THE SPEECHDAT-CAR MULTILINGUAL SPEECH DATABASES FOR IN-CAR APPLICATIONS: SOME FIRST VALIDATION RESULTS

Henk van den Heuvel(1), Jerôme Boudy (2) Robrecht Comeyne (3),
Stephan Euler (4), Asuncion Moreno (5), Gaël Richard (2)

(1) SPEX, Nijmegen, Netherlands;
(2) Matra Nortel Communications, Bois d'Arcy, France;
(3) Lernout & Hauspie Speech Products, Ieper, Belgium;
(4) Robert Bosch GmbH, R&D Division, Stuttgart, Germany;
(5) UPC, Barcelona, Spain

H.v.d.Heuvel@let.kun.nl

ABSTRACT

The main objective of SpeechDat-Car is to develop a set of speech databases to support training and testing of multilingual speech recognition applications in the car environment. SpeechDat-Car started in April 1998 in the 4th EC framework under project code LE4-8334. The duration of the project is 30 months. Equivalent and similar resources for nine languages will be created: Danish, English, Finnish, Flemish/Dutch, French, German, Greek, Italian and Spanish. For each language 600 sessions will be recorded from at least 300 speakers. SpeechDat-Car commits itself to a strict validation protocol to ensure optimal quality and exchangeability of the databases. The first milestone in this respect is the validation of the recording platform and of a small subset of initial recordings. This paper briefly describes the database design and the recording platforms; next, it focuses on the objectives, the procedure, and some of the results of the early validation stage.

1. INTRODUCTION

The emergence of multiple 'in-car' accessories (radio, telephone, navigation systems,....) provides the driver of a modern car with additional functionalities but also puts him (or her) in a difficult situation since the manipulation of these accessories clearly distracts him from his main task (i.e. to drive). Automatic speech recognition (ASR) appears to be a particularly well adapted technology for providing voice-based interfaces (based on hands-free mode) that will enable such applications to develop while taking care of safety aspects. ASR applications for the car are nowadays seriously being investigated [4,5]. However, the car environment is known to be particularly noisy (street noise, car engine noise, vibration noises, bubble noise, etc...). To obtain an optimal performance for speech recognition, it is necessary to train the system on large corpora of speech data recorded in context (i.e. directly in the car). For this reason, language-specific initiatives for database collections have been developed since about 1990 (for an overview see [8]). The European project SpeechDat-Car aims at providing a set of uniform, coherent databases for nine European languages.

SpeechDat-Car continues the success of the SpeechDat project [2,6,7] 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 [9,11]. The applications are aimed at accessing remote teleservices and voice driven services from car telephones, controlling car accessories and voice dialling with mobile telephones in cars.

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 Nortel Communications, 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.

SpeechDat-Car commits itself to a strict validation protocol to ensure optimal quality and exchangeability of the databases. The first milestone in this respect is the validation of the recording platform and of a small subset of initial recordings. This paper briefly describes the database design and the recording platforms; next, it focuses on the objectives, the procedure, and some of the results of the early validation stage.

2. SPECIFICATIONS OF THE DATABASES

The databases are intended to provide material for both training and testing of speech recognisers for a large variety of products. In order to cover these products and also to provide a basis for future applications the following items are included in each of the databases:

- Application words spoken in isolation
- Navigation words: cities, regions, and road names including spellings
- Digits and numbers: e.g. telephone numbers, credit card numbers
- Dates and times
- Phonetically rich sentences
- Spontaneous sentences

Thus, in each session a total of 129 utterances are recorded. These utterances contain both spontaneous and read items. A total of 600 sessions per database will be recorded. The items are selected such that, counted over all sessions, an even distribution is achieved. E.g. in each session 4 isolated digits are prompted. Over the 600 session we obtain 2400 digits, i.e. 240 repetitions of each of the 10 digits. Each speech file in the databases will come with an orthographic transcription of all speech utterances and a pronunciation lexicon with a phonemic representation of all words in the transcriptions.

In automotive applications the driving conditions have a significant impact on the speech input. In the SpeechDat-Car project we distinguish seven environment conditions. In terms of amount of noise the conditions range from a stopped car with running engine up to driving with high speed on a highway with audio equipment (radio) switched on. Each environment condition should be represented by at least 10% of all sessions in a database.

Recruiting speakers and instructing them for the recording session is a very time consuming task and therefore each speaker records two sessions in different environments. The required 300 speakers are balanced with respect to age, gender and regional accent. For each country main dialectal regions are defined. Based on the specifications in the SpeechDat project between four and six regions are used per country. Preferably, the speaker should drive the car; in countries where this is forbidden the speaker should be the co-driver.

Exact details about the design of the databases can be found in [3].

