Speech Recognition for Mass-Production Vehicles

Nils Rohrer[†] and Eike Gegenmantel[*]

[†]*ScanSoft Aachen GmbH, D-52072 Aachen, Germany*

[*]*Philips Research Laboratories, D-52066 Aachen Germany*

Email: nils.rohrer@scansoft.com, eike.gegenmantel@philips.com

Introduction

In many areas speech is a considerable choice for user interfaces to devices and services. In a vehicle, speech control has specific value: It is a true additional channel to the many already existing buttons and displays. Speech control does not need the hands, they can stay at the steering wheel, and with an acoustic feed back, the eyes are not forced to a display and remain on the road. There are high demands on speech control in vehicles: The noise level is comparably high, the underlying electrical devices have to fulfill high physical requirements, and the cost pressure is very tough when targeting the mass market. This paper will give an overview over how the combination of microphone position and direction, and electronic and algorithmic filters copes with the acoustical conditions. The paper will also show that mature digital signal processors fulfill the physical requirements at low cost, however at another price: Sophisticated software engineering is needed to prepare complex speech recognition and filter algorithms in a way that - despite limitations of memory space, computing power and precision - the required quality of speech control is ensured.

Applications

Applications for cars are navigation, car-audio- and telephone-control. These applications differ in complexity of the dialog and the number of used words in the vocabulary. In table 1 the number of words and the type of the acoustic model is listed for different usages. Since in a recognizer for navigation many more words are active at the same time compared to a speech recognizer for telephone control, not only the memory usage but also the necessary computation power is much bigger. Currently most automotive speech applications are for telephone control as add-on to hands-free car kits available for the mass market. With decreasing hardware costs more sophisticated applications will target this market. This will lead to new challenges for user-interface design and recognition accuracy.

Application	words	vocabulary models
Navigation	$\gg 1000$	phonemes
Car-audio	≈ 100	phonemes & whole word
Telephone	≈ 30	whole word

Table 1: Different applications for speech recognizer

User interface

The quality of the user interface is essential for the acceptance and the comfortable use of recognition systems. The structure of the dialogue as well as the feedback by sound prompts must be designed to lead the user in a natural way of communication. For instance the choice of voice commands in prompts and recognition should be matched to avoid mix-ups and disorientation of the user. The choice of non-similar words also can help to increase the recognition rate. In modern systems the combination of display and voice prompts assists the user in controlling the system.

Hardware requirements

A limitation for the user interface in respect to the sound prompts and the display is given by the hardware costs. If the recognizer is integrated in the design of the car, it can make use of existing displays. This is not possible for after-market kits. Due to cost limitations they rarely come with their own display. Another limitation is the length and quality of the sound prompts. Long sound prompts need lots of memory or, if compressed in a reasonable quality, computation power for playback. If the system is produced for more than one language, the memory cost again increases. In modern recognizers the data for vocabulary and sound prompts often demands more memory than the recognizer engine.

Solutions for better noise robustness

The main challenge for speech recognition in automotive environments is the high noise level and the resulting comparably bad SNR. There are several solutions to deal with this problem which will be discussed in the following. In general, they can be categorized into signal-processing, training of acoustic models, and hardware solutions.

Signal processing

Signal processing is done at recognition time and therefore time-critical. Consequently, signal processing algorithms have to be checked carefully whether they can deal with limitations on computational power and memory access times. Some methods are listed in the following:

- Spectral Subtraction [1] can handle stationary noise at a marginal amount of calculation power. Due its favorable cost-value ratio it is a frequently used algorithm.
- Beamforming [2], blind source separation [3] and adaptation to automatically detected driving conditions [4] improve the SNR but need more than one microphone,

[*] Eike Gegenmantel was with ScanSoft Aachen GmbH and is now with Philips Research Laboratories

which is currently out of scope for most mass-production cars.

- Speaker Adaptation and Noise Adaptation [5] is a promising algorithm which results in an increased need of calculation power and used memory.
- Parallel recognizers can be used to improve the recognition rate [6] but definitely increase the needed calculation power and exceed most cost limitations.

