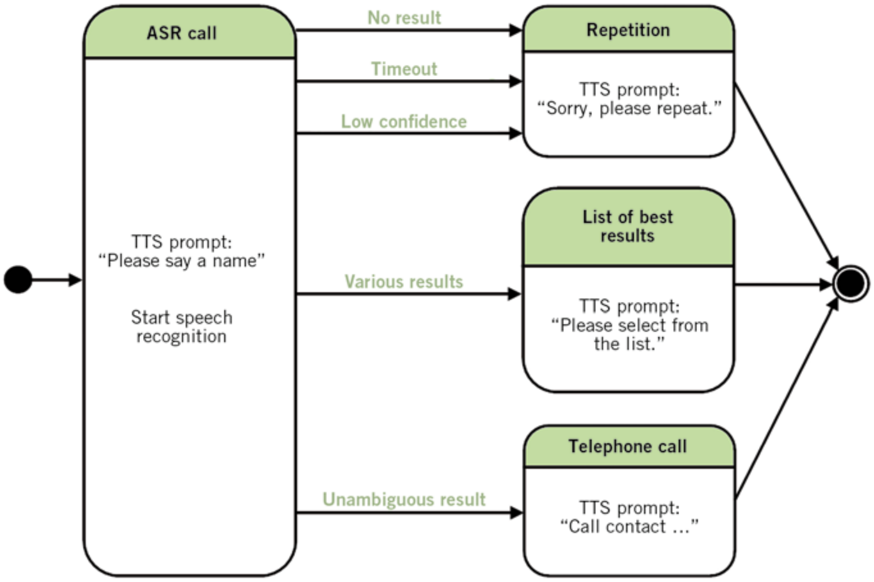
CLOUD-BASED SPEECH RECOGNITION

These functions and the general requirement for additional computing resources and storage capacity, as well as the more effective support of multilingualism, have prompted considerations on how some aspects of speech recognition can be outsourced to the cloud. Users of smartphone voice technology such as Apple's Siri, Google Now and Samsung S-Voice are already familiar with cloud services.

The advantages are obvious – the cloud has vast resources which can cope with more complex entries and speech recognition databases can be quickly updated with new content from news reports, sports events, stock market data and similar. Since thousands of people use the system in different contexts, the quality of recognition continuously improves. Users of cloud-based systems in several countries profit from the fact that they can recognise and process several languages.

However, there are also disadvantages. Access to cloud services necessitates a high-speed internet connection, which can only be achieved in a car if there is very good wireless coverage. In roaming situations, high data transmission costs add to the problem of network reception. The use of cloud services is associated with a number of privacy issues. Recognised voice commands relating to navigation destinations, contacts or calendar entries are private data which can be clearly allocated to the user. In addition to the service providers, mobile communications providers and cloud infrastructure providers such as Amazon or Microsoft are involved in the transmission, storage and in some cases processing of the user's personal data. OEMs also have to take into account that different countries have different technical and legal frameworks.

The decision on whether and to what extent cloud-based speech recognition systems will be used ultimately lies with the OEM who commissions the system.



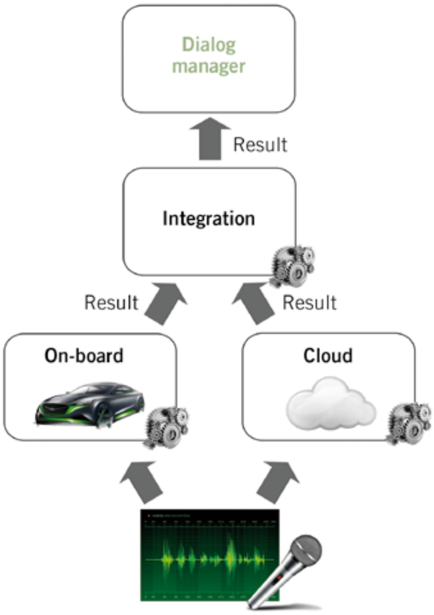
❷ Speech dialogue process for „Please call…in my address-book"

AVOIDING DRIVER DISTRACTION AND WORKLOAD

As already mentioned, voice commands are an important means of reducing driver distraction and workload in certain driving and operation situations. However, the use of voice command in the car isn't just part of the solution, it's also part of the problem. Legislators around the world are currently working towards tightening up legislation on the use of information and communication systems while driving. Also, the increasing integration of news and information from the internet can lead to the driver being overtaxed, despite operating concepts such as voice commands, in certain situations. Recent research in this field also indicates that drivers have to be maintained in an optimum workload corridor in which the vehicle systems neither over nor undertax them, because boredom and monotony have a negative impact on attention and performance. When designing human-machine interfaces (HMI), one of which is the voice command system, it is therefore necessary to take driving situation and context into account. Other in-vehicle systems that manage information about traffic, route and driving situation can influence the scope of functions and operating options. Fatigue recognition and warnings which obtain data from steering wheel use and other operating procedures are also standard features in some cars.

Even today, certain functions such as TV reception are blocked when the car is in motion. In future, infotainment systems could be more fine tuned to respond to situations and the driver's concentration/fatigue status, and regulate access to certain functions depending on the situation or status.

Biometric recognition mechanisms are still a future vision, though feasible in this connection. For example, the



❸ Software architecture for hybrid speech recognition

user's voice pattern, a camera image and/or other parameters could be used to ascertain driver attention levels and workload. Although we are still some years away from introducing such mechanisms, other biometric processes such as voice and/or camera based driver recognition may soon be incorporated in vehicles.

OUTLOOK – HYBRID SYSTEMS WITH OR WITHOUT SMARTPHONE USE

Practical considerations make it impossible to depend entirely on the cloud for speech recognition and an on-board implementation should at least be feasible as a fall-back solution. A hybrid architecture, ❸, could incorporate a comprehensive catalogue of local functions, supplemented by other functions in the cloud.

In this connection, it is important that the voice commands are compatible with all other HMI applications. Everything that can be operated by voice command, whether locally or in the cloud, has to be included in the dialog and statistical language analysis systems. The complexity and footprint of a language dialog system are therefore today very similar to manually operated HMIs with graphical displays. This is why some people in the sector call voice command systems "second HMIs".

This is the context in which we have to view the integration of mobile phones and smartphones in the vehicle. The on-board infotainment system retains control, but can access smartphone content and, in future, also smartphone functions. For example, it would be feasible to integrate the smartphone display in the HMI or to outsource certain infotainment system functions to smartphone apps. In future, therefore, the smartphone will be a third factor, alongside the on-board system and the cloud, to be taken into account when designing hybrid HMI concepts. Recent announcements from Apple (CarPlay) and Google (Android in the Car) point towards this development. Since smartphone systems have their own speech recognition systems, it will be necessary to solve the problem of whether and how Siri, Google Now, S-Voice and Co. can be made compatible with the on-board dialog systems in future.