3. RECORDING PLATFORMS

The configuration used in SpeechDat-Car to gather speech resources is based on two recording platforms :

1. a ‘mobile’ recording platform (PltM) that is installed inside the car, recording multi-channel speech utterances in a high bandwith mode (16kHz sample frequency)
2. a ‘fixed’ recording platform (PltF) located at the far-end fixed side of the GSM communications simultaneously recording the speech utterances coming from the car (8 kHz sample frequency)

PltM is the master platform; it uses a PC to drive the recording process and to control the remote PltF. Data acquisition is performed by some dedicated hardware in the PC and storage takes place directly onto the built-in hard disk. The PC is operated by the experiment leader in the car who calls for the prompts by pressing a key. The recordings are always made on four microphones: one close-talk microphone as reference and three far-talk microphones at fixed positions in the car which are identical for all databases. If the car radio is switched on, the two stereo loudspeaker signals will be recorded instead of two far-talk signals.

A complex synchronisation protocol was devised for the communication between the two platforms.

A GSM speech signal is sent from the car to a fixed platform connected in the far end of the GSM communication system. Before recording an item, PltM always checks whether PltF is alive; in case of a transmission interrupt, it tries to restore the connection and restart the recording at the item where the connection was lost in the previous session. The main characteristics of the fixed platform are:

- Connected to an ISDN line, either BRI or PRI
- Speech samples are stored onto the disk in the incoming A-law format.
- DTMF detection
- Full duplex operation

4. VALIDATION

4.1 General Validation Scenario

The validation scenario consists of two main parts: platform validation and database validation. The platforms are validated by submitting them to an expert test, which is a test of the platform equipment (section 4.2.1), and to a functional test, which is a test of the recording script by means of a questionnaire (section 4.2.2).

After approval of the platform, a database pre-validation is carried out, which has as its main goal the detection of major design errors before the actual recordings start (section 4.3).

Upon completion of the full database the producing partner sends a CD-ROM with all files, except the speech files to the validation centre for final validation. A selection of 16 calls is then made for which also the speech files are checked.

If the database is not approved by the consortium, or the producing partner wants to add some modifications to the database after it is accepted, then a revalidation by the validation centre takes place. The full list of validation criteria and the validation protocol is contained in [10].

Below we will only consider platform validation and pre-validation.

4.2 Platform Validation

This step in the process concerns the methodology for evaluating and validating the recording platforms and the recording script. The recording script is the program which prompts and records all 129 items of a complete session guided by the experiment leader. The complete platform validation is described in [1].

4.2.1 Expert Test

This test is the developers internal verification test of the platform. It is performed completely at each partner's own responsibility.

4.2.1.1 Listening test

The procedure for this test is as follows:

1. Record one expert speaker on all items according to the recording script;
2. Listen by one or more expert persons to all items previously recorded in order to detect errors in the recording chain: high clipping rate, truncations, highly distorted speech and very low SNR that is not generated by the environment;
3. Re-iterate previous steps if corrections or modifications of the platform are performed as long as the quality of the recorded speech is not judged as correct by the expert(s).

4.2.1.2 Load test

This test consists in stressing the platform to detect problems with logging or data transfer. This is executed by making recordings of 20 seconds for at least 100 items. In practice a script with 100 (or more) items of 20-sec each should be completed without any errors.

4.2.1.3 Interrupt Test

This test entails a power supply breakdown (electricity supply) and a communication (e.g. transmission) breakdown. After rebooting it is then verified that the platform is able to restart at the item where the connection was lost before.

Normally the platform must inspect all the files already created and will only record the remaining items for an identified session. If session identification fails then the system has to record all the items again. In the latter case the test consists only in checking that the platform re-starts the session at the beginning.

4.2.1.4 Stability Test

The average time between successive platform failures is measured under simulated conditions of real traffic over a long period of time. This test is passed if six sessions are recorded and no failures occurred that could not be diagnosed and corrected. The script to be used is the final script concerning structure, length of utterances and content.

Only if all four above tests are passed can the platform enter the functional test to be presented next.

4.2.2 Functional Test

For this test each partner has to find six test-speakers who completely perform one real life recording session. These test-speakers (three male / three female) can be selected in the partner's organisation provided they are not familiar with the recording platform to be tested. After the recording session these persons have to fill in a questionnaire which contains some detailed questions addressing the clarity of the instructions and the appreciation of the recording procedure. The questions are listed in Appendix A.

Instructions and the related questionnaire must be provided to each test person in his own language. It is mandatory that the questionnaire be filled after the test recording itself.

Each platform owner evaluates his own collection of questionnaires. For this purpose all the obtained marks/answers have to be entered into a table. In a first analysis each partner should present the main tendencies of the collected answers and also explain reasons of negative answers obtained in the questionnaires, if any.

Then all collections of questionnaires are centralised for global analysis and reporting.

4.3 Database Pre-Validation

For the pre-validation each partner sends a complete mini database of the six recorded sessions to the validation centre. This mini database contains all speech and label files and all other files that are required for a normal validation, but, of course, tailored to the six sessions included only. The main goal of the pre-validation is to detect errors in the database design before the main series of recordings start. It also serves to test (parts of) the software which the validation centre intends to use for the validation of the complete databases.

4.4 First Validation Results

By the end of April 1999 (the submission date for this paper) three databases successfully passed the platform test (in terms of expert and functional tests). One database was delivered for pre-validation, but not pre-validated at that time. These numbers are fairly low due to unexpected delays in the installation of the recording platforms.

