User Requirements for Speech in Automotive Applications

Dr.-Ing. Hans-Wilhelm Rühl, VDO Car Communication, Wetzlar

Hans-Wilhelm.Ruehl@de3.vdogrp.de

Summary

Speech technologies are currently finding their way into the car environment. Compared to other environments, the car situation is characterised by high ambient background noise and a hands-busy eyes-busy situation of the driver. New advanced information systems like navigation and telematic services require speech recognition and speech output to serve the driver conveniently and safely. The related requirements for speech technology are presented.

Constraints by car environment, market, and user view

In a car, the primary task for the driver is to control the car and the traffic situation. The eyes should be on the road, and the hands preferably on the steering wheel. Navigation, telematics and entertainment are only secondary appliances intended to assist the driver, and not to distract his attention from the road.

Acoustically, the environment covers a wide range of SNRs from standing car with nonstationary street noise at rather low level up to SNRs around 0 dB for high speed driving or ventilators at full speed or when windows are open. Voice-control for sound functions must be operable without sound output to be muted.

Technical constraints for speech technology in automotive environment differs strongly from e.g. the PC environment. Table 1 shows some major differences between automotive and PC market.

In the future, the user perception of the infotainment functions inside the car is going to change. Current "separate box" applications like radio, telephone, or navigation, will merge into an integrated infotainment system with a unified user interface These systems will offer new applications like access to internet-based and internet-like telematics services, e.g. via WAP. They will allow interaction of car based applications with remote services. Personal appliances like PDAs and mobiles may be linked to the car information system, allowing to synchronise or share databases, or to provide an automotive MMI to some relevant PDA applications.

Required Speech Technologies

The full range of speech technology is needed for the future car environment:

Hands-free talking including acoustic echo cancellation, noise reduction, and microphone array processing is used for telephone conversations, and as preprocessor to speech recognition and voice capturing.

Speech recognition must be operable hands-free, hence has to be robust against car noise. Speaker independent command & control will be standard, using isolated words or a finite-state syntax. For name dialling, personal address books or favourite radio stations, speaker dependent recognition is required.

For navigation address entry or future server based applications, recognition should employ phonemes to construct new vocabularies online.

Continuous speech is required to enter phone numbers as digit strings, or to spell the begin of a name. Keyword activation will replace push-to-talk operation, and keyword spotting will make operation more reliable. Future speech recognition will also use natural language technologies.

Speaker identification and verification modules will offer both anti-theft protection and comfort by activating driver specific presets like seat configuration, phone repertories, favourite radio stations etc.

Voice coding schemes that are robust against background noise will be used to record voice notes. Voice decoders optimised for sound quality and low bit rates will reproduce speech permanently stored for guidance and help messages.

	computers	*car information systems*
development lead time	0.5 years	3 ... 5 years
next generation after	0.5 years	~ 7 years, facelift after 3 years
product lifetime	2 ... 4 years	10 years support after end of prod.
CPU clock	400 ... 1000 MHz	100 ... 200 MHz (automotive chips!)
RAM	64 or more MB	2 ... 8 MB
permanent mem.	6 ... 24 GB	4 ... 16 MB Flash (no harddisk!)
mean time between breakdowns	~ days	~ years

Table. 1: Technological and industrial constraints for computers and car information systems

Text-to-speech will offer improved navigation guidance ("Turn right into Main Street"), language independent traffic information based e.g. on TMC data broadcast, or will read aloud emails and user manuals.

Multilinguality

As car models more and more are distributed globally, multiple languages have to be supported. E.g., VDO Dayton navigation systems communicate in British or American English, French, German, Dutch, Spanish, Italian, Swedish and Danish, and one European car manufacturer issues user manuals for cars in 14 languages. To preferably have the same car and configuration in every country, all languages should be available inside the system. As currently the necessary data has to be stored in Flash memory, compromises have to be taken for the time being.

Address Entry and Translinguality

For address entry, dynamic vocabularies using phoneme based recognition are mandatory. To be able to enter city or street names by voice, the related phonetic names have to be stored as part of the map database. Related vocabulary sizes may become quite large, e.g. the VDO map of Germany contains 5000 location names, and Berlin has about 8000 streets.

As these vocabulary sizes are impractical both with respect to recognition accuracy and real-time performance, temporary by-passes like spelling, vocabulary reduction using area codes or region preselection are introduced. Finally, more powerful recognisers and more intelligent MMIs will solve the problem.

Availability of adequate phonetic transcriptions is a problem in several areas. Easiest to solve is browsing related to WWW or WAP. Currently, the related communication protocols (HTML, XML, WAP, ...) do not provide phonetic names to allow voice based browsing. But fortunately, vocabulary sizes are quite low, so even very simple text-to-phoneme conversion tools may help intermediately. On mid terms, voice extensions will be added to XML.

A more serious problem is adequate phonetic transcription of addresses for use in foreign countries. While most Europeans are able to correctly pronounce English locations and some French locations, location and street names in other countries typically will be pronounced using the driver's native pronunciation rules. E.g. in one experiment executed at VDO, speech recognition accuracy for German speakers was found to be 20% better on a 100-Dutch-cities task, if the Dutch cities were transcribed in German pronunciation instead of correct Dutch transcription.

Tools to provide related native phonetic transcriptions can be provided, but several questions remain:

- Do the TTP tools need additional target language specific rules?
- How can drivers with some knowledge in the foreign language be handled?
- Do we finally need n TTP tools for a set of n languages with n^2 optimised rule sets?

It is possible to provide simplified solutions for the major languages on short term, but for a high-quality solution, there is still considerable work to do.

The complexity is still increased when looking at on-line template construction for name-dialling, using e.g. SIM card or PDA based contact databases. In this case, no phonetic names are available in stored form, the language origin (i.e. the underlying pronunciation rules) of the names is unknown.

Even worse, the driver knows how to pronounce the names, and assumes the application to be knowledgeable, too. This is also critical in case that text-to-speech is used to echo the name of a person as recognition result.

As PDA based contact databases tend to grow with time, sizes of up to 1000 names will not be unusual in the future. To get sufficient recognition accuracy for such large vocabularies, a high correlation between estimated pronunciation and user pronunciation is crucial.

As for the time being, there are no efforts to place phonetic names into the contact databases, a text-to-phoneme conversion tool probably will have to serve as an intermediate solution. This tool will have to be enhanced by a decision tool to estimate the origin of the names. But even with these tools, former immigrants like e.g. "Monsieur Schneider" and "Mister Pierracini" still will cause problems.

Conclusions

Current speech technology is mature enough for use in car based products. The use in cars will grow rapidly with falling prices for the technology, and with a more widespread use of voice MMIs in other environments. Research is well organised to solve the remaining problems that are primarily accuracy and vocabulary size related. Decreasing memory prices and increasing processor power will support problem solution.

Concerning multilinguality, more and more databases will become available for more and more languages. This requires little inspiration, but a lot of transpiration.

In the area of transnational and translingual use of speech recognition and speech synthesis, the correct pronunciation and the most likely expected pronunciation of person names, location names and street names is often unclear. Temporary tools may help to find intermediate solutions. But to derive proper solutions, still basic research is needed to both examine pronunciation behaviour, and to specify translingual pronunciation rule sets.