SPEECH RECOGNITION IN THE CAR
From Phone Dialing to Car Navigation

Dirk Van Compernolle
Lernout & Hauspie Speech Products NV
St Krispijnstraat 7, 8900 Ieper, Belgium
Tel. +32 2 456 05 00, Fax +32 2 460 01 72, E-mail Dirk.VanCompernolle@lhs.be

ABSTRACT

This paper focuses on the evolving demands for speech recognition in the car and its corresponding impact on algorithmic and technological development. Till today the major demand for speech recognition in the car was related to hands free operation of the telephone. This functionality could be provided in a satisfactory way with a word based system, at the same time allowing for more simplistic noise suppression algorithms. Fully new speech recognition systems are required today to be able to cope with the demands for voice control of car navigation systems. These systems require noise robust phoneme based large vocabulary recognition systems and a much more advanced user interface. The very large perplexity of a car navigation task requires inherent embodiment of a spelling recognizer. Hardware and software design for this new application must also be tackled from the point of view that it will be one, though central part of a fully integrated speech control inside the car.

1. INTRODUCTION

The automobile has been hailed as one of the most promising application environments for speech recognition since a long time. The motivation for this is simple. It is the hands busy - eyes busy environment in which hundreds of millions of people operate daily. While there have been several speech enabled products on the market, acceptance has been slow and penetration low. This has been caused by a combination of factors: the high noise environment is a great challenge, there was much confusion about which functions were actually the most promising to be handled by voice, extra hardware cost was a critical consideration and the automotive industry is slow acceptor in general. Some poorly accepted introductions of speech output for signaling more than 10 years ago, caused designers to be even more careful with the introduction of speech recognition.

There is a continuous drive to bring more and more new technologies into the car, in particular aspects of the information technology society. By doing so one hopes to make some use of the hundreds of millions of hours that people spend in their car daily. On the other hand ergonomists claim that today's dashboards are already overcrowded and that there is strictly no room for new devices, or at least not for any new switches or controls. It is obvious that those who are dreaming of bringing the office into the car will not be able rely on a classical keyboard and screen man machine interface but that they will need to rely almost exclusively on speech recognition for control and for text-to-speech for system feedback. The high end navigation systems that have been introduced over the past years are a good example of such complex systems. Operation of the devices is cumbersome at best and will therefore preferably happen in a standing vehicle. While navigation is the prime feature that needs voice control today, it will just be one of many in the future. Hence adding speech recognition to the navigation device should be tackled from the point of view of adding speech recognition to the car in general. Functional controls in the car can be put in following logical groups: driving functions, secondary controls (heat, radio), telephone and navigation. It is commonly accepted that the demand for speech recognition can be ranked for those groups as: non existing, small, high and essential. These categories also define the different types of speech recognition products that might be envisaged and what there cost might be.

In this paper we examine the needs and solutions for speech recognition for 2 real world applications of the latter classes: dialing and navigation. We show that for these 2 applications with very different functional requirements the solution to the noise problem can be handled in different ways and that MMI considerations are as critical as plain recognizer design. This paper will be organized as follow. In section 2 we will analyze the requirements of the applications in greater detail both from acoustic and functional spec. In section 3 we will propose prototypical designs for both systems and finally we will make some concluding remarks.

2. REQUIREMENTS FOR SPEECH RECOGNITION IN THE CAR

2. 1 ACOUSTICAL ROBUSTNESS

A major challenge for speech recognition systems in the car is that the recognizer should be robust against a wide variety of noises coming from the engine, wind, tires, car radio, fan, windshield, horn,..., each with different frequency distributions (e.g. noise from the engine is in the low frequencies situated compared to noise from wind) and spatial characteristics of the noise.

2.2 FUNCTIONAL SPECIFICATIONS

2.2.1 Carkit extension for handy's

The primary request for speech recognition today comes for use with the telephone. Hands-free dialing is the number one feature that is requested. Full hands-free operation comes second. Full hands-free operation has very clear benefits to safety and comfort. It is clearly possible with today's technology, though it may involve a more expensive solution. This functionality will find its way into many carkits in the very near future. Most likely lawmakers in Europe and elsewhere will forbid using a phone in a driving car unless used in a hands-free fashion. In a few countries and states this is already the case today.

Hands-free dialing can be provided as a name dialing option only for the most common contacts or as a combination of name and number dialing as a full solution. The first requires speaker dependent word based recognition only, while the second requires both isolated word speaker dependent recognition and continuous digit speaker independent recognition. We assume the latter to be the real solution that will survive in the longer term.

