SPEECHDAT-CAR: TOWARDS A COLLECTION OF SPEECH DATABASES FOR AUTOMOTIVE ENVIRONMENTS

Henk van den Heuvel (1), Antonio Bonafonte (2),
Jerome Boudy (3), Sandra Dufour (3), Philip Lockwood (3),
Asuncion Moreno (2), Gaël Richard (3)

(1) SPEX, Nijmegen, Netherlands;
(2) Universitat Politechnica de Catalunya, Barcelona, Spain;
(3) Matra Nortel Communications, Bois d'Arcy, France;
E-mail address: Jerome.Boudy@matranortel.com

ABSTRACT

The SpeechDat-Car project is a 4th framework EC project in the Language Engineering programme. It aims at collecting a set of nine speech databases to support training and testing of robust multilingual speech recognition for in-car applications. The consortium participants are car manufacturers, telephone communications providers, and universities. This paper describes the background of the project, its organisation, and the design of the databases in terms of contents, speaker and environment coverage. It further addressses the recording platforms, the validation scenario, and the links with other projects.

1. INTRODUCTION

Speech recognition technology has reached a point where it is sufficiently robust to operate in relatively noisy environments, if sufficient efforts are made to cope with the noise problem. A notoriously noisy environment is the car cockpit. Here, speech recognition presents an attractive option. For instance, operating devices like telephones in the car can require substantial mental effort, dangerously detracting the concentration of the driver. It has already become illegal to operate a telephone by hand whilst driving in several European countries. Speech recognition is an obvious, safety-enhancing alternative. The R&D of reliable in-car voice driven applications requires an extensive set of representative data, both for training and testing speech recognisers. SpeechDat-Car continues the success of the SpeechDat project [1] in developing large-scale speech resources for a wide range of European languages. Whereas SpeechDat developed resources for the fixed and cellular telephone networks, SpeechDat-Car specifically addresses the challenge of in-car voice processing. The main objective of SpeechDat-Car is the development of a set of speech databases to support training of robust multi-lingual speech recognition for in-car applications. The applications are aimed at:

- Accessing remote teleservices and voice driven services from car telephones
- Controlling car accessories
- Voice dialling with mobile telephones in cars

In order to approach this challenge adequately, at least two types of recordings are necessary. First, wideband recordings (60-7000 Hz) are required for systems which are installed and operate in the car itself; second, narrow band recordings (300-3400 Hz) are needed for systems that operate centrally outside the car and obtain their spoken input from the driver over the cellular telephone network.

Both types of recordings are furnished by the forthcoming SpeechDat-Car databases (see section 3). SpeechDat-Car started in April 1998 in the 4th EC framework under project code LE4-8334 with a 30 months' project duration. It will produce resources for nine EU languages: Danish, English, Finnish, Flemish/Dutch, French, German, Greek, Italian, and Spanish. The consortium of the project comprises car manufacturers (BMW, FIAT, Renault, SEAT-Volkswagen), companies active in mobile telephone communications and voice-operated services (Bosch, Alcatel, Knowledge, Lernout & Hauspie, Matra, Nokia, Sonofon, Tawido, Vocalis), and universities (CPK, Denmark; DMI, Finland; IPSK, Germany; IRST, Italy; SPEX, Netherlands; UPC, Spain; WCL, Greece). The project management is with Matra Nortel Communications.

2. DATABASE DESIGN

Each of the nine databases will entail 600 sessions, recorded from at least 300 speakers with a maximum of two sessions per speaker. Each of the sessions comprises 119 items. Each item is recorded by several microphones,

and stored in a multiplexed speech file with a corresponding label file. The label file contains, among other things, a manually checked orthographic transcription of the item. Details on the database setup follow below.

2.1 Contents

Table 1 lists the items to be recorded and their distributions over the sessions.

Items	# per session	Tot. expected number of repetitions
Isolated digit	4	200 per digit
Digit strings		-
10 digits in isolation	1	600 per digit
Telephone number	3	-
Spontaneous tel number	1	-
Credit card number (set of 150)	1	4 per number
.PIN code (set of 150)	1	4 per number
Sheet number	1	-
Natural numbers	1	-
Money amounts	1	-
Dates		
Spontaneous	1	-
Absolute	1	50 per month name; 85 per day name
Relative (set of 10)	1	60 per date
Times		
Spontaneous	1	-
Analog	1	-
Names		
Birth city (spontaneous)	1	-
Most important cities (set of 150)	2	8 per name
Most important companies (set of 150)	2	8 per name
Forename or surname (spontaneous)	1	-
Forename + surname (set of 150)	1	4 per name
Spelling		
Forename/surname	1	-
Word/name	4	-
Artificial word	1	600 per letter
City name	1	-
Phonetically rich words	4	200 per phone
Phonetically rich sentences	9	250 per phone
Application words (set of 201)	67	200 per word
Spont. phrase with application word	2	-
Voice activation keywords (set of 5)	2	240 per word
Additional language-dependent application words (set of 10)	2	120 per word
TOTAL	119	

Table 1: SpeechDat-Car corpus contents

Adhering to these general specifications each database producing partner will compose a language-specific document, detailing the exact words used in the database. This document will be included in the final CD-ROMs.