The test results obtained so far show that a recording session takes about 45 minutes, instruction time not included. Most test persons did not mind to participate (indeed would participate again) and expressed a positive judgement about the recording procedure as a whole, although quite a few of them perceived the recording time as long.

In general the system records all items on PltM, also after an interrupt. But most sessions on PltF miss one or two items, due to a temporal GSM disconnection.

5. THE FUTURE

The pre-validation phase of the project was concluded in the Summer of 1999. Recordings and annotations are being made since Spring of 1999. Each partner has established a recording and annotation schedule which is checked on a monthly basis and monitored by the consortium through its Web pages [13]. According to our planning all databases will be delivered for final validation in the period January-April 2000 and be validated until project end at 1 October 2000.

After that, a set of nine high quality speech databases with in-car recordings will be available for the speech technology community via ELRA/ELDA [12]. These databases allow unique R&D activities due to the homogeneity of their designs which opens up a realm of comparative ASR studies in a variety of languages.

ACKNOWLEDGEMENT

Part of the SpeechDat-Car project is funded by the Commission of the European Communities, Telematics Applications Programme, Language Engineering, Contract LE4-8334.

REFERENCES

[1] Comeyne, R., Tsopanoglou, A., Boudy, J. Van den Heuvel, H., Chatzi, I. (1998) Validation of the recording platform, part A: methodology. *SpeechDat-Car Technical Report*, D2.3.a.

[2] Draxler, C., Van den Heuvel, H., Tropf , H. (1998) SpeechDat Experiences in creating Large Multilingual Speech Databases for Teleservices. *Proceedings LREC 98,* Granada, pp 361-366.

[3] Dufour, S. (1999) Specification of the car speech database. *SpeechDat-Car Technical Report*, D 1.12.

[4] Fischer, A. Stahl, V. (1999) Database and online adaptation for improved speech recognition in car environments. *Proc. ICASSP 99*, pp. 445-448.

[5] Hirtenberger, L. (1998) Man machine interaction in car information systems. *Proceedings LREC 98,* Granada, pp. 179-182.

[6] Höge, H., Tropf, H.S., Winski, R., Van den Heuvel, H., Haeb-Umbach, R. & Choukri, K. (1997) European speech databases for telephone applications. *Proceedings ICASSP 97*, Munich, pp. 1771-1774.

[7] Höge, H., Draxler, C., Van den Heuvel, H., Johansen, F.T., Sanders, E., Tropf, H. (1999) SpeechDat multilingual databases for teleservices: across the finish line. *Proceedings Eurospeech 99, Budapest.* (These Proceedings).

[8] Langmann, D., Pfitzinger, H. Schneider, Th., Grudszus, R., Fischer, A., Westphal, M., Crull, T., Jekosch, U. (1998) CSDC – the MoTiV car speech data collection. *Proceedings LREC 98, Granada,* pp. 1107-1110.

[9] Sala, M., Sanchez, F., Wengelnik, H., Van den Heuvel, H., Moreno, A., Le Chevalier, E., Deregibus, E., Richard, G. (1999) Speechdat-Car: Speech databases for voice driven teleservices and control of in-car applications. *Proceedings EAEC 99, Barcelona.*

[10] Van den Heuvel, H. (1999) Validation criteria. *SpeechDat-Car Technical Report*, D1.3.1.

[11] Van den Heuvel, H., Bonafonte, A., Boudy, J., Dufour, S., Lockwood, Ph., Moreno, A., Richard, G. (1999) SpeechDat-Car: towards a collection of speech databases for automotive environments. *Proc.s of the Workshop for Robust Methods for Speech Recognition in Adverse Conditions, Tampere*.

[12] ELDA: http://www.icp.grenet.fr/ELRA/home.html

[13] SpeechDat Family: http://www.speechdat.org

APPENDIX A: QUESTIONNAIRE FOR THE TEST SPEAKERS

Each test speaker in the functional test has to answer the following questionnaire. For each question the mark must range between the appreciation levels given in brackets.

a) Have you participated in this type of recording before?
Never; Once; More than Once

b) From the instructions given to you, did you understand the goal of these recordings?
Yes; No - If no, why?

c) How did the system response time appear to you?
1 (very long); 2 (long); 3 (medium); 4 (short); 5 (very short)

d)Did you at some point in the session want to end the recordings?
Yes; No - If yes, why?

e) How well could you follow the displayed information during the session?
1 (bad); 2 (poor); 3 (fair); 4 (good); 5 (excellent)

f) How well did you appreciate the screen readability?
1 (bad); 2 (poor); 3 (fair); 4 (good); 5 (excellent)

g) Were there any items that you found difficult to pronounce?
Yes; No - If yes, which ones and why?

h) Were there any items that you did not want to pronounce?
Yes; No - If yes, which ones and why?

i) What did you think of the actual length of the sessions?
1 (very long); 2 (long); 3 (medium); 4 (short);5 (very short)

j) What is your general impression of the whole procedure?
1 (bad); 2 (poor); 3 (fair); 4 (good); 5 (excellent)

k) Would you be ready to participate in another session of this type of recording?
Yes; No - If no, why not?

l) General or specific comments concerning the overall procedure
free text