Improved Training methods

In general, improved training methods will be made offline without the restrictions of the embedded target platforms. This is a good way to minimize the cost for mass-production and to increase the recognition rate. This methods increase the recognition rate but cannot substitute good signal processing and good audio hardware. Some known training methods are:

- Corrective and rival training [7] as well as discriminative and robust training [8] can be used to improve noise robustness.
- For noisy environments the use of noisy training data is helpful. Best results are achieved when taking the original noise of the targeted car.

Hardware solutions

Hardware solutions normally do not decrease the noise level itself. They are designed to receive as little as possible of the noise and at the same time as much as possible of the speech signal. Some methods are described in general.

- High pass filter with a cutting frequency at around 200 Hz suppresses a major part of typical car noise. If this part is not recorded, this also enhanced the dynamic range for the wanted signal. This solution is simple in principle, in practice it has to be well adjusted to avoid the deletion of energetic parts in speech. The filter can be realized in hard- or software. If it is realized in software, the dynamic range will not be expanded.
- The positioning of the microphone [9] has a deep impact to the quality of the received sound. It affects also the level of the recorded signal and the SNR. Not only the room acoustics but also structure-borne noise has influence to the right placement. The position must be planned in the design phase of a car to avoid subsequent conflicts.
- The directivity of the microphone has influence on the SNR. If a strong directivity to the drivers position is possible, it should be chosen. In this case recognition rates for the driver will be better, but for the co-driver will be worse.
- The choice of the microphone is important for the recognition quality. It must be ensured that the frequency response is nearly linear in tight limitations; the variation between different exemplars of the chosen microphone has to be minimal. For the layout of the audio path it has to be considered, too, that modern voice recognizers increased their sampling frequency from former 8 kHz to currently 16 kHz so that the requirements on frequency response changed.
- Audio codecs, pre-amplifier, the complete audio-path and the cabling must be designed for a reasonable sound quality. Shielding against influences from other car electronics is required.

Summary

Several solutions for voice recognition can be found in the market today. Their number and quality will increase in the future if car manufacturers plan them as an integral part of the user interface. Cheaper and more powerful hardware will allow to realize some of the mentioned approaches for better noise robustness and recognition performance. In the actual situation with strong limitations on memory size and DSP power, the focus is on small solutions like telephone control which can be produced with satisfying results for the mass market now.

References

[1] Kotnik, B., Vlaj, D., Kačič, Z. and Horvat, B., “Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures”, Proc. ICSLP-02, Denver, pp. 445-448, USA, 2002

[2] Betlehem, T. and Williamson, R.C., “Acoustic beamforming exploiting directionality of human speech sources”, Proc. ICASSP-03, Hong Kong, vol. 5, pp. 365-368, PRC, 2003

[3] Saruwatari, H., Sawai, K., Lee, A., Shikano, K., Kaminuma, A. and Sakata, M., “Speech enhancement in car environment using blind source separation”, Proc. ICSLP-02, Denver, pp. 1781-1784, USA, 2002

[4] Banno, H., Shinde, T., Takeda, K. and Itakura, F., “In-car speech recognition using distributed microphones - Adapting to automatically detected driving conditions”, Proc. ICASSP-01, Hong Kong, vol. 1, pp. 324-327, PRC, 2002

[5] Yao, K., Zhu, D.-L. and Nakamura, S., “Evaluation of a noise adaptive speech recognition system on the AURORA 3 database”, Proc. ICSLP-02, Denver, pp. 457-460, USA, 2002

[6] Cristoforetti, L., Matassoni, M., Omologo, M. and Svaizer, P., “Use of parallel recognizers for robust in-car speech interaction”, Proc. ICASSP-03, Hong Kong, vol. 1, pp. 320-323, PRC, 2003

[7] Meyer, C. and Rose, G., “Improved noise robustness by corrective and rival training”. Proc. ICASSP-01, Salt Lake City, pp. 293-296, USA, 2001

[8] Hong, W.-T., “A discriminative and robust training algorithm for noisy speech recognition”, Proc. ICASSP-03, Hong Kong, vol. 1, pp. 8-11, PRC, 2003

[9] Kettler, F. “Einbauort des Freisprechmikrofons im Fahrzeug: Erster Schritt in Richtung guter Sprachqualität”, Automotive Engineering Partners, Vieweg Verlag/GWV Fachverlag (5/2003), pp. 36-39