2.2 Speakers and environments

The recorded database must be suitably sampled over the population. Each corpus should have a sufficient demographic coverage.

- **Speaker coverage**

At least 300 speakers per database are recorded. The speaker information (sex, age and dialect region) have an important impact on the final speech database, so they have to fit with the following requirements :

- *sex* : For each database the number of male and female speakers must be in the range 45-55% of the total number of speakers.

- *age* : the distribution of the speaker ages should be conform to the following table :

Age	0-15	16-30	31-45	46-60	61+
Proportion	0	≥ 20%	≥ 20%	≥ 15%	≥ 0%

All drivers should have a driving license (and should be older than 17 yrs for this reason).

- *regions* :

Considering the fact that only 300 speakers per database are recorded, the number of dialectal regions per language has to be limited to six, whilst a minimum of 50 speakers per region has to be included.

Remark : The constraints for speaker sex, age and dialect region are considered as being independent of each other.

- **Environment coverage**

Seven different recording conditions were defined for the SpeechDat-Car databases:

- car stopped with motor running;
- car in town traffic;
- car in town traffic, with open windows;
- car moving at a low speed with rough road conditions (freeway, out-town roads);
- car moving at a low speed with rough road conditions (freeway, under noisy conditions);
- car moving at a high speed with good road conditions (smooth asphalt - highway);
- car moving at a high speed with good road conditions (smooth asphalt – highway), with audio equipment on

Each speaker should be recorded in a maximum of two of these environments. The design should be such that each

of the seven conditions should be represented by at least 60 sessions. Other conditions that may significantly effect recording quality (such as rain, wind, or running car fan) will also be annotated.

3. RECORDING PLATFORMS

For each database two recording platforms are used in the project. The first is located in the car. It picks up the signals from a set of microphones attached to several optimal positions in the car cockpit and from a close-talk microphone. The other platform is at some fixed end outside the car and picks up the simultaneous signal transmitted over the GSM network. Both platforms are described below. Special attention is given to their interaction.

3.1 Car Platform

The mobile recording platform in the car (PltC) collects the wideband recordings to train services that operate in the car itself.

PltC is the master platform. Running on a PC it drives the recording process and controls the remote recordings on the far-end fixed platform (PltF). The PC is a shock-proof device which stores the in-car recordings directly in files on a hard disk. Recordings are made on four channels, each sampled at 16 kHz: one close-talk microphone (reference) and three far-talk microphones attached to the ceiling of the car: a. near the A-pillar; b. in front of the driver behind the sunvisor; c. over the midconsole near the rear window. In the condition with the car audio equipment switched on, both stereo loudspeakers are recorded instead of two of the far-talk signals. Adjustable microphone amplifiers are used (one for each microphone) and accommodated to the loudness of a speaker's voice and to the noise background of the recording environment.

A GSM phone is installed in the car with a commercial hands-free car-kit including a special hands-free microphone. The GSM phone is connected to the serial port of the PC. The recordings on the PltF are made via the GSM connection. The required remote control of the GSM phone runs via the PC and comprises the following functions: a. make a call; b. generate DTMF tones; c. hang up; d. detect dropped connection to PltF.

A 8.4'' display will be connected to the PC and attached to the top of the dashboard in front of the speaker. The experiment leader, who operates the PC, sends the prompts to the display. The PC contains a database with all prompt sheets for the individual sessions.

3.2 Fixed Platform

The fixed platform (PltF) records each item spoken in the car via the GSM channel with an ISDN connection. The data is stored at 8 kHz sampling frequency in A-law coding. The system operates as the slave of PltC. It features DTMF recognition (to receive messages from PltC), and 2. optimal synchronisation between PltC and PltF recordings under both full duplex and half duplex operation.

3.3 Communication and Synchronisation

The objectives of the communication between the two platforms are: 1. To detect if PltF is still alive during the recordings (and to repair a hang up); 2. to allow synchronisation of the recordings on the two platforms; 3. To allow separation of the items in individual files.

In the project several protocols have been proposed to meet these goals. These protocols comprise a series of beeps and DTMF-codes transmitted by PltC to PltF to ensure that each recorded item is preceded by a simultaneous beep on all recording channels to allow rapid off-line synchronisation of the recordings on both platforms.

4. VALIDATION

A quality check is carried out on each database produced in the project, in order to ascertain that the high quality requirements are met. Thus, it is warranted that the databases can be exchanged within the project on a basis of equality. This quality check is termed "validation" in the project, and it is carried out by an independent validation centre, SPEX, which is also in charge of the database validation in the other SpeechDat projects. "Independent" is to say that the validation centre itself does not produce a database in the project, neither as consortium member nor as third party.