2.2.2 Speech Recognition for Navigation

A fully different type of speech recognition is required to meet the demands of new generation car navigation systems. In first generation navigation systems speech operation has been provided with a limited vocabulary word based system. This had the advantage of being a cost effective and available solution. Nevertheless from a user interface point of view, there is clearly lot's of room for improvement if a higher end recognizer can be integrated.

Ease of use of a car navigation system requires that a user is able to identify geographical names (cities, streets, ..) out of lists of ten thousand and larger. While all this data might be highly structured (district, city, county, province, …) it is not obvious that one can make use of all that structure within the user interface. Eg you shouldn't have to tell a navigator product that you want to go to Munich, Bavaria. The latter should be more than obvious to everyone, including the navigation device. Hence user interface design will be a nontrivial issue. Because of the large vocabulary we clearly need a phoneme based speaker independent system and we surely will need fallback to spelling mode at several instances. Other issues that might trouble the speech recognition design is the phonetic lexicon as one has to deal with proper names, quite often borrowed from a different language.

2.3 HARDWARE REQUIREMENTS

Carkits are standalone black boxes as far as the car is concerned. They interact with the phone but not with the car as such. Major issues of course are the size and cost of the speech recognition module. As power can be drawn from the car's battery power needs will not be as critical as in other consumer applications. For the time being carkits will be largely part of the after market sales.

For the navigation kit we need to look more in the direction of adapting the design to be part of the control of the car radio system or even to interface with the digital bus onto which more devices could be plugged in. Speech recognition design has to take into account that other devices than the navigator may want to make use of speech I/O in the near future. Hence an open software architecture is required in which voice navigation is the prime application of interest.

2.4 SUMMARY

	Dialing	Navigation
Spk. Dependent Isolated Words	Yes	No
Spk. Indep. Cont Dig	Yes	Yes
Continuous Spk. Indep. Phoneme based	No	Yes
Spelling	No	Yes
Acoustic Robustness	High	High
Vocabulary size	< 100	> 10.000
Standalone Hardware	Yes	No
Open Architecture	No	Yes

3. SPEECH RECOGNITION DESIGN

3.1 SPEECH INPUT AND PREPROCESSING MODULES

3.1.1 Microphone Selection

The quality of the speech input is of utmost importance. Given the high noise environment a good quality microphone or a microphone array will be selected. In the case of a single microphone we have frequently used an AKG-400-II mouse microphone with a high pass filter characteristic built in, making it quite suitable for operation in the car.

Microphone arrays can give significant noise reduction, especially with non stationary interference [3]. However, because of cost considerations only a single microphone is acceptable with almost any car manufacturer. Also, multiple microphones are particular cumbersome to install with products of the after market sales, making them again not a very attractive solution. Not the number of microphones by itself is the cost, but as much all the wiring related to it. Given these considerations only single microphone solutions have been acceptable for all low and medium cost solutions.

3.1.2 Acoustic Echo Cancellation

Another source of noise comes from the car radio. Within the VODIS project [8] it has been shown that the acoustic echo can be removed efficiently within the car using a modified NLMS algorithm [7]. This is required if voice activated operation is required while the radio or CD is on. Today, canceling the echo of a single source in a single microphone can be solved; however, cancellation of stereo sounds and/or echo cancellation in the case of microphone arrays is still problematic problems and is not mature yet for deployment. Hence one of following compromises is possible today:

- switching to mono sound output at the detection of a speech input signal (or push-to-talk button); this can be combined with a single microphone speech input
- full suppression of all sound output at the detection of a speech input signal (or push-to-talk button); this input mode can be combined with multi-microphone speech input

3.1.3 Noise Suppression

Some kind of noise suppression will be present in any recognizer for high noise environments. Though a great deal can be covered with masking alone, explicit noise removal will help.

Several variants of spectral subtraction are possible [1,2]. A particular choice may depend on the fact if there is a need to resynthesize the speech from a cleaned up spectrum, as is the case if full hands-free operation is required of the telephone handset. If no resynthesis is required then the suppression will be performed inside the recognition preprocessor and on filterbank outputs rather than on individual FFT components.

3.2 ENGINE DESIGN FOR CARKIT

Major constraints are put on CPU and especially memory because of cost considerations. Hence compactness of code and feature representation may favor some algorithms over others even if they yield marginally better results.

As general framework a discrete density word based HMM system is used on the basis of cepstral features. Multiple codebooks containing cepstra, derivatives and an energy related codebook are used. Some modifications from a classical 'text-book' design will be required to meet all the criteria. Masking at a preset SNR and bandpass or highpass) filtering of spectral/cepstral features are 2 such techniques which are very trivial to apply and reach already a high degree of the required noise robustness [4,5,6]. These filters typically have a relatively large time constant (on the order of 200msec). This causes significant leakage of features from one phoneme to another and makes its use difficult in the context of a phoneme recognizer, but is perfectly acceptable in the case of a word based recognizer. This leakage phenomenon might actually improve the recognition performance in the latter case. For continuous speech recognition (continuous digits) it follows that cross word modeling must be tackled with care and content dependent word models are necessary to yield satisfactory performance (word error rate on the order of 2%).