The following aspects of a database are checked and compared to the validation criteria as agreed by the consortium: completeness and correctness of documentation; compliance to the database format specifications; completeness of recordings; correctness of the distributions of individual items; quality of the speech signals; balances of speaker and environmental distributions; completeness of the lexicon; quality of the orthographic transcriptions.

The validation procedure comprises three steps:

1. Pre-validation. Each partner sends a complete minidatabase of 6 speakers to the validation centre after the platform is installed and before the main recordings start. This minidatabase contains all speech and label files

and all other files that are required for a normal validation, but, of course, tailored to the 6 speakers included only. The goals of the pre-validation are:

- To detect errors in the database design before the main series of recordings start;
- To stimulate partners to write their database formatting software in an early stage of the project;
- To stimulate the validation centre to write the validation software in an early stage of the project.

2. Validation of the complete database. This validation is not performed on the full data set itself, but on the text files (e.g. label files, session tables, contents lists) that describe the full (speech) data set. Subsequently, a subset of complete sessions is selected for which the correspondence to the information in the metafiles is examined in detail for all speech files in the subset. The subset is also used to check the orthographic transcriptions in the label files. This is done by a native speaker of the language concerned who listens to the corresponding speech files and corrects the transcriptions if necessary.

3. Re-validation. This phase is entered only if the consortium and/or the producer considers it necessary that (part of) a database be rectified.

5. PROJECT LINKS AND INITIATIVES

The development of databases, like those of the SpeechDat-Car project, involves a prohibitive cost for any individual entity. This project puts together the efforts of individual partners to generate multilingual databases for the in-car environment. For each language, there is one partner responsible for the recordings. This is the owner of the database. The other partners have access to the nine databases produced in the project. The consortium edited a document which describes the procedures for third parties that are interested to produce a similar car database for another language and to exchange this database with consortium members on an equal-by-equal basis [2].

The databases produced in SpeechDat-Car will also be used in several European initiatives. To save cost and time, the VODIS-II project [8] will use the Italian and English databases produced within the SpeechDat-Car consortium. The SpeechDat-Car specifications include the main specifications of the VODIS-II project. These partners from both VODIS-II and SpeechDat-Car will have free access to the databases of both projects.

The results of the project will be presented in the European work group of the ETSI and in particular within the AURORA Group defining Distributed Speech Recognition, which particularly deals with the mobile environment, and thus also with the car environment. The SpeechDat-Car results will also be presented to the International work group ISO (TC22WG8). In particular, it could be interesting to make recommendations about the different type of speech utterances to be used for activation of the car-equipment (including telephone).

In order to produce reusable resources, the consortium has procured to define pre-standard specifications with respect to expected functionality and command words. A first draw-up of the specifications was widely distributed and the feedback was considered in order to fix the specifications.

When completed and fully validated the SpeechDat-Car databases will be made available via the European Language Resources Association (ELRA).

6. WHERE ARE WE NOW?

SpeechDat-Car started in April 1998. The first four months were used to formulate in detail the specifications that the databases should meet. A substantial part of the effort was devoted to the integration of the VODIS-II requirements into the SpeechDat-Car database design. Another time-consuming factor was the specification of the recording platforms and their interaction, together with the implementation of the hardware and software in the cars.

At the time this paper was written the specifications were finalised. They are laid down in a number of reports [3-6]. These public reports will be made available after conclusion of the prevalidation. This allows the consortium some fine-tuning of the specifications taking into account the experiences of a full recording and validation cycle. The prevalidation phase is expected to be completed by May 1999.

It is expected that at the time of this conference the recording platforms are installed and validated, that the first sessions are recorded and lined up for prevalidation, and that the recording of the major parts of the databases has commenced. Please visit our WWW-site for additional information [7].

7. REFERENCES

[1] Höge, H., Tropf, H., Winski, R., Van den Heuvel, H., Haeb-Umbach, R. & K. Choukri. *European speech databases for telephone applications*. Proceedings ICASSP 1997, München, Vol. III, pp.1771-1774, 1997.

[2] Richard, G. & J. Boudy *Baseline document for interested candidates for recording additional languages*. SpeechDat-Car Technical Report D0.1, 1999.

[3] Dufour, S. *Specification of the car speech databases*. SpeechDat-Car Technical Report D1.12, 1998.
[4] Van den Heuvel, H. *Validation criteria*. SpeechDat-Car Technical Report D1.3.1, 1998.
[5] Van den Heuvel, H. *Orthographic transcription conventions*. SpeechDat-Car Technical Report D1.3.2, 1998.
[6] Draxler, Chr. *Specification of database interchange format*. SpeechDat-Car Technical Report D1.3.3, 1998.
[7] URL: http://speechdat.phonetik.uni-muenchen.de/SP-CAR/
[8] URL:http://www2.echo.lu/langeng/en/le1/vodis/vodis.html