Reduction in memory may be achieved by clustering of states based on cross entropy measures and by a coarse quantization of the individual discrete probabilities. Another important way to insure success is to reduce the acoustic mismatch between the training database and the operational car environment. This is achieved by using diverse training material obtained from in car recordings (which is very expensive and a compromising factor in getting speech recognition into the car) and which is further augmented with office quality speech artificially degraded with environmental noises. Appropriate robustness for speaker dependent training can equally be achieved by applying the latter technique.

3.3 ENGINE DESIGN FOR NAVIGATION

A significant part of the work for engine design for navigation at L&H has been performed within the framework of the EC LE project VODIS [8]. The ultimate goal of the project is to demonstrate an intuitive voice operated navigation system. Several key problems need to be tackled in the car which have not been demonstrated before: *phoneme* based speaker independent recognition of *very large* vocabularies and *spontaneous* mixed initiative dialogue.

3.3.1 The acoustic phonetic recognizer

Given the extreme requirements no compromises can be made on today's state of the art recognizers. Because of the need for a phoneme based recognizer, simple noise adaptation schemes which were appropriate in a word based recognizer, may no longer be sufficient and the spectral subtraction techniques are no longer an option but a necessity. Other competitive techniques such as parallel model combination [9] may not be selected because of implementation issues. Also a continuous density HMM recognizer may be needed to deal with the high task demands. Optimization of cost / performance is the critical issue.

3.3.2 The MMI

The target of mixed initiative and spontaneous input looks extreme at first sight. Initial trials within VODIS indicate that it may turn out not to be as bad as initially anticipated as there are only a few ways of saying that one wants to go from place A to place B.

The very high branching factors associated with geographical terminologies poses a much more series problem. A state of the art recognizer may be assumed to be capable of handling a few hundred words in parallel but not a few thousand as is required. Therefore a 2-tier mechanism in required in which plain speech recognition needs to handle all the more common names and where spelling (or partial spelling of words) can be used as fallback for less frequently used names. Rejection of out of vocabulary words need to achieve high accuracy to bring this to the user in an intuitive way. For the plain recognition of geographical names specialized phonetic dictionaries need to be constructed as many of those names or foreign or follow less than standard pronunciation rules. This problem is further stressed by the fact that users of navigation systems will often be unfamiliar with the places they want to go to (including their pronunciation) and may be from foreign origin. These latter problems are definitely some of the most challenging one in bringing a successful voice navigation product to the market.

ACKNOWLEDGEMENTS

This paper is partially based on work of and informal discussions with J. Smolders, H. Van hamme, F. Vanpoucke (L&H), L. Arevalo (Bosch), A. Waibel (U Karlsruhe), C. Antweiler (RWTH) and other members of the VODIS team.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech and Signal Processing 1984, pp 1109-1121.

[2] D. Van Compernolle, "Spectral Estimation using a Log-Distance Error Criterion Applied to Speech Recognition", Proc. Int. Conf. Acoust. Speech and Signal Processing, pp. 258-261, 1989

[3] D. Van Compernolle, W. Ma, F. Xie and M. Van Diest, "Speech Recognition in Noisy Environments with the Aid of Microphone Arrays", Speech Communications, Volume 9, 1990, pp 433-442

[4] H.G. Hirsch, P. Meyer and H.W. Ruehl, "Improved Speech Recognition using High-Pass Filtering of Subband Envelopes", Eurospeech 91, pp 413-416.

[5] H. Hermansky, "Compensation for the effect of the communication channel in auditory-like analysis of speech", EUROSPEECH 91, Genova, Italy

[6] J. Smolders, D. Van Compernolle, "In Search for the Relevant Parameters for Speaker Independent Speech Recognition", Proc. Int. Conf. Acoust. Speech and Signal Processing, 1993, pp 684-687

[7] C. Antweiler et al., "Optimal Step Size Control for Acoustic Echo Cancelation", Proc. Int Conf. Acoust. Speech and Signal Processing 1997, pp 295-298.

[8] VODIS, "Advanced Speech Technologies for Voice Operated Driver Information Systems", EC Language Engineering Project LE 1-2277.

[9] M. Gales and S. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", IEEE Transactions on Speech and Audio Processing 1996, pp. 